C

Carbon Nanotubes, Thermo-mechanical and Transport Properties of

H. Rafii-Tabar[1,2]

[1] Department of Medical Physics and Biomedical Engineering, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[2] Department of Nano-science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Article Outline

Glossary

Scanning tunneling microscope (STM) A nano-technology instrument capable of imaging the topography of conducting surfaces with atomistic resolution. It can also be used to manipulate individual atoms and molecules, and construct nano-scale structures.

Atomic force microscope (AFM) A nano-technology instrument for investigation of material surfaces on atomistic and molecular scales. It can be used to map the topography of non-conducting surfaces, by sensing the inter-atomic forces, and produce three-dimensional images of these surfaces.

Spintronics Refers to the field of spin-based electronics rather than charge-based electronics. It will lead to a new type of device that is based on the use of electron spin for transfer and storage of information.

Brillouin zone In solid state physics several Brillouin zones can be defined. The first zone is defined as the Wigner–Seitz primitive cell of the reciprocal lattice. The *n*th Brillouin zone is defined as the set of points that are reached from the origin by crossing $(n-1)$ Bragg planes.

Monte Carlo (MC) method In computational modeling, this method provides a probabilistic scheme for solving a variety of problems by employing powerful sampling techniques. In nano-science and condensed matter physics, one application of this method is for computing the minimum energy state of a nano-structure.

Stochastic dynamics (SD) method This refers to the computer simulation method wherein the Langevin equation of motion, describing the random behavior of a particle, is solved as opposed to the deterministic MD method in which Newton's equations of motion are solved.

Nano-electromechanical systems (NEMS) These are nano-technology based systems that are the smaller versions of the micro-electromechanical systems (MEMS). They are capable of measuring small displacements, sensing minute amount of substances, and performing rotary motions. NEMS can be constructed via either the top-down approach, i. e., via miniaturization of the micro-scale devices, or via the bottom-up approach, i. e., by positioning single atoms or molecules so that a complex and functional nano-system is built from below the nano-scale.

Ab initio approach This is the first-principles approach to the computation of the properties, especially the electronic-structure properties, of nano-scale systems using quantum-mechanical concepts and methods. In this method, the structure of a molecule, for instance, is obtained purely from a knowledge of its composition by solving the Schrödinger equation.

Definition of the Subject

Carbon nanotubes form the fourth allotrope of crystalline carbon after graphite, diamond, and a variety of caged-like fullerene molecules, and were discovered in the early 1990s. Their mechanical properties make them stronger than steel, and their thermal conductivity is faster than copper. They have very exotic electronic-conduction properties, namely, by changing their geometry, or introducing topological defects into their structure, their electronic conductance can change from metals to semiconductors. They can also be used to store gases and transport fluids. Furthermore, nano-scale robots, machines, and sensors can be constructed from them, and these can be used to deliver drugs to specific locations in the body, or detect individual cancer cells, or be used as molecular filters to separate minute particles from the environment. Carbon nanotubes are referred to as the fabric of nanotechnology, and will play a central role in the future development of this technology. Understanding the properties of nanotubes, via computational simulation studies, has been one of the most intensive areas of research in physical sciences during the past 20 years.

Introduction

The current scientific-technological arena is distinguished by the presence of *four* distinct and highly complex disciplines, namely:

1) information science, underlying the fields of information technology, high performance computing, and computational science,
2) molecular genetics and molecular biology, underlying the fields of genetic engineering and bio-technology,
3) neuro and cognitive sciences, and their related fields of neural- and brain-based technologies,
4) nano-science, and its related fields of nano-technology and molecular-scale engineering, making atom-by-atom manipulation of the physical, biological and smart matter possible.

These four areas are strongly correlated, and the post-modern era in science and technology will be characterized by research efforts to establish, step-by-step, a close synergy among these fields, leading to their total integration within a *unified* and holistic framework, promoting the convergence of all these seemingly separate and compartmentalized branches and paving the way for the emergence of a *convergent* technology [49]. Such a technology will produce advanced man-made machines, devices, components, and materials that carry some of the most essential characteristics of the biological and smart systems, such as self-repair, self-assembly, self-correction of internal faults, re-production, ability to establish communication, and adaptation to unfamiliar environments.

The convergence takes place at the nano-scale (1–100 nanometers) since the main building blocks of the physical, biological and smart matter, i. e., the physical and biological *nano-structures*, and the laws governing their evolution, are formed at this scale. It is, therefore, no exaggeration to state that nano-scale science and nano-scale technology [17] represent the main components of the 21st century science and technology. It is now accepted that nano-science can be defined as *the study of structures, dynamics, and properties of systems wherein one or more of spatial dimensions varies in the nanoscopic range*. At this scale, the dynamics and properties of systems are distinctly different, very often in quite unforeseen ways, from those in microscopic and macroscopic systems.

Nanoscopic structures operate on very reduced time and energy scales, and are constructed from a *countable* number of atoms and molecules. Their sizes are located between those of molecules and micro-structures, and their distinguishing feature is their high aspect, or surface-to-volume, ratio. Nano-technology represents the combination of atomic and molecular level techniques and instrumentations for constructing nano-scale devices, systems, and components from the *assemblies* of individual nano-structures. On the other hand, nano-technology is an *enabling* technology, injecting nano-structures into micro- and macro-scale, devices, materials, and systems to enhance, or radically change, their electronic, mechanical, optical and thermal properties. For instance, nano-structured materials, i. e., materials with nano-sized grains, or materials injected with nano-grains, can show very different mechanical, thermal, electronic and optical properties. For instance, it is known that nano-structured Fe, Cu and Ni have electrical resistances respectively 55%, 15% and 35% higher than the coarse-grained polycrystalline samples [38].

Nano-structures can be assembled by a precise positioning of the atoms at specified locations, using such highly sensitive devices as the scanning tunneling microscope (STM) [9], and the atomic force microscope (AFM) [8]. These devices can provide access to the detailed topography, crystal structure, and the electronic-structure maps at material surfaces and of individual nano-structures. Their use has led to the design and fabrication of devices, and materials, that manifest novel physical, chemical and biological properties.

Among the multitude of nano-structures currently used in nano-technology, carbon nano-structures occupy a central position, with wide-ranging applications in practically all the nano-technology sub-fields. Carbon is a fundamental element, crystallizing in several allotropes. In the 1970s, *diamond* and *graphite* were the only known forms of condensed carbon, and a third allotrope, the cage-like *fullerene* molecules, was discovered and synthesised in macroscopic amounts by mid 1980s. In 1990s, the fourth allotrope, the multi-walled carbon *nanotube* (MWCNT), and the single-walled carbon nanotube (SWCNT) were synthesized [20,21]. Several growth techniques, namely, the arc-discharge, the laser ablation, and recently the chemical vapor deposition method, have been developed. The appearance of nanotubes has led to the emergence of new and very active areas of research within several fields; in computational and experimental nano-science, industrial nano-technology, theoretical and experimental condensed matter physics and chemistry, materials science and engineering, bio-physics, medical nano-technology, molecular genetics, nano-biotechnology, information technology device fabrication, optics, electro-mechanical systems and electronics, carbon nanotubes are viewed as highly relevant nano-structures with extensive potential applications.

Originally, the carbon nanotubes were of *multi-walled* variety, with outer diameters in the range of 4–30 nm and lengths of up to 1 μm [21]. These consisted of seamless concentric cylindrical shells, separated by 3.4 Å. Each shell was a rolled-up version of a two-dimensional graphene sheet. The *single-walled* nanotubes were synthesized later [6,22], and consisted of single graphene sheets with a typical diameter on the order of 1.4 nm, similar to the diameter of a C_{60} molecule, and lengths usually on microscopic orders [21]. It was also found that these SWCNTs can bundle together to form a *rope*, normally patterned as a hexagonal array, with a morphology very similar to porous materials, and membranes, with nanometer spaces available both inside the nanotubes and in the interstitial channels between them that can be utilized for *storage* of adsorbed gases, and *flow* of fluids, turning them into filtering channels and molecular sieves.

Another form of SWCNTs, called single-walled carbon *nanohorn* (SWCNH), has also been synthesized [23]. This horn-shaped material has a closed structure, with a cylindrical part and a horn-tip part. Its internal space is not normally available as a storage medium, but heat-treatment within an oxygen environment promotes the appearance of windows on the its walls, allowing for gas and liquid particle transfer to the interior.

Nanotubes incorporating fullerenes, such C_{60} molecules, have also been synthesized [55], and are referred to as *peapods*. The internal space of the SWCNT is filled with a regularly-arranged formation of fullerene molecules [21]. Peapods, when heated to 1000–1200°C, transform into double-walled carbon nanotubes (DWCNTs), as the encapsulated molecules coalesce [21]. Peapods provide a good model of nano-platforms for targeted drug delivery.

As will be discussed later, an SWCNT is identified by two *chiral* indices (n, m) that determine its diameter, and depending on their values, three *different* classes of nanotube are observed. These are the $(n, 0)$, or the *zigzag*, nanotube, the (n, n), or the *armchair* nanotube, and the $(2n, n)$, or the general *chiral* nanotube. They have very different *electronic* conduction properties. The graphene sheet is a semi-metal with a zero band-gap. The electronic states of an infinitely long nanotube are continuous in its axial direction, but are quantized along its circumference. For the (n, n) armchair nanotubes, there always exist electronic states that cross the corner points of the first Brillouin zone, making these nanotubes a *metallic* material. For the (n, m) nanotubes, if $(n - m)/3 \neq$ an integer, the electronic states miss the corner points, and these nanotubes behave as semi-conductors with a band-gap that scales with the diameter d of the tube as $1/d$, and which is on the order of 0.5 eV. If $(n - m)/3 =$ an integer, however, certain electronic states are located on the corner points, and these nanotubes behave as semi-metals, but become small-band semi-conductors due to the curvature-induced rehybridization [14]. These remarkable electronic-conduction properties, that are linked with the geometry of the material, are also very partial to local deformations in the geometry. Recent exploitation of the electronic properties of carbon nanotubes includes the field of spin-electronics, or *spintronics* that utilizes the spin of the electrons to transfer and store of information.

Nanotubes have also extraordinary mechanical, thermal, mass-transport, and gas-storage properties. For example, their Young's modulus, being a measure of their stiffness, is estimated to be of the order of 1.5 to nearly 5 TPa, and their tensile strength is far higher than that of steel. They are, therefore, strong candidates as functional units in molecular-scale machinery, and highly complex nano-electromechanical systems (NEMS), and as probe *tips* in STM and AFM [56]. Their high thermal conductivity, far exceeds that of diamond. They can also act as media for the transport and storage of various types of gases and fluids.

The investigation into nanotube properties has prompted an intensive experimental and theoretical/computational research, leading to major discoveries, and the appearance of several thousand patents and publications in fields of basic sciences, engineering, and medicine. One of the most active areas has involved the use of predictive computational modeling of their mechanical, thermal and mass transport properties [44]. These computational modelings and simulations have been performed via two distinct methodologies:

1) using quantum mechanical methods, such as the density functional theory (DFT) of atoms and molecules [36] that provides an *ab initio* approach to computation of the properties of nano-scale systems composed of several *ten* to several *hundred* atoms,
2) using classical statistical mechanics methods, such as the molecular dynamics (MD), Monte Carlo (MC) and stochastic dynamics (SD) simulation methods [2] that can handle nano-scale structures composed of several *thousand* to several *billion* atoms. These simulation methods require the use of phenomenological inter-atomic potentials to model the energetics of the system. These potentials play a very crucial role, and the more accurate they are, the closer the simulation results approach the experimental data. A good deal of efforts have been focused to develop highly accurate inter-atomic potentials for different classes of materials, particularly the covalently-bonded ones. Some of these potentials are *many-body* potentials, and most of the simulations concerning carbon nanotubes have used these many-body potentials.

Research into the properties of carbon nanotubes has led to the appearance of several informative reviews, among which one can mention those by [1,5,16,40,43,45]. In addition, an impressive number of encyclopedias and handbooks [7,34,48,51] covering most sub-fields of this topic are also available.

In this review I have surveyed the field of computational modeling of the thermo-mechanical, transport and storage properties of nanotubes. In the interest of saving space, I have only described, albeit very briefly, one of the computational methods that has been extensively employed in these research studies, namely the MD simulation method. The interested reader can find a very good summary of the other methods in [43]. An extensive part of the modeling-based research into the mechanical properties has employed concepts from the field of continuum elasticity theory, using such structures such as *curved plates, shells, vibrating rods, and bent beams*. An easy-to-follow introduction to these topics can found in my book [44].

The organization of this review is as follows. In Sect. "Geometry of SWCNT, MWCNT and SWCNH", the description of the geometrical structure of the SWCNTs, MWCNTs and the SWCNHs is briefly considered. In Sect. "Simulation at Nano-Scale" the basic principles of the classical MD simulation method is presented. In Sect. "Modeling Fluid Transport and Gas Storage Properties of Nanotubes", the flow of fluids and the storage of gases in nanotubes are considered, while in Sect. "Modeling the Mechanical Properties of Nanotubes", the mechanical properties are considered. Section. "Modeling the Thermal Properties of Nanotubes" summarizes the thermal conductivity and specific heat properties of nanotubes, and Sect. "Concluding Remarks" presents an overall summary together with the future directions.

Geometry of SWCNT, MWCNT and SWCNH

The SWCNT

We can construct an SWCNT by rolling a 2D graphene sheet into a cylinder. Consider a lattice point O as the origin in a graphene sheet, as shown in Fig. 1. The 2D Bravais lattice vector of the graphene sheet $\mathbf{C}_h$, referred to as the *chiral* vector, can be used to reach any other equivalent point Z on the sheet. The vector $\mathbf{C}_h$ is constructed from the pair of unit-cell vectors $\mathbf{a}_1$ and $\mathbf{a}_1$ as

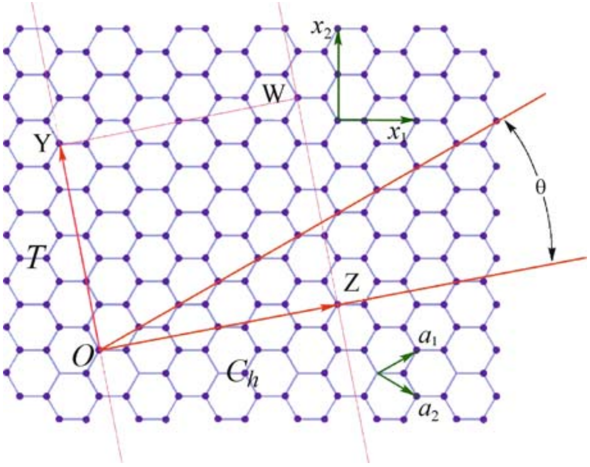
$$\mathbf{C}_h = n\mathbf{a}_1 + m\mathbf{a}_2 \,, \tag{1}$$

where m and n are a pair of integers, and

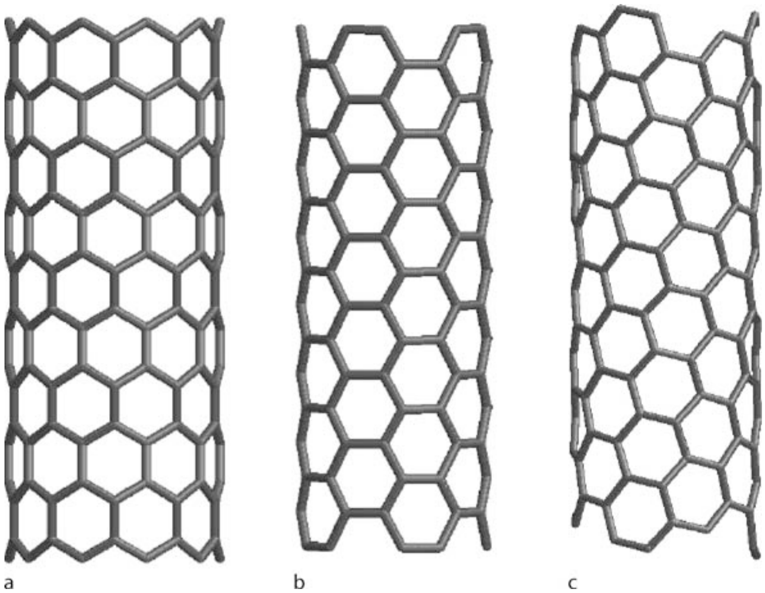
$$\mathbf{a}_1 = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)a \,, \quad \mathbf{a}_2 = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)a \,, \tag{2}$$

where $a = 2.46$ Å is the lattice constant of graphite

$$a = \sqrt{3}a_{\text{C–C}} \,, \tag{3}$$



Carbon Nanotubes, Thermo-mechanical and Transport Properties of, Figure 1
The geometry of a two-dimensional graphene sheet, showing the relevant vectors that characterize a single-walled carbon nanotube (SWCNT). Figure based on [15]



Carbon Nanotubes, Thermo-mechanical and Transport Properties of, Figure 2
The outlines of three types of nanotube: a a (10,0) zig-zag nanotube; b a (5,5) armchair nanotube; c a (7,3) general chiral nanotube

and a_{C-C} is the carbon-carbon bond length. The angle θ that $\mathbf{C}_h$ makes with the zigzag axis of the sheet is referred to as the chiral angle, and (n, m) are called the chiral indices. If the sheet is rolled, and the point characterized by the pair of values of (n, m) coincides with the origin O, then an (n, m) nanotube is generated. The *zigzag* direction corresponds to $\theta = 0$, and when the direction of rolling is along this axis, a zigzag nanotube is formed. When $\theta = \pi/6$, the direction is called the *armchair* direction, and rolling along this direction generates an armchair nanotube. A nanotube generated for any other value $0 < \theta < \pi/6$ is referred to as a general chiral nanotube. Figure 2 shows the schematic representations of these three types of nanotube.

The circumference of the nanotube, i. e., the length of the chiral vector, is given by [15]

$$L = |\, \mathbf{C}_h \,| = a(n^2 + m^2 + nm)^{\frac{1}{2}} \,, \tag{4}$$

and, hence, the diameter is

$$d_t = \frac{L}{\pi} = \frac{(n^2 + m^2 + nm)^{\frac{1}{2}}}{\pi} a \,. \tag{5}$$

Other relationships connecting the chiral angle with the chiral indices are [15]

$$\sin\theta = \frac{\sqrt{3}m}{2(n^2 + m^2 + nm)^{\frac{1}{2}}} \,,$$

$$\cos\theta = \frac{2n + m}{2(n^2 + m^2 + nm)^{\frac{1}{2}}} \,,$$

$$\tan\theta = \frac{\sqrt{3}m}{2n + m} \,. \tag{6}$$

Hence, if $n = m$, then $\theta = \pi/6$ and the resulting nanotube is an (n, n), or an armchair, nanotube. If $m = 0$, then $\theta = 0$, and an $(n, 0)$, or a zigzag, nanotube is obtained. The nanotubes in the general chiral category are $(2n, n)$.

The rectangle bounded by the chiral vector $\mathbf{C}_h$ and the vector $\mathbf{T}$, i. e., OYWZ, where Y is the first lattice point through which the vector passes, is the unit-cell of the nanotube. This vector is written as [15]

$$\mathbf{T} = t_1\mathbf{a}_1 + t_2\mathbf{a}_2 \,, \tag{7}$$

where t_1 and t_2 are a pair of integers, related to m and n via

$$t_1 = \frac{2m + n}{d_R} \,, t_2 = -\frac{2n + m}{d_R} \,, \tag{8}$$

and d_R denotes the highest common divisor (HCD) of $(2n + m, 2m + n)$, given by

$$d_R = \begin{cases} = d & \text{if } (n - m) \text{ not a multiple of } 3d \,, \\ = 3d & \text{if } (n - m) \text{ multiple of } 3d \,, \end{cases} \tag{9}$$

where d is the HCD of (n, m). The length T of the vector $\mathbf{T}$ is given by

$$T = |\, \mathbf{T} \,| = \frac{\sqrt{3}L}{d_R} \,. \tag{10}$$

The unit-cell of the nanotube contains $2N$ atoms, where

$$N = \frac{2(n^2 + m^2 + nm)}{d_R} \,. \tag{11}$$

Carbon nanotubes can be capped at both ends [15]. This can be done by bisecting a C_{60} molecule into two hemispheres and joining these with a cylindrical nanotube of the same diameter as the molecule.

An SWCNT is a low-energy structure with a distinct topology, brought about by distorting the geometry of the 2D graphene sheet into the third dimension. The act of rolling the sheet calls for the expenditure of a modest amount of strain, or curvature, energy, which is the difference between the total energy of a carbon atom in an SWCNT and that in a graphene sheet. In nanotubes with a diameter less than 1 nm, this energy penalty is more significant since the bond angles deviate far below the ideal 120° angles in the graphene sheet.

The MWCNT

Two forms of MWCNTs have been found in experiments. These are a thermodynamically-stable system composed of nested coaxial cylinders [20], and a metastable type with a scroll structure [26]. The transition from the latter type to the nested type is attributed [26] to the presence of the dislocation-like defects in the scroll type. The inter-shell gap in an MWCNT is estimated to be ≈3.4 Å [50], very close to the inter-planar gap in the graphite crystal. Several other values of this gap, in the range 3.42 Å to 3.75 Å, have also been reported [25], with the gap increasing when the nanotube diameter decreases. The simplest type of an MWCNT is a DWCNT. To construct a DWCNT, the chiral vector of the inner SWCNT, with indices (n_1, m_1), is related to the chiral vector of the outer SWCNT, with indices (n_2, m_2). These indices are related via

$$m_2^2 + m_2 n_2 + (n_2^2 - \kappa(n_1, m_1)) = 0 \,, \tag{12}$$

where

$$\kappa(n_1, m_1) = \left[\frac{2\pi(r_{t1} + r_g)}{a}\right]^2 \,, \tag{13}$$

and r_{t1} and r_g are respectively the radius of the inner nanotube, and the inter-shell gap. Then n_2 and m_2 can be obtained by solving (12). For example, for a (9,6) inner nanotube, and $n_2 = 15$, we find $m_2 = 9.96$ which rounds to $m_2 = 10$ and, therefore, the nanotube (9,6) can be nested inside the nanotube (15,10).

The SWCNH

These objects are distinguished by their horn shape. Pure SWCNH samples are relatively easier to produce than pure SWCNT samples [33]. SWCNHs are always produced with closed tips, incorporating pentagons into their hexagonal lattices to close the ends. They consist of a tubule part, and a conical cap whose average angle is 20°, implying that the caps contain five pentagon rings and several hexagon rings. The average length of SWCNHs is 30-50 nm, and the separation between neighboring SWCNH is about 0.35 nm. The average diameter of the tubular parts is 2-3 nm, larger than the 1.4 nm diameter of a typical SWCNT. The interstitial spaces in an SWCNH assembly provide the external micropore and mesopore spaces [33,35], where pores with a width of less than 2 nm are called micropores, and those with a width between 2 and 50 nm are called mesopores. The width of the internal pores in SWCNHs is close to the critical size of 2 nm [33]. Oxidation produces windows on the walls of a closed SCWNH [35], making available 11% and 36% of the closed spaces at $T = 573$ K and 623 K respectively [32].

Simulation at Nano-Scale

The time-evolution of the energetics and dynamics of complex nano-structures and nano-processes, involving several billions of atoms, can be accurately studied via simulation methods that use concepts from classical statistical mechanics. The information gained from these simulations can be profitably used in the design of nano-scale components, nano-structured materials, and structures dominated by nano-interfaces. Furthermore, the construction of functional assemblies of nano-structures requires a deep understanding of the interaction between individual nano-structures, and this can be handled by the use of such simulation methods. We briefly consider the essential elements of one of the most popular simulation methods at the nano-scale, namely the MD simulation method.

MD Simulation Method

Classical MD simulation [2,18,46] studies the motion of individual atoms in an assembly of N atoms or molecules employing either the Newtonian, or the stochastic dynamics, when the initial spatial configuration and velocities of these atoms are given. In MD, the N-atom structure is confined to a simulation cell of volume V, and this cell is replicated in all dimensions, generating the periodic images of itself and of the N atoms. This periodic boundary condition (PBC) is necessary to compensate for the undesirable effects of the artificial surfaces associated with the finite size of the simulated system. The energetics of the system are modeled by two- or many-body inter-atomic potentials $H_I(r_{ij})$, and in the simulation the force experienced by each atom is obtained from

$$\mathbf{F}_i = -\sum_{j>i} \nabla_{\mathbf{r}_i} H_I(r_{ij}) \,, \tag{14}$$

where r_{ij} is the distance between atoms i and j. The $3N$ coupled equations of motion are integrated numerically via a variety of numerical techniques, all based on the finite-difference method. One popular integration scheme is the velocity Verlet algorithm [2], whereby the positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$ of the atoms of mass m_i are updated at each simulation time step $\mathrm{d}t$ according to

$$\begin{aligned} \mathbf{r}_i(t+\mathrm{d}t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\mathrm{d}t + \frac{1}{2}(\mathrm{d}t)^2\frac{\mathbf{F}_i(t)}{m_i} \,, \\ \mathbf{v}_i(t+\frac{1}{2}\mathrm{d}t) &= \mathbf{v}_i(t) + \frac{1}{2}\mathrm{d}t\frac{\mathbf{F}_i(t)}{m_i} \,, \\ \mathbf{v}_i(t+\mathrm{d}t) &= \mathbf{v}_i(t+\frac{1}{2}\mathrm{d}t) + \frac{1}{2}\mathrm{d}t\frac{\mathbf{F}_i(t+\mathrm{d}t)}{m_i} \,. \end{aligned} \tag{15}$$

At each $\mathrm{d}t$, the exact *instantaneous* values of thermodynamical observables, such as the temperature and pressure

$$\begin{aligned} T_{\text{ins}} &= \frac{1}{3Nk_\text{B}} \sum_i^N \frac{|\mathbf{p}_i^2|}{m_i} \,, \\ P_{\text{ins}} &= \frac{1}{3V}\left(\sum_i^N \frac{|\mathbf{p}_i^2|}{m_i} + \sum_{i=1}^N \sum_{j>i}^N \mathbf{r}_{ij}.\mathbf{F}_{ij}\right) \,. \end{aligned} \tag{16}$$

can be computed, where $\mathbf{p}_i$ is momentum of particle i, k_B is the Boltzmann constant and $\mathbf{F}_{ij}$ is the force experienced by atom i due to atom j. These instantaneous data then allow for the computation of time-averaged values at the conclusion of the simulation.

Constant-Temperature MD: The Nosé–Hoover Method

MD simulations are generally performed on closed isothermal systems, represented in statistical mechanics by canonical ensembles wherein N, V, and the temperature T of the members are all fixed [37]. A constant-temperature MD simulation can be realized in a variety of ways [42], and a method that generates the canonical ensemble distribution in *both* the configuration and momentum parts of the phase space was proposed by Nosé [42]. Its application leads to a modification of the equations of motion (15) to the following forms [59]

$$\begin{aligned} \mathbf{r}_i(t+\mathrm{d}t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\mathrm{d}t \\ &\quad + \frac{1}{2}\mathrm{d}t^2\left[\frac{\mathbf{F}_i(t)}{m_i} - \eta(t)\mathbf{v}_i(t)\right] \,, \\ \mathbf{v}_i(t+\frac{\mathrm{d}t}{2}) &= \mathbf{v}_i(t) + \frac{\mathrm{d}t}{2}\left[\frac{\mathbf{F}_i(t)}{m_i} - \eta(t)\mathbf{v}_i(t)\right] \,, \\ \eta(t+\frac{\mathrm{d}t}{2}) &= \eta(t) + \frac{\mathrm{d}t}{2\,Q}\left[\sum_i^N m_i\mathbf{v}_i^2(t) - gk_\text{B}\,T\right] \,, \\ \eta(t+\mathrm{d}t) &= \eta(t+\frac{\mathrm{d}t}{2}) \\ &\quad + \frac{\mathrm{d}t}{2\,Q}\left[\sum_i^N m_i\mathbf{v}_i^2(t+\frac{\mathrm{d}t}{2}) - gk_\text{B}\,T\right] \,, \\ \mathbf{v}_i(t+\mathrm{d}t) &= \frac{2}{2+\eta(t+\mathrm{d}t)\,\mathrm{d}t} \\ &\quad \cdot \left[\mathbf{v}_i(t+\frac{\mathrm{d}t}{2}) + \mathrm{d}t\frac{\mathbf{F}_i(t+\mathrm{d}t)}{2\,m_i}\right] \,. \end{aligned} \tag{17}$$

where Q is given by

$$Q = g\,k_\text{B}\,T\,\tau^2 \,, \tag{18}$$

and η is the friction coefficient of the heat bath connected to the system, τ is the relaxation time of this bath, normally of the same order of magnitude as $\mathrm{d}t$, and $g = 3(N-1)$ is the number of degrees of freedom. The parameter τ controls the speed with which the bath damps down the fluctuations in the temperature.

Modeling Fluid Transport and Gas Storage Properties of Nanotubes

Fluid flow and gas storage in nano-scale objects, like carbon nanotubes, is significantly different as compared with the corresponding activities in microscopic and macroscopic structures. For instance, fluid flow in a nano- scale machine is not similar to the flow in a large-scale machine, since in the latter case, there is no need to consider the atomic structure of the fluid, and it can be characterized by its viscosity, density and other bulk properties [61]. Furthermore, in large-scale objects, the so-called *no-slip* boundary condition is often invoked, according to which the fluid velocity is negligibly small at the fluid-wall interface.

Reduction in length scales immediately introduces new phenomena, such as *diffusion*, into the system. Moreover, for flow, or storage, in nano-scale objects, the dynamics of the walls, the fluid, or the gas, and their mutual interaction, must all be considered. We note that the dynamics of the walls can be strongly size-dependent.

An added difficulty is that, at nano-scale, standard classical notions such as pressure and viscosity may also not have a clear definition. For example, the surface area of a nano-scale object, such as a nanotube, may be ambiguous. Notwithstanding these difficulties, modeling fluid flow and gas storage in nanotubes has been an active area of research, with practical consequences. In this section we consider a set of studies, each highlighting one aspect of transport and storage properties of nanotubes.

Modeling Fluid Flow

Let us consider the MD simulation of the flow of fluids, composed of Ar and He particles, through an SWCNT [61]. In simulations, the fluid atoms were always dynamic, whereas for the nanotube, the first 10 and the last 10 rings were always frozen, and the rings in the middle section were treated either as static or as dynamic rings, giving rise to either a static or a dynamic nanotube. The significant result was that the dynamic nanotube slowed down the fluid particles faster than the static nanotube. It took about 42 ps for the fluid velocity to slow down to 2.5% of its initial velocity in the static nanotube, while in the dynamic nanotube this time was 15 ps. Furthermore, the motion of the nanotube perturbed the motion of the adjacent fluid atoms, causing faster randomization and promoting hard collisions with the walls that slowed down the fluid steadily until it essentially came to a stop. Similar results were also obtained for the Ar fluid, whose atoms are some 10 times heaver than He atoms. In this case, it was found, however, that the velocity damping was even more pronounced. It was also found that the higher the fluid density, the faster it slowed down.

Another MD simulation has addressed the rapid imbibition of oil (decane) in a (13,13) SWCNT, at temperature $T = 298$ K [58], modeling the SWCNT as a rigid cylinder. The computed fluid density ρ inside the nanotube showed that the nanotube first filled rapidly with the low density fluid, followed by the higher density fluid at lower speeds tending to ≈ 150 m/s for $\rho > 0.2$. Such speeds are comparable to the speed of sound in air (340 m/s). It was found that, the flow terminated at the end of the nanotube, i. e., although the nanotube was open, its end acted as a barrier to further flow. It was also found that the imbibition was much faster than predicted by the classical Washburn equation which relates the penetration length of the fluid, in a macroscopic tube, to its radius, the surface tension, the viscosity and the time.

The effect of confinement of liquid water in a nanotube on its vibrational and rotational spectra has been also been investigated via MD simulations [30] for four different-sized nanotubes (6,6), (8,8), (10,10) and (12,12). The liquid water density was an input. The significant aspect of the results was the presence of a frequency band, between 3640 and 3690 cm^{-1} , which is absent in bulk water. This was referred to as a vibration frequency, and its position shifted to smaller values when the nanotube radius increased. It was concluded that this *vibration band was purely an effect of confinement*. Adsorption line shape in the domain between 300 and 1000 cm^{-1}, that corresponds to the rotational motion of the molecules, was also obtained, and again a shift to lower frequencies was observed compared with those in the bulk. Hence, there was a direct connection between the radius of the nanotube and some of the observed frequency shifts.

Transport of water through a solvated carbon nanotube has also been modeled [19]. The central channel of the nanotube is strongly hydrophobic in character. It was, however, found that the initially empty channel rapidly filled up with water from the surrounding container, and was occupied during the entire simulation run. The water density inside the cylinder was found to exceed the bulk density. Since the water molecules interact weakly with the carbon atoms, the continuous hydration of the nanotube interior was unexpected.

Modeling Gas Storage

Gas storage can take place in single nanotubes, as well as in their bundles, where three adsorption sites are available, namely the interstitial channels between the SWCNTs, the outer surfaces of the nanotubes in the bundle, and the groove channels, i. e., the wedge-shaped spaces on the outer surface of the bundle where two SWCNTs meet [60]. Simulation results indicate that while H_2, He and Ne atoms can adsorb in the interstitial channels, other types of atoms are too large to fit into these spaces. The clarification of adsorption sites and the adsorbed amount has been the focal point of research in this area.

Hydrogen Storage The storage of atomic and molecular hydrogen has occupied a very prominent position in the field of gas adsorption in nanotubes. The adsorption of para-hydrogen in (9,9) and (18,18) SWCNTs and their bundles has been modeled at several temperatures [62]. The inter-nanotube spacing within a bundle, called the van der Waals gap, is approximately 3.2 Å, measured from the center of the nanotube walls. The gap g is defined as ($g = a - D$), where a is the lattice spacing, and D is the diameter of the nanotube. The total storage in the bundle is a sum of interior and interstitial adsorptions. The computed isotherms at $T = 77$ K show that, in the low pressure range, the (9,9) nanotube bundle gives a higher gravimetric and volumetric densities than the (18,18) nanotube bundle, since the hydrogen-nanotube interaction in the former case is stronger. At pressures above 5 and 10 atm, the situation is reversed. The volume of a (9,9) nanotube is of such a size that it can hold one layer of adsorbed molecules on the interior surface and a column of molecules in the center of the nanotube, whereas the volume of the (18,18) nanotube can accommodate three layers of hydrogen in three concentric rings and a column in the center of the nanotube.

It was found that the interstitial adsorption formed 14% of the total adsorption in the (18,18) nanotube bundle, in the (9,9) nanotube bundle this proportion was less than 1%. Increasing the temperature from $T = 77$ to 298 K lowered the adsorbed amount in both the (9,9) and the (18,18) bundles by approximately a factor of 5 at 100 atm. The computed adsorption isotherms on the external and internal surfaces of an isolated (18,18) SWCNT showed that it adsorbed slightly more than the (9,9) SWCNT at $T = 77$ K, except at the lowest pressures. A significant portion of the total adsorption took place on the external surfaces.

Quantum-mechanical based computation of adsorption of hydrogen atoms in (5,5) and (10,10) SWCNTs and in a (5,5)@(10,10) DWCNT has also been performed [27]. The results show that there were two chemisorption sites on the exterior and the interior of the nanotube, and that a form of H_2 *molecule* formed within the empty pore inside the nanotube. Furthermore, hydrogen capacity increased linearly with the diameter of the nanotube. The results for the DWCNT showed that the SWCNTs were better for higher hydrogen storage than the DWCNT.

Another study on the physisorption of *molecular* hydrogen H_2 in SWCNT bundles [64] has shown that small-diameter bundles are preferable for storage, and that the delamination of the nanotube bundle increases the gravimetric storage capacity.

In another simulation [63], the question of optimization of the gap in a bundle of SWCNTs for maximum storage of H_2 has been addressed for two bundle geometries, namely, a square bundle and a triangular bundle, each bundle composed of four nanotubes of types (9,9), (12,12), (18,18). The gap was varied to obtain the optimum separation for adsorption. The results on the gravimetric and volumetric densities, as a function of the gap, for $T = 298$ K and $P = 50$ atm showed that for $g = 3.2$ Å, i. e., the smallest gap, much of the volume and surface area in the bundles were unavailable for adsorption in both packing geometries. Increasing this gap allowed adsorption to take place on the external surface of the nanotubes. It was found that, the optimum value of g, as a function of temperature, was $g = 6$ Å at $T = 298$ K, and $g = 9$ Å at $T = 77$ K in a *triangular* bundle of the (9,9) nanotubes.

The adsorption of H_2 molecules in *charged* (9,9) SWCNTs, and their two-dimensional rhombic-shape bundles, has also been investigated [53]. The results for isolated SWCNTs showed more second layer adsorptions in the charged nanotubes, as compared with the uncharged nanotubes, and that the difference in adsorption was about 10%–20% at $T = 298$ K and 15%–30% at $T = 77$ K. Furthermore, it was noticed that the negatively charged nanotubes adsorbed more than the positively charged nanotubes at $T = 77$ K. At the higher temperature of $T = 298$ K, there was no observable difference. Volumetric densities for the adsorption in the bundles of charged and neutral nanotubes, as a function of the van der Waals gap, showed that the charged bundle had an enhanced adsorption of H_2 because of charge-multipole interactions. The results from these simulations imply that an enhancement of the hydrogen storage capacity of nanotubes can be achieved by using electric fields and gradients.

Adsorption of Other Gases The quantum ground-state energy of He atoms, as well as the isosteric specific heat of low-density He gas, adsorbed on the external groove sites in a bundle of 37 nanotubes have been computed [52]. The energy was found to be -22.7 meV, in close agreement with the experimental value, but larger than the computed energy for the He atom adsorbed in the interstitial channels, which is -27.9 meV. The data on specific heat for adsorption in the grooves were computed, and were seen to be very different from the data for adsorption in the interstitial positions.

The adsorption of Xe atoms in (10,10) SWCNTs, and their hexagonal bundle, at $T = 95$ K was modeled [54]. The results showed that the rather strong interaction between highly polarizable Xe atoms and the nanotubes led to significant adsorption even at very low pressures. The adsorption isotherms for internal and external surfaces showed a significant difference. Adsorption in the bundle of SWCNTs was also considered, and as the interstitial space was only 2.9 Å wide, for a large bundle, adsorption was expected to take place only in the interior of the nanotubes.

Adsorption of NO_2, O_2, NH_3, CO_2, H_2O gas molecules on (10,0), (17,0), (5,5) and (10,10) SWCNTs has also been obtained [66]. Different adsorption sites, such as the top of a carbon atom, the top of the center of a C-C bond, and the top of a carbon hexagon, were considered. The results showed that, in general most of the gases were weakly bound to the nanotubes, i. e., they were physisorbed. Furthermore, most of the molecules were charge donors with a small charge transfer, whereas O_2 and NO_2 were charge acceptors. The influence of adsorption on the electronic properties of SWCNTs was also investigated, and it was found that the adsorption of NH_3 and NO_2 did not significantly affect either the conduction or the valence band of a (10,0) nanotube. The density of states (DOS) of the nanotube decorated with NH_3 was very close to that of a pure nanotube. This behavior was also observed for all charge donor molecules, i. e., N_2, H_2O, CO_2 etc. The adsorption of NO_2 significantly changed the shape of DOS, implying that semi-conducting SWCNTs can be turned into p-type conductors following adsorption of NO_2 or O_2.

Another study [12] has considered the influence of lattice dilation, due to adsorption of He, Ne, H_2, D_2, Ar and CH_4 gases in the interstitial channels of the SWCNT triangular bundles, on the further uptake of these gases. The basic idea was that, lattice dilation allows small molecules to significantly enhance their binding energies in the interstitial channels without a substantial increase in the internanotube interaction energy. The results revealed that the chemical potential for H_2 was about 200 K lower than the value in the absence of dilation, i. e., there was a greater tendency for the adsorption in a dilated medium than in an undilated one. In the case of D_2, the lattice dilated less than in the case of H_2. Also the dilation in the He and Ne adsorptions was less than 0.5%. For H_2, Ar and CH_4 the consequences of dilation were significant.

Adsorption of Gases in SWCNH Assemblies Since SWCNHs are closed structures, access to their interior must first be achieved by opening potential entry points. The most common method for pore opening is heat treatment in oxygen [3], and oxidization at $T = 573$ and 623 K respectively leads to the opening of 11% and 36% of the spaces. Assemblies of SWCNHs also possess interstitial channels.

Adsorption of nitrogen, at $T = 77$ K was used to estimate the volumetric porosity of aggregates of heat-treated SWCNHs [32]. The interaction of the corn (apex) parts of the horns was not taken into account, and only the contribution of tube parts was considered. It was found that the interstitial site in the middle of three SWCNHs provided the strongest site. However, since only a one-dimensional array of N_2 molecules could be packed at this site the adsorption capacity was limited there. Strong adsorption could also occur on the sites within the SWCNHs near their walls, leading to the formation of a monolayer. A third site, of weaker type, also located in the interior of SWCNHs, was also identified. Therefore, three sites were found for adsorption.

Adsorption of N_2 in the internal pore space of *individual* SWCNHs has been addressed [35], with the interaction of the corn part of the SWCNH included. Adsorption isotherms were obtained for different diameters D of the tube part of the SWCNH, and since the SWCNHs were partially oxidized at $T = 693$ K, almost all of the pore spaces were available for adsorption. The computed isotherms showed two gradual steps around the normalized pressures of 10^{-5} and 10^{-1}, and the smaller the value of D, the shaper was the step. The average pore width w of the internal space was found to be $w = 29$ Å. Experimental images of three regions, namely the tip, the neck and the tube parts of the SWCNHs showed that the adsorption in these regions was considerably different, and that it began at the tip and then moved to the neck space, and in the tube space the molecules were adsorbed on the internal wall in multilayers.

Modeling the Mechanical Properties of Nanotubes

Modeling the mechanical properties of carbon nanotubes occupies a central position in research on nanotubes, and computational simulation in this field has employed highly sophisticated atomistic and continuum-elasticity models. These studies reveal that nanotubes have high tensile strength, large bending flexibility and high aspect ratios, properties that make them an ideal superstrong material. Research has focused on the elastic constants, Poisson's ratio and Young's modulus, and their dependence on the diameter and chirality of the nanotubes. A very interesting part of this research has used well-established continuum-based theories of curved plates, thin shells, bending beams and vibrating rods. An extensive presentation of these topics, as applied to nanotubes, can be found in [44]. These theories have been used either as independent computational tools to model the elastic properties and deformation modes of nanotubes, or in conjunction with the atomistic models to interpret the results obtained from them. *A key issue in this respect is the identification of the total energy of an atomistic system with the elastic strain energy of an equivalent continuum system.*

Some research problems pose a challenge to both types of the modeling. Large MWCNTs are a case in point, as their sizes present a challenge to MD simulations, and the continuum mechanics is most successful in the limit of a thin-shell. A very fruitful approach would be a *multiscale* modeling that couples the continuum, the atomistic and the quantum description of the nanotubes within a unified model, and until such models are developed, a pragmatic attitude would be the application of both atomistic and continuum modeling to extract as much useful information as possible.

Structural Deformations of Nanotubes

Modeling the axial compression, bending and torsion modes of a (7,7) SWCNT has been performed via both MD simulations, and the continuum elasticity theory of plates [65]. Let us consider the axial compression modeling, whereby a compression was applied to the ends of the nanotube by shifting the end atoms along the axis in small steps, and then relaxing the nanotube while keeping each end constrained. The variation of the strain with the strain energy showed a set of discontinuities that were associated with the rather exotic shape changes undergone by the nanotube. The evolution of this deformation mode within the continuum theory was examined via the expression for the deformation energy of a curved plate. Using the MD generated information as input, the curved plate parameters, i. e., its flexural rigidity and in-plane stiffness, that appear in the continuum energy expression were obtained. It was found that at a critical level of imposed strain, and employing the plate parameters, the nanotube buckled sideways, very similar to what happened in the MD simulation, and that the lowest value of the strain was close to that obtained in the MD simulation. The overall results from all the simulations showed that nanotubes could sustain extreme strains without developing any brittle or plastic transitions or dislocations, and beyond the range covered by Hooke's law, their behavior can be described by continuum-based modeling.

The mechanism whereby uniaxially strained (5,5) and (10,10) nanotubes release their stored strain energy has also been investigated [11]. The nanotube was subjected to a 10% tensile strain, and following equilibration at 1800 K, the first signs of mechanical yielding appeared. This yield was accompanied by the formation of *topological* defects, beyond a critical tension. The first defect to appear was a 90° rotation of a C–C bond about its center which transformed four hexagons into a double pentagon-heptagon (5-7-7-5) defect. This bond rotation is referred to as the Stone–Wales transformation. The energy barriers to the bond rotation that produced the defect were calculated, and it was found that for both types of nanotubes it decreased with an increase in the strain. Furthermore, the results showed that by annealing the (5,5) nanotube, subject to 10% strain, under the condition of high temperature ($T = 2500$ K), the (5-7-7-5) defect was reversible and the nanotube recovered its original hexagonal network.

In a further simulation [10], a (10,10) nanotube containing an initial (5-7-7-5) defect was strained up to 15%, and it was observed that octagonal defects and higher order rings appeared, i. e., the initial defect had acted as a nucleation center for other defects. The simulation showed that armchair nanotubes could display a ductile behavior via plastic flow of the (5-7) dislocation cores. These topological defects could change the chirality of the nanotube, and when dislocations covered a sizable distance of the nanotube wall, they left behind a nanotube whose chirality changed according to the dislocation rules, i. e., $(n, n) \rightarrow (n, n-1) \rightarrow (n, n-2) \ldots$, i. e., an armchair nanotube transformed into a zigzag geometry.

The effect of hydrogen storage on the deformation properties of (10,10) and (17,0) SWCNTs has also been investigated [67]. The SWCNTs were pre-stored with 4.17 or 8.34 wt % H_2 gas, and the maximum strain and maximum tensile force as a function of hydrogen wt %, and temperature were obtained, showing that both the maximum strain and maximum force decreased with the storage of H_2, and that the reduction at $T = 600$ K was much bigger than that at $T = 300$ K. The reductions were attributed to the competition between the C–C bonds and the C–H bonds. The effect of storage on the (17,0) nanotube did not seem to be as significant as on the (10,10) nanotube.

Elastic Properties of Nanotubes

Elastic properties of nanotubes, i. e., their elastic constants, Young's, bulk and shear moduli, have been computed [28,29] for a number of SWCNTs and MWCNTs. An inspection of the results shows that the shear and bulk moduli of SWCNTs and MWCNTs of different diameters were respectively around 0.4 TPa and 0.75 TPa, and Young's modulus for these nanotubes was around 1 TPa, comparing well with the experimental value of 1.3 TPa. Comparison of the SWCNT and MWCNT results shows that:

1) the elastic moduli change very little with the number of walls,
2) the inter-shell van der Waals interaction does not affect significantly the elastic moduli of multi-walled nanotubes,
3) there is a large anisotropy in elastic properties of both SWCNTs and MWCNTs,
4) the elastic properties are basically the same for all nanotubes with radii greater than 1 nm.

An MD simulation [41] has also computed Young's modulus and Possion's ratio ν of the SWCNTs, giving

$$\begin{aligned} &E = 1.20\,\text{TPa}\,, \quad \nu = 0.27 \text{ for a (10,10) nanotube}\,, \\ &E = 1.10\,\text{TPa}\,, \quad \nu = 0.28 \text{ for a (5,5) nanotube}\,, \end{aligned} \tag{19}$$

Another MD simulation [24], using an energy approach (EA) and a force approach (FA), has computed the axial E_z and the rotational E_θ Young's moduli of a set of (n, n) SWCNTs, for $n = 6$ to 19, and has found that E_z and E_θ had little dependence on the radius of the nanotubes in both approaches, and that both moduli were very close. The average values for different nanotubes in both approaches were

$$E_z = E_\theta = \begin{cases} 1.344 \pm 0.02 \quad \text{TPa (EA)} \,, \\ 1.238 \pm 0.01 \quad \text{TPa (FA)} \,. \end{cases} \tag{20}$$

The bulk moduli of (6,6), (10,10) and (8,4) SWCNT hexagonal bundles subject to hydrostatic pressure have also been computed [47], giving the bulk modulus of the bundle K_{bund}, and the individual nanotubes K_{tube} as

$$\begin{aligned} K_{\text{bund}} &= 37\,\text{GPa} \,, \\ K_{\text{tube}} &= 230\,\text{GPa} \,. \end{aligned} \tag{21}$$

Modeling the Thermal Properties of Nanotubes

Thermal stability of nanotubes is crucial to the safe operation of nano-scale devices. Measurements of the specific heat and thermal conductivity of microscopic *mats*, covered with compressed ropes of nanotubes have been made, providing information on the ensemble-average thermal properties of bulk-phase materials, rather than on individual nanotubes. Equilibrium and non-equilibrium types of MD simulations have also been applied. In the former case, the equilibrium time-correlation functions of the heat flux operator are computed and are used in the Green–Kubo relation to obtain the thermal conductivity. In the latter method, hot and cold reservoirs are coupled to the two ends of the nanotube, and the thermal conductivity is obtained via the average heat flux .

Thermal Conductivity of Nanotubes

In an MD simulation [4], combining both equilibrium and non-equilibrium methods, the temperature-dependence of the thermal conductivity λ of an isolated (10,10) nanotube has been computed, showing that at $T = 100$ K the nanotube displayed an unusually high value $\lambda = 37\,000$ W/m K, far exceeding that of pure diamond, i. e., $\lambda = 3320$ W/m K. Comparing the thermal conductivity of an isolated nanotube with those of an isolated graphene sheet and bulk graphite showed that the nanotube had a very similar heat transport properties as those of a hypothetical isolated graphene sheet. The results on the three carbon structures showed that below $T = 300$ K, the thermal conductivity of the graphene sheet was higher than that of the nanotube, and between $T = 300$ K and 400 K they were nearly equal. For graphite, the thermal conductivity of the basal plane also peaked near the $T = 100$ K point, and its value in the range $T = 200$ K to 400 K was always smaller than the values for other two structures.

Another MD simulation [13] investigated the dependence of λ on the nanotube length, and the presence of defects and vacancies, showing that after an initial rise in λ, up to a nanotube length of $L = 100$ Å, it converged to a fairly constant value $\lambda = 29\,800$ W/m K along the nanotube axis for L in the range 100 to 500 Å. This was similar to the value obtained by [4]. In the computations, the thickness of the nanotube was taken to be 1 Å. The results on the variation of λ with the vacancy concentration showed that it decreased smoothly from $\lambda = 14\,000$ W/m K for a concentration of 0.002 to $\lambda = 4000$ W/m K for a concentration of 0.01. Furthermore, the role of Stone–Wales (5-7-7-5) defects was examined, showing that λ decreased from $\lambda = 35000$ W/m K at concentration of 0.5 (on the 1/1000 scale) to about $\lambda = 12\,000$ W/m K for a concentration of 3. Both the rate of decrease and its absolute amount were less than the case with vacancies.

For a bundle of (10,10) nanotubes, it was found that $\lambda = 9500$ W/m K along the axis, very close to that of the in-plane bulk graphite at $\lambda = 10\,000$ W/m K, while perpendicular to the axis, $\lambda = 56$ W/m K, similar to the out-of-plane conductivity in graphite at $\lambda = 55$W/m K.

Specific Heat of Carbon Nanotubes

The low-temperature specific heats C_V of individual SWCNTs, and of their bundles composed of n SWCNTs of type (9,9), with $n = 1, 2, \ldots, 7$, and of MWCNTs have been computed [39], based on the phonon contribution to C_V. The results showed three different patterns for C_V of an individual SWCNT for $T < 100$ K, namely at very low temperatures,

$$C_V \propto T^{\frac{1}{2}} \,, \tag{22}$$

while with an increase in temperature,

$$C_V \propto T \,, \tag{23}$$

and above $T \approx 5$ K, the optical phonons began to make a contribution. Furthermore, it was found that, the $T^{\frac{1}{2}}$ behavior diminished when nanotubes were added to the bundle, and for a bundle composed of 7 nanotubes, this form of dependency on T was no longer observable and $C_V(T)$ approached the experimental value.

For MWCNTs, the specific heat of a five-walled nanotube, (5,5)@(10,10)@(15,15)@(20,20)@(25,25), was calculated, showing that by starting with the (5,5) nanotube and adding more shells to it, the $T^{\frac{1}{2}}$ part diminished and disappeared and was replaced by a linear dependence on T. If the number of shells were increased, C_V tended to that of graphite. It should be remarked that the variation of C_V with T for MWCNTs differed significantly from the experimental behavior.

Concluding Remarks

In this survey we have reviewed the modeling studies on the mechanical, thermal and transport properties of carbon nanotubes. We have seen that in modeling the fluid flow correctly, the motion of both the fluid and the walls, and their mutual coupling must be considered. In this field, the future research must address the use of such classical concepts as *viscosity* and *pressure* that are not clearly defined at nano-scales. Furthermore, flow-induced structural instability poses another important research problem.

In the storage of gases, we have seen that adsorption can take in both the internal spaces and in the interstitial channels, and that SWCNTs provided a better storage medium for hydrogen than the MWCNTs. Furthermore, the small-diameter ropes were more preferable than large-diameter ones. The future research should help clarify the question of volume storage capacity of nanotubes and the ways this capacity could be improved as the data obtained in different studies are not yet unitary. They indicate, however, that the adsorption in nanotubes is higher than in most carbonaceous materials.

In the studies of mechanical and elastic properties, both the atomistic and continuum theories have been used successfully in providing deep insights into the deformation and elastic properties of nanotubes. as well as providing quantitative estimates of the pertinent variables. Studies focusing on obtaining Young's modulus of the nanotubes have shown that the estimates obtained depended on the theoretical models or the inter-atomic potentials used. Computed values ranged from 1 TPa to 5.5 TPa for SWCNTs, and for the MWCNTs an average experimental value of about 1.8 TPa seemed to be the accepted result. Future research could employ better continuum-based models to help reduce the number of very time-consuming and costly MD simulations that are currently employed to obtain the estimates of mechanical and elastic properties.

Finally, we have seen that the thermal conductivity of the SWCNT as a function of temperature is unusually high, implying that the mean free path of phonons was quite large in nanotubes. The presence of defects and vacancies also substantially affected this conductivity. In the computation of the specific heat, the main finding concerned the significant contribution of phonons to the specific heat as compared with the electrons, and in the low temperature regime, several important scaling laws on the behavior of the specific heat have been derived. The future research could consider such problems as the computation of the thermal properties of nanotubes containing defects, or decorated with adsorbed molecules, or nano-scale devices composed of carbon nanotubes.

Bibliography

1. Ajayan PM, Ebbesen TW (1997) Nanometre-size tubes of carbon. Rep Prog Phys 60:1025–1062
2. Allen MP, Tildesley DJ (1987) Computer simulation of liquids. Clarendon Press, Oxford
3. Bekyarova E, Kaneko K, Kasuya D, Murata K, Yudasaka M, Iijima S (2002) Oxidation and porosity evaluation of budlike single-wall carbon nanohorn aggregates. Langmuir 18:4138–4141
4. Berber S, Kwon YK, Tománek D (2000) Unusually high thermal conductivity of carbon nanotubes. Phys Rev Lett 84:4613–4616
5. Berbholc J, Brenner D, Nardelli MB, Meunier V, Roland C (2002) Mechanical and electrical properties of nanotubes. Annu Rev Mater Res 32:347–75
6. Bethune DS, Kiang CH, de Vries MD, Gorman G, Savoy R, Vazquez J, Beyers R (1993) Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. Nature 363:605–607
7. Bhushan B (ed) (2004) Springer handbook of nanotechnology. Springer, Berlin
8. Binnig G, Quate CF, Gerber C (1986) Atomic force microscope. Phys Rev Lett 56:930–933
9. Binnig G, Rohrer H (1982) Scanning tunnelling microscopy. Helv Phys Acta 55:726–735
10. Buongiorno Nardelli M, Yakobson BI, Bernholc J (1998) Brittle and ductile behaviour in carbon nanotubes. Phys Rev Lett 81:4656–4659
11. Buongiorno Nardelli M, Yakobson BI, Bernholc J (1998) Mechanism of strain release in carbon nanotubes. Phys Rev B 57:R4277–R4280
12. Calbi MM, Toigo F, Cole MW (2001) Dilation-induced phases of gases adsorbed within a bundle of carbon nanotubes. Phys Rev Lett 86:5062–5065
13. Che J, Cagin T, Goddard WA III (2000) Thermal conductivity of carbon nanotubes. Nanotechnology 11:65–69
14. Dai H (2002) Carbon nanotubes: opportunities and challenges. Surf Sci 500:218–241
15. Dresselhaus MS, Dresselhaus G, Saito R (1995) Physics of Carbon Nanotubes. Carbon 33:883–891
16. Dresselhaus MS, Dresselhaus G, Saito R, Jorio A (2005) Raman spectroscopy of carbon nanotubes. Phys Rep 409:47–99
17. Drexler KE (1992) Nanosystems: Molecular Machinery, Manufacturing and Computation. Wiley, New York
18. Haile JM (1992) Molecular dynamics simulation: Elementary methods. Wiley, New York
19. Hummer G, Rasaiah JC, Noworyta JP (2001) Water conduction through the hydrophobic channel of a carbon nanotube. Nature 414:188–190

20. Iijima S (1991) Helical microtubules of graphitic carbon. Nature 354:56–58
21. Iijima S (2002) Carbon nanotubes: past, present, and future. Physica B 323:1–5
22. Iijima S, Ichihashi T (1993) Single-shell carbon nanotubes of 1-nm diameter. Nature 363:603–605
23. Iijima S, Yudasaka M, Yamada R, Bandow S, Suenaga K, Kokai F, Takahashi K (1999) Nano-aggregates of single-walled graphitic carbon nano-horns. Chem Phys Lett 309:165–170
24. Jin Y, Yuan G (2003) Simulation of elastic properties of single-walled carbon nanotubes. Compos Sci Technol 63:1507–1515
25. Kiang CH, Endo M, Ajayan PM, Dresselhaus G, Dresselhaus MS (1998) Size effects in carbon nanotubes. Phys Rev Lett 81:1869–1872
26. Lavin JG, Subramoney S, Ruoff RS, Berber S, Tománek D (2002) Scrolls and nested tubes in multiwall carbon nanotubes. Carbon 40:1123–1130
27. Lee SM, An KH, Lee YH, Seifert G, Frauencheim T (2001) A hydrogen storage mechanism in single-walled carbon nanotubes. J Am Chem Soc 123:5059–5063
28. Lu JP (1997) Elastic properties of carbon nanotubes and nanoropes. Phys Rev Lett 79:1297–1300
29. Lu JP (1997) Elastic properties of single and multilayered nanotbes. J Phys Chem Solids 58:1649–1652
30. Martí J, Gordillo MC (2001) Effects of confinement on the vibrational spectra of liquid water adsorbed in carbon nanotubes. Phys Rev B 63:165430-1–165430-5
31. Mizel A, Benedict LX, Cohen ML, Louie SG, Zettl A, Budraa NK, Beyermann WP (1999) Analysis of the low-temperature specific heat of multiwalled carbon nanotubes and carbon nanotube ropes. Phys Rev B 60:3264–3270
32. Murata K, Kaneko K, Steele WA, Kokai F, Takahashi K, Kasuya D, Hirahara K, Yudasaka M, Iijima S (2001) Molecular potential structures of heat-treated single-wall carbon nanohorn assemblies. J Phys Chem B 105:10210–10216
33. Murata K, Kaneko K, Steele WA, Kokai F, Takahashi K, Kasuya D, Hirahara K, Yudasaka M, Iijima S (2001) Porosity evaluation of intrinsic intraparticle nanopores of single wall carbon nanohorn. Nano Lett 1:197–199
34. Nalwa HS (ed) (2004) Encyclopedia of nanoscience and nanotechnology, vol 1–10. American Scientific Publishers, Cambridge
35. Ohba T, Murata K, Kaneko K, Steele WA, Kokai F, Takahashi K, Kasuya K, Yudasaka M, Iijima S (2001) N_2 adsorption in an internal nanopore space of single-walled carbon nanohorn: GCMC simulation and experiment. Nano Lett 1:371–373
36. Parr RG, Yang W (1989) Density functional theory of atoms and molecules. Oxford University Press, Oxford
37. Pathria RK (1972) Statistical mechanics. Pergamon, Oxford
38. Pekala K, Pekala M (1995) Low temperature transport properties of nanocrystalline Cu, Fe and Ni. NanoStructured Mater 6:819–822
39. Popov VN (2002) Low-temperature specific heat of nanotube systems. Phys Rev B 66:153408-1–153408-4
40. Popov VN (2004) Carbon nanotubes: properties and applications. Math Sci Eng R 43:61–102
41. Prylutskyy YI, Durov SS, Ogloblya OV, Buzaneva EV, Scharff P (2000) Molecular dynamics simulation of mechanical, vibrational and electronic properties of carbon nanotubes. Comp Math Sci 17:352–355
42. Rafii-Tabar H (2000) Modeling the nano-scale phenomena in condensed matter physics via computer-based numerical simulations. Phys Rep 325:239–310
43. Rafii-Tabar H (2004) Computational modeling of thermo-mechanical and transport properties of carbon nanotubes. Phys Rep 390:235–452
44. Rafii-Tabar H (2008) Computational physics of carbon nanotubes. Cambridge University Press, Cambridge
45. Rao CNR, Satishkumar BC, Govindaraj A, Nath M (2001) Nanotubes. Chem Phys Chem 2:78–105
46. Rapaport DC (1995) The art of molecular dynamics simulation. Cambridge University Press, Cambridge
47. Reich S, Thomsen C, Ordejón P (2002) Elastic properties of carbon nanotubes under hydrostatic pressure. Phys Rev B 65:153407-1–153407-4
48. Rieth M, Schommers W (ed) (2007) Handbook of theoretical and computational nanotechnology, vol 1–10. American Scientific Publishers, Cambridge
49. Roco MC, Bainbridge WS (2002) Convergent technologies for improving human performance. NSF/DOC-sponsored Report. WTEC Inc, Arlington
50. Saito Y, Yoshikawa T, Bandow S, Tomita M, Hayashi T (1993) Interlayer spacings in carbon nanotubes. Phys Rev B 48:1907–1909
51. Schwarz JA, Contescu CI, Putyera K (ed) (2004) Dekker encyclopedia of nanoscience and nanotechnology, vol 1–5. Marcel Dekker Inc, New York
52. Siber A (2002) Adsorption of He atoms in external grooves of single-walled carbon nanotube bundles. Phys Rev B 66:205406-1–205406-6
53. Simonyan VV, Diep P, Johnson JK (1999) Molecular simulation of hydrogen adsorption in charged single-walled carbon nanotubes. J Chem Phys 111:9778–9783
54. Simonyan VV, Johnson JK, Kuznetsova A, Yates Jr JT (2001) Molecular simulation of xenon adsorption on single-walled carbon nanotubes. J Chem Phys 114:4180–4185
55. Smith BW, Monthoux M, Luzzi DE (1998) Encapsulated C_{60} in carbon nanotubes. Nature 396:323–324
56. Stevens RMD, Frederick NA, Smith BL, Morse DE, Stucky GD, Hansma PK (2000) Carbon nanotubes as probes for atomic force microscopy. Nanotechnology 11:1–5
57. Stojkovic D, Zhang P, Crespi VH (2001) Smallest nanotube: breaking the symmetry of sp^3 bonds in tubular geometries. Phys Rev Lett 87:125502–125505
58. Supple S, Quirke N (2003) Rapid imbibition of fluids in carbon nanotubes. Phys Rev Lett 90:214501-1–214501-4
59. Sutton AP, Pethica JB, Rafii-Tabar H, Nieminen JA (1994) Mechanical properties of metals at nanometre scale. In: Pettifor DG, Cottrell AH (eds) Electron theory in alloy design. Institute of Materials, London, pp 191–233
60. Talapatra S, Zambano AZ, Weber SE, Migone AD (2000) Gases do not adsorb on the interstitial channels of closed-ended single-walled carbon nanotube bundles. Phys Rev Lett 85:138–141
61. Tuzun RE, Noid DW, Sumpter BG, Merkle RC (1996) Dynamics of fluid flow inside carbon nanotubes. Nanotechnology 7:241–246
62. Wang Q, Johnson JK (1999) Molecular simulation of hydrogen adsorption in single-walled carbon nanotubes and idealised carbon slit pore. J Chem Phys 110:577–586
63. Wang Q, Johnson JK (1999) Optimisation of carbon nanotube arrays for hydrogen adsorption. J Phys Chem B 103:4809–4813

64. Williams KA, Eklund PC (2000) Monte Carlo simulation of H_2 physisorption in finite-diameter carbon nanotube ropes. Chem Phys Lett 320:352–358
65. Yakobson BI, Brabec CJ, Bernholc J (1996) Nanomechanics of carbon tubes: instabilities beyond linear response. Phys Rev Lett 76:2511–2514
66. Zhao J, Buldum A, Han J, Lu JP (2002) Gas molecule adsorption in carbon nanotubes and nanotube bundles. Nanotechnology 13:195–200
67. Zhou LG, Shi SQ (2002) Molecular dynamic simulations on tensile mechanical properties of single-walled carbon nanotubes with and without hydrogen storage. Comp Math Sci 23:166–174

Catastrophe Theory

WERNER SANNS
University of Applied Sciences, Darmstadt, Germany

Article Outline

Glossary

Singularity Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable map defined in some open neighborhood of the point $p \in \mathbb{R}^n$ and $J_p f$ its Jacobian matrix at p, consisting of the partial derivatives of all components of f with respect to all variables. f is called singular in p if rank $J_p f < \min\{n, m\}$. If rank $J_p f = \min\{m, n\}$, then f is called regular in p. For $m = 1$ (we call the map $f\colon \mathbb{R}^n \to \mathbb{R}$ a differentiable function) the definition implies: A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is singular in $p \in \mathbb{R}^n$, if grad $f(p) = 0$. The point p, where the function is singular, is called a singularity of the function. Often the name "singularity" is used for the function itself if it is singular at a point p. A point where a function is singular is also called critical point. A critical point p of a function f is called a degenerated critical point if the Hessian (a quadratic matrix containing the second order partial derivatives) is singular at p; that means its determinant is zero at p.

Diffeomorphism A diffeomorphism is a bijective differentiable map between open sets of $\mathbb{R}^n$ whose inverse is differentiable, too.

Map germ Two continuous maps $f\colon U \to \mathbb{R}^k$ and $g\colon V \to \mathbb{R}^k$, defined on neighborhoods U and V of $p \in \mathbb{R}^n$, are called equivalent as germs at p if there exists a neighborhood $W \subset U \cap V$ on which both coincide. Maps or functions, respectively, that are equivalent as germs can be considered to be equal regarding local features. The equivalence classes of this equivalence relation are called germs. The set of all germs of differentiable maps $\mathbb{R}^n \to \mathbb{R}^k$ at a point p is named $\varepsilon_p(n, k)$. If p is the origin of $\mathbb{R}^n$, one simply writes $\varepsilon(n, k)$ instead of $\varepsilon_0(n, k)$. Further, if $k = 1$, we write $\varepsilon(n)$ instead of $\varepsilon(n, 1)$ and speak of function germs (also simplified as "germs") at the origin of $\mathbb{R}^n$. $\varepsilon(n, k)$ is a vector space and $\varepsilon(n)$ is an algebra, that is, a vector space with a structure of a ring. The ring $\varepsilon(n)$ contains a unique maximal ideal $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. The ideal $\mu(n)$ is generated by the germs of the coordinate functions $x_1, \ldots, x_n$. We use the form $\mu(n) = \langle x_1, \ldots, x_n \rangle \varepsilon(n)$ to emphasize, that this ideal is generated over the ring $\varepsilon(n)$. So a function germ in $\mu(n)$ is of the form $\sum_{i=1}^{n} a_i(x) \cdot x_i$, with certain function germs $a_i(x) \in \varepsilon(n)$.

r-Equivalence Two function germs $f, g \in \varepsilon(n)$ are called r-equivalent if there exists a germ of a local diffeomorphism $h\colon \mathbb{R}^n \to \mathbb{R}^n$ at the origin, such that $g = f \circ h$.

Unfolding An unfolding of a differentiable function germ $f \in \mu(n)$ is a germ $F \in \mu(n + r)$ with $F|\mathbb{R}^n = f$ (here | means the restriction. The number r is called an unfolding dimension of f. An unfolding F of a germ f is called universal if every other unfolding of f can be received by suitable coordinate transformations, "morphisms of unfoldings" from F, and the number of unfolding parameters of F is minimal (see "codimension").

Unfolding morphism Suppose $f \in \varepsilon(n)$ and $F\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ and $G\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ be unfoldings of f. A right-morphism from F to G, also called unfolding morphism, is a pair (Φ, α), with $\Phi \in \varepsilon(n + k, n + r)$ and $\alpha \in \mu(k)$, such that:

1. $\Phi|\mathbb{R}^n = \mathrm{id}(\mathbb{R}^n)$, that is $\Phi(x, 0) = (x, 0)$,
2. If $\Phi = (\phi, \psi)$, with $\phi \in \varepsilon(n + k, n)$, $\psi \in \varepsilon(n + k, r)$, then $\psi \in \varepsilon(k, r)$,
3. For all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^k$ we get $F(x, u) = G(\Phi(x, u)) + \alpha(u)$.

Catastrophe A catastrophe is a universal unfolding of a singular function germ. The singular function germs are called organization centers of the catastrophes.

Codimension The codimension of a singularity f is given by $\operatorname{codim}(f) = \dim_{\mathbb{R}} \mu(n)/\langle \partial_x f \rangle$ (quotient space). Here $\langle \partial_x f \rangle$ is the Jacobian ideal generated by the partial derivatives of f and $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. The codimension of a singularity gives of the minimal number of unfolding parameters needed for the universal unfolding of the singularity.

Potential function Let $f: \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable map (that is, a differentiable vector field). If there exists a function $\varphi: \mathbb{R}^n \to \mathbb{R}$ with the property that $\operatorname{grad} \varphi = f$, then f is called a gradient vector field and φ is called a potential function of f.

Definition of the Subject

Catastrophe theory is concerned with the mathematical modeling of sudden changes – so called "catastrophes" – in the behavior of natural systems, which can appear as a consequence of continuous changes of the system parameters. While in common speech the word catastrophe has a negative connotation, in mathematics it is neutral.

You can approach catastrophe theory from the point of view of differentiable maps or from the point of view of dynamical systems, that is, differential equations. We use the first case, where the theory is developed in the mathematical language of maps and functions (maps with range $\mathbb{R}$). We are interested in those points of the domain of differentiable functions where their gradient vanishes. Such points are called the "singularities" of the differentiable functions.

Assume that a system's behavior can be described by a potential function (see Sect. "Glossary"). Then the singularities of this function characterize the equilibrium points of the system under consideration. Catastrophe theory tries to describe the behavior of systems by local properties of corresponding potentials. We are interested in local phenomena and want to find out the qualitative behavior of the system independent of its size.

An important step is the classification of catastrophe potentials that occur in different situations. Can we find any common properties and unifying categories for these catastrophe potentials? It seems that it might be impossible to establish any reasonable criteria out of the many different natural processes and their possible catastrophes. One of the merits of catastrophe theory is the mathematical classification of simple catastrophes where the model does not depend on too many parameters.

Classification is only one of the mathematical aspects of catastrophe theory. Another is stability. The stable states of natural systems are the ones that we can observe over a longer period of time. But the stable states of a system, which can be described by potential functions and their singularities, can become unstable if the potentials are changed by perturbations. So stability problems in nature lead to mathematical questions concerning the stability of the potential functions.

Many mathematical questions arise in catastrophe theory, but there are also other kinds of interesting problems, for example, historical themes, didactical questions and even social or political ones. How did catastrophe theory come up? How can we teach catastrophe theory to the students at our universities? How can we make it understandable to non-mathematicians? What can people learn from this kind of mathematics? What are its consequences or insights for our lives?

Let us first have a short look at mathematical history. When students begin to study a new mathematical field, it is always helpful to learn about its origin in order to get a good historical background and to get an overview of the dependencies of inner-mathematical themes. Catastrophe theory can be thought of as a link between classical analysis, dynamical systems, differential topology (including singularity theory), modern bifurcation theory and the theory of complex systems. It was founded by the French mathematician René Thom (1923–2002) in the sixties of the last century. The name 'catastrophe theory' is used for a combination of singularity theory and its applications. In the year 1972 Thom's famous book "Stabilité structurelle et morphogénèse" appeared. Thom's predecessors include Marston Morse, who developed his theory of the singularities of differentiable functions (Morse theory) in the thirties, Hassler Whitney, who extended this theory in the fifties to singularities of differentiable maps of the plane into the plane, and John Mather (1960), who introduced algebra, especially the theory of ideals of differentiable functions, into singularity theory. Successors to Thom include Christopher Zeeman (applications of catastrophe theory to physical systems), Martin Golubitsky (bifurcation theory), John Guckenheimer (caustics and bifurcation theory), David Schaeffer (shock waves) and Gordon Wassermann (stability of unfoldings). From the point of view of dynamical systems, the forerunners of Thom are Jules Henry Poincaré (1854–1912) who looked at stability problems of dynamical systems, especially the problems of celestial mechanics, Andrei Nikolaevic **K**olmogorov, Vladimir I. **A**rnold and Jürgen **M**oser (KAM), who influenced catastrophe theory with their works on stability problems (KAM-theorem).

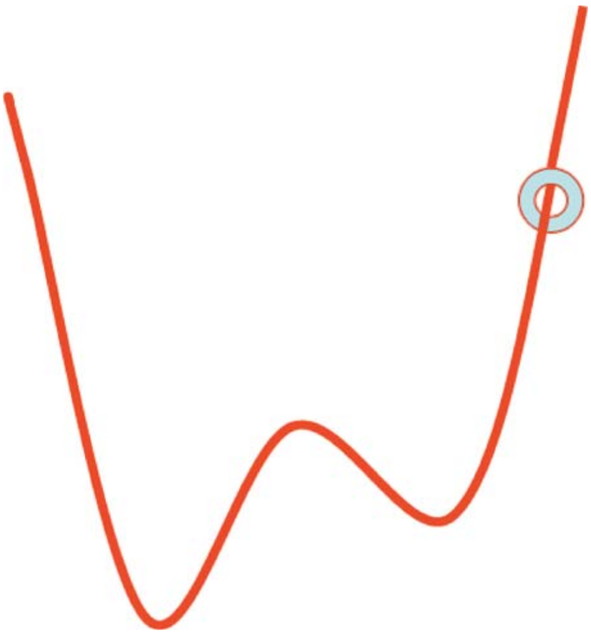
From the didactical point of view, there are two main positions for courses in catastrophe theory at university level: Trying to teach the theory as a perfect axiomatic system consisting of exact definitions, theorems and proofs or trying to teach mathematics as it can be developed from historical or from natural problems (see [9]). In my opinion the latter approach has a more lasting effect, so there is a need to think about simple examples that lead to the fundamental ideas of catastrophe theory. These examples may serve to develop the theory in a way that starts with intuitive ideas and goes forward with increasing mathematical precision.

When students are becoming acquainted with catastrophe theory, a useful learning tool is the insight that continuous changes in influencing system parameters can lead to catastrophes in the system behavior. This phenomenon occurs in many systems. Think of the climate on planet earth, think of conflicts between people or states. Thus, they learn that catastrophe theory is not only a job for mathematical specialists, but also a matter of importance for leading politicians and persons responsible for guiding the economy.

Introduction

Since catastrophe theory is concerned with differentiable functions and their singularities, it is a good idea to start with the following simple experiment: Take a piece of wire, about half a meter in length, and form it as a 'curve' as shown in Fig. 1. Take a ring or screw-nut and thread the wire through it so that the object can move along the wire easily.

The wire represents the model of the graph of a potential function, which determines the reactions of a system by its gradient. This means the ring's horizontal position represents the actual value of a system parameter x that characterizes the state of the system, for example, x might be its position, rotation angle, temperature, and so on. This parameter x is also called the state parameter. The form of the function graph determines the behavior of the system, which tries to find a state corresponding to a minimum of the function (a zero of the gradient of f) or to stay in such a 'stable' position. (Move the ring to any position by hand and then let it move freely.) Maxima of the potential function are equilibrium points too, but they are not stable. A small disturbance makes the system leave such an equilibrium. But if the ring is in a local minimum, you may disturb its position by slightly removing it from that position. If this disturbance is not too big, the system will return to its original position at the minimum of the function when allowed to move freely. The minimum is a stable point of the function, in the sense that small disturbances of the system are corrected by the system's behavior, which tries to return to the stable point. Observe that you may disturb the system by changing its position by hand (the position of the ring) or by changing the form of the potential function (the form of the wire). If the form of the potential function is changed, this corresponds to the change of one or more parameters in the defining equation for the potential. These parameters are called external parameters. They are usually influenced by the experimenter. If the form of the potential (wire) is changed appropriately (for example, if you pull upward at one side of the wire), the ring can leave the first disappearing minimum and fall into the second stable position. This is the way catastrophes happen: changes of parameters influence the potentials and thus may cause the system to suddenly leave a stable position.



Catastrophe Theory, Figure 1
Ring on the wire

Some typical questions of catastrophe theory are:

- How can one find a potential function that describes the system under consideration and its catastrophic jumps?
- What does this potential look like locally? How can we describe the changes in the potential's parameters?
- Can we classify the potentials into simple categories?
- What insights can such a classification give us?

Now we want to consider two simple examples of systems where such sudden changes become observable. They are easy to describe and lead to the simplest forms of catastrophes which Thom listed in the late nineteen-sixties. The first is a simple physical model of an eccentric cylinder on an inclined plane. The second example is a model for the formation of a traffic jam.

Example 1: The Eccentric Cylinder on the Inclined Plane

We perform some experiments with a cylinder or a wide wheel on an inclined plane whose center of gravity is eccentric with respect to its axes (Fig. 2).

You can build such a cylinder from wood or metal by taking two disks of about 10 cm in radius. Connect them with some sticks of about 5 cm in length. The eccentric center of gravity is generated by fixing a heavy piece of metal between the two discs at some distance from the axes of rotation. Draw some lines on the disks, where you can read or estimate the angle of rotation. (Hint: It could be interesting for students to experimentally determine the center of gravity of the construction.)

If mass and radius are fixed, there are two parameters which determine the system: The position of the cylinder can be specified by the angle x against a zero level (that is, against the horizontal plane; Fig. 4). This angle is the "inner" variable of the system, which means that it is the "answer" of the cylinder to the experimenter's changes of the inclination u of the plane, which is the "outer" parameter. In this example you should determine experimentally the equilibrium positions of the cylinder, that is, the angle of rotation where the cylinder stays at rest on the inclined plane (after releasing it cautiously and after letting the system level off a short time). We are interested in the stable positions, that is, the positions where the cylinder returns after slight disturbances (slightly nudging it or slightly changing the inclination of the plane) without rolling down the whole plane. The search for the equilibrium position x varies with the inclination angle u. You can make a table and a graphic showing the dependence of x and u.

The next step is to find an analytical approach. Refer to Fig. 3: Point P is the supporting point of the cylinder on the plane. Point A marks the starting position of the center of gravity. When the wheel moves and the distance between M and A is big enough, this center can come into two equilibrium positions S_1 and S_2, which lie vertically over P.

Now, what does the potential function look like? The variables are:

***R*:** radius of the discs,

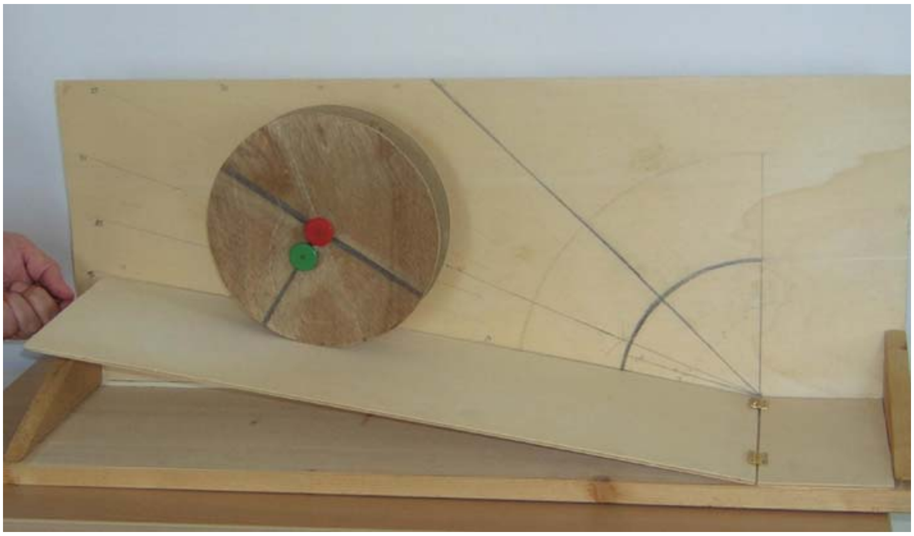
***u*:** angle of the inclined plane against the horizontal plane measured in radians,

***x*:** angle of rotation of the cylinder measured against the horizontal plane in radians,

***r*:** distance between the center of gravity S of the cylinder from the axes marked by point M,

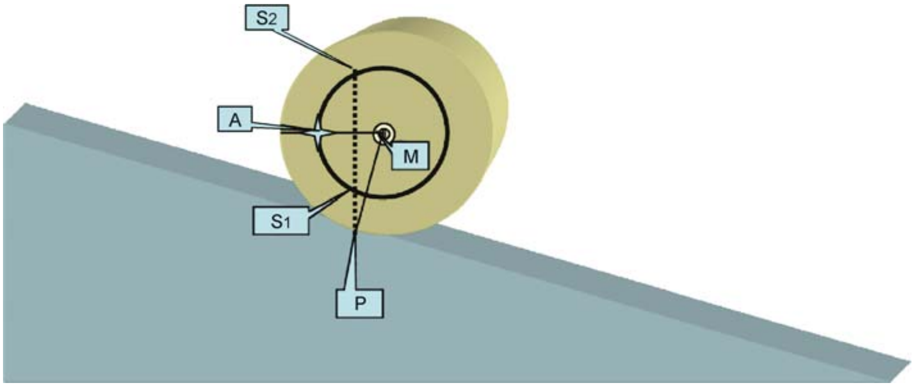
x_0 (resp. u_0): angles where the cylinder suddenly starts to roll down the plane.

Refer to Fig. 4. If the wheel is rolled upward by hand from the lower position on the right to the upper position, the center of gravity moves upward with the whole wheel, or

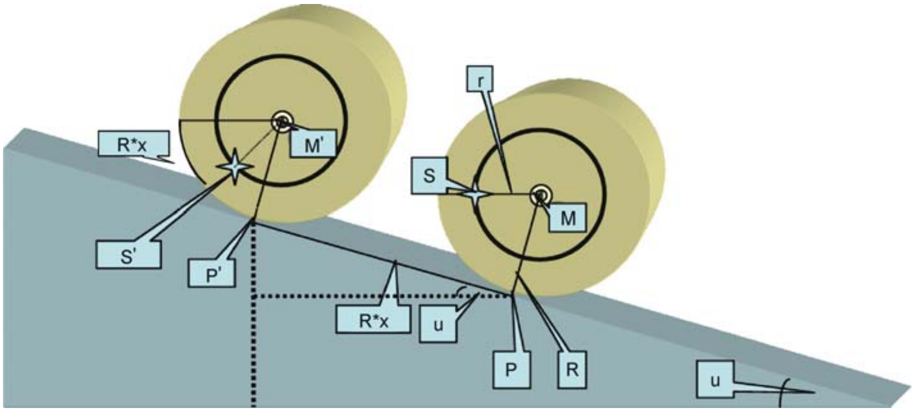


Catastrophe Theory, Figure 2

The eccentric cylinder on the inclined plane. The inclination of the plane can be adjusted by hand. The position of the disc center and the center of gravity are marked by colored buttons



Catastrophe Theory, Figure 3
Two equilibrium positions (S_1 and S_2) are possible with this configuration. If the point A lies "too close" to M, there is no equilibrium position possible at all when releasing the wheel, since the center of gravity can never lie vertically over P



Catastrophe Theory, Figure 4
The movement of S is a composite of the movement of the cylinder represented by the movement of point M and the rotation movement with angle *x*

as the midpoint M or the point P. But point S also moves downward by rotation of the wheel with an amount of $r \cdot \sin(x)$.

The height gained is

$$h = R \cdot x \cdot \sin(u) - r \cdot \sin(x) . \qquad (1)$$

From physics one knows the formula for the potential energy E_{pot} of a point mass. It is the product $m \cdot g \cdot h$, where $m =$ mass, $g =$ gravitational constant, $h =$ height, where the height is to be expressed by the parameters x and u and the given data of the cylinder.

To a fixed angle u_1 of the plane, by formula (1), there corresponds the following potential function:

$$f(x) = E_{\text{pot}} = m \cdot g \cdot h = m \cdot g \cdot (R \cdot x \cdot \sin(u_1) - r \cdot \sin(x)) .$$

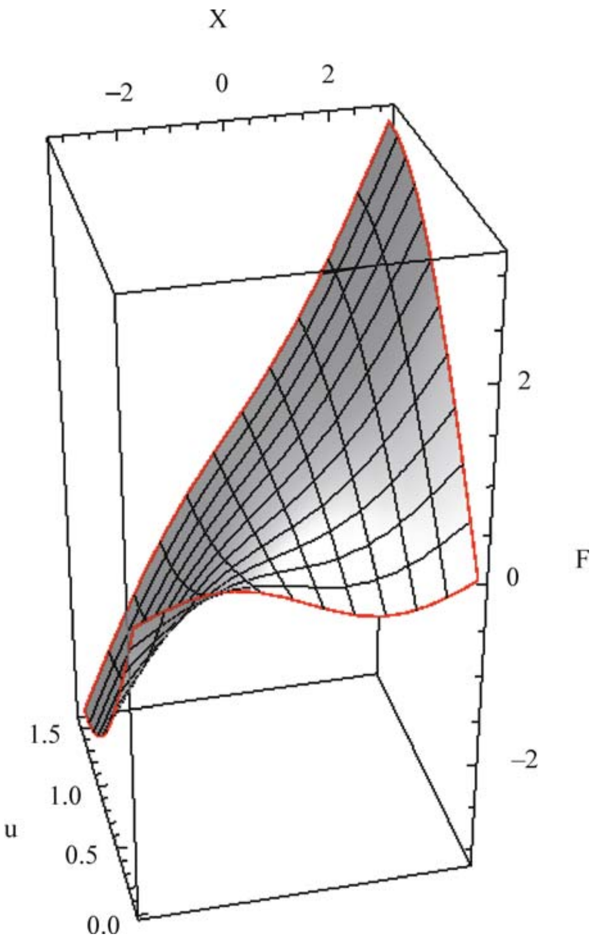
If the inclination of the plane is variable, the angle u of inclination serves as an outer parameter and we obtain a whole family of potential functions as a function F in two variables x and u:

$$F(x,u) = F_u(x) = m \cdot g \cdot (R \cdot x \cdot \sin(u) - r \cdot \sin(x)) .$$

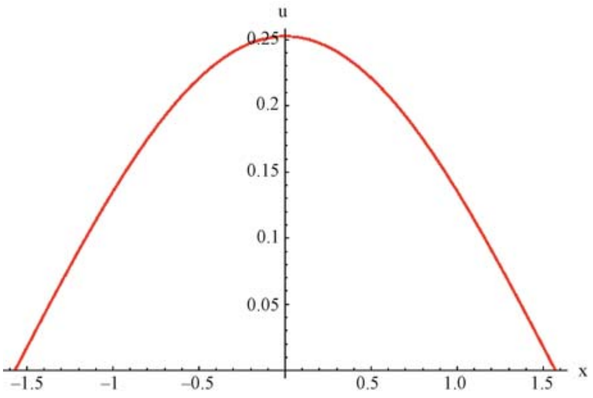
Thus, the behavior of the eccentric cylinder is described by a family of potential functions. The extrema of the single potentials characterize the equilibria of the system. The saddle point characterizes the place where the cylinder suddenly begins to roll down due to the slightest disturbance, that is, where the catastrophe begins (see Fig. 5).

Example. The data are $m = 1\,\text{kg}$, $g = 10\,\text{m/sec}^2$, $R = 0.1\,\text{m}$, $r = 0.025\,\text{m}$.

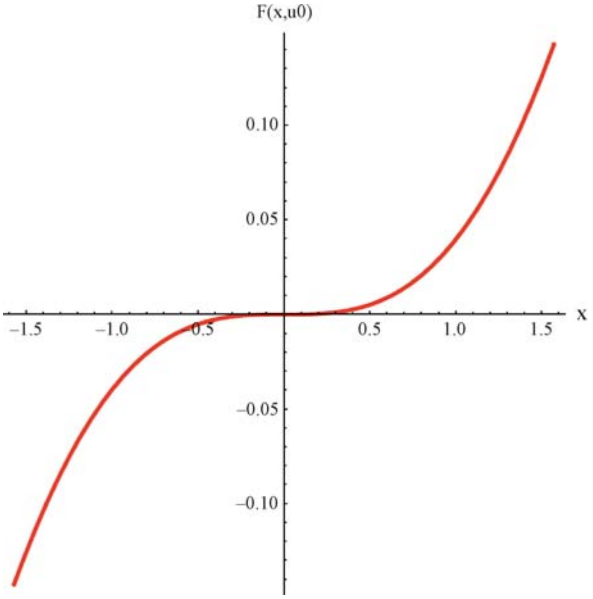
For constant angle of inclination u, each of the functions $f(x) = F(x, u = \text{const})$ is a section of the function F.



Catastrophe Theory, Figure 5
Graph of $F(x, u)$ for a concrete cylinder



Catastrophe Theory, Figure 6
Curve in x–u-space, for which $F(x, u)$ has local extrema



Catastrophe Theory, Figure 7
Graph of the section $F(x, u_0)$

To determine the minima of each of the section graphs, we calculate the zeros of the partial derivative of F with respect to the inner variable x. In our particular example above, the result is $\partial_x F = -0.25\cos(x) + \sin(u)$. The solution for $\partial_x F = 0$ is $u = \arcsin(0.25\cos(x))$ (see Fig. 6).

Calculating the point in x–u-space at which both partial derivatives $\partial_x F$ and $\partial_{xx} F$ vanish (saddle point of the section) and using only positive values of angle u gives the point $(x_0, u_0) = (0, 0.25268)$. In general: $x_0 = 0$ and $u_0 = \arcsin(r/R)$. In our example this means the catastrophe begins at an inclination angle of $u_0 = 0.25268$ (about 14.5 degrees). At that angle, sections of F with $u > u_0$ do not possess any minimum where the system could stably rest. Making u bigger than u_0 means that the cylinder must leave its stable position and suddenly roll down the plane. The point u_0 in the u-parameter space is called catastrophe point.

What is the shape of the section of F at u_0? Since in our example

$$f(x) = F(x, u_0) = 0.25x - 0.25\sin(x) ,$$

the graph appears as shown in Fig. 7.

The point $x = 0$ is a degenerated critical point of $f(x)$, since $f'(x) = f''(x) = 0$. If we expand f into its Taylor polynomial $t(x)$ around that point we get

$$t(x) = 0.0416667x^3 - 0.00208333x^5 + 0.0000496032x^7 \\ -6.88933 \times 10^{-7}x^9 + O[x]^{11} .$$

Catastrophe theory tries to simplify the Taylor polynomials of the potentials locally about their degenerated critical points. This simplification is done by coordinate transformations, which do not change the structure of the singularities. These transformations are called diffeomorphisms (see Sect. "Glossary"). A map $h\colon U \to V$, with U and V open in $\mathbb{R}^n$, is a local diffeomorphism if the Jacobian determinant $\det(J_p h) \neq 0$ at all points p in U. Two functions f and g, both defined in some neighborhood of the origin of $\mathbb{R}^n$, are called r-equivalent if there exists a local diffeomorphism $h\colon U \to V$ at the origin, such that $g = f \circ h$.

The Taylor polynomial $t(x)$ in our example starts with a term of 3rd order in x. Thus, there exists a diffeomorphism (a non-singular coordinate transformation), such that in the new coordinates t is of the form x^3. To see this, we write $t(x) = x^3 \cdot a(x)$, with $a(0) \neq 0$. We define $h(x) = x \cdot a(x)^{1/3}$. Note that the map h is a local diffeomorphism near the origin since $h'(0) = a(0)^{1/3} \neq 0$. (Hint: Modern computer algebra systems, such as Mathematica, are able to expand expressions like $a(x)^{1/3}$ and calculate the derivative of h easily for the concrete example.)

Thus for the eccentric cylinder the intersection of the graph of the potential $F(x, u)$ with the plane $u = u_0$ is the graph of a function, whose Taylor series at its critical point $x = 0$ is $t = x^3 \circ h$, in other words: t is r-equivalent to x^3.

Before we continue with this example and its relation to catastrophe theory, let us introduce another example which leads to a potential of the form x^4. Then we will have two examples which lead to the two simplest catastrophes in René Thom's famous list of the seven elementary catastrophes.

Example 2: The Formation of Traffic Jam

In the first example we constructed a family F of potential functions directly from the geometric properties of the model of the eccentric cylinder. We then found its critical points by calculating the derivatives of F. In this second example the method is slightly different. We start with a partial differential equation which occurs in traffic flow modeling. The solution surface of the initial value problem will be regarded as the surface of zeros of the derivative of a potential family. From this, we then calculate the family of potential functions. Finally, we will see that the Taylor series of a special member of this family, the one which belongs to its degenerated critical point, is equivalent to x^4.

Note: When modeling a traffic jam we use the names of variables that are common in literature about this subject. Later we will rename the variables into the ones commonly used in catastrophe theory.

Let $u(x, t)$ be the (continuous) density of the traffic at a street point x at time t and let $a(u)$ be the derivative of the traffic flow $f(u)$. The mathematical modeling of simple traffic problems leads to the well known traffic flow equation:

$$\frac{\partial u(x,t)}{\partial t} + a(u)\frac{\partial u(x,t)}{\partial x} = 0\,.$$

We must solve the Cauchy problem for this quasilinear partial differential equation of first order with given initial density $u_0 = u(x, 0)$ by the method of characteristics. The solution is constant along the (base) characteristics, which are straight lines in xt-plane with slope $a(u_0(y))$ against the t-axes, if y is a point on the x-axes, where the characteristic starts at $t = 0$. For simplicity we write "a" instead of $a(u_0(y))$. Thus the solution of the initial value problem is $u(x, t) = u_0(x - a \cdot t)$ and the characteristic surface $S = \{(x, t, u) | u - u_0(x - a \cdot t) = 0\}$ is a surface in xtu-space. Under certain circumstances S may not lie uniquely above the xt-plane but may be folded as is shown in Fig. 8. There is another curve to be seen on the surface: The border curve of the fold. Its projection onto the xt-plane is a curve with a cusp, specifying the position and the beginning time of the traffic jam.

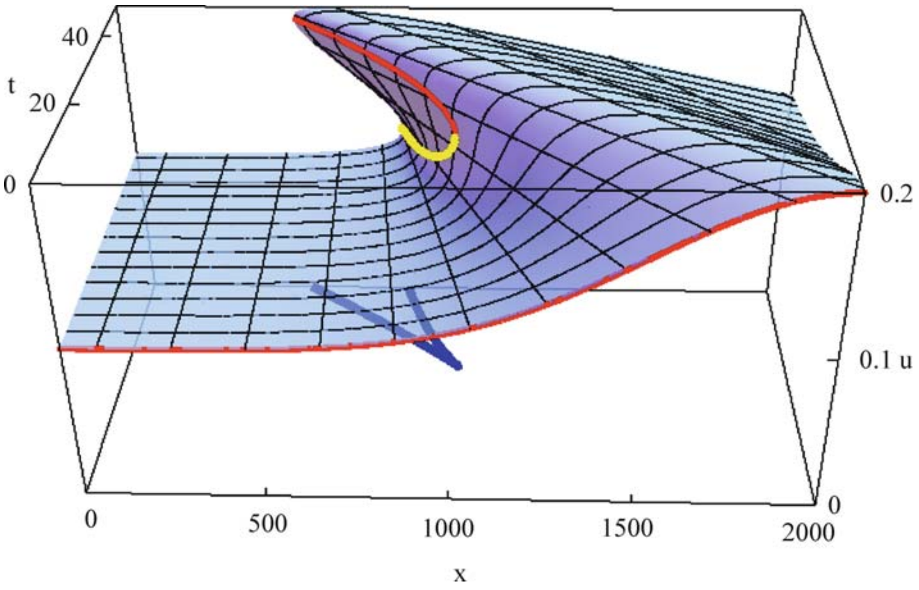
The equations used in the example, in addition to the traffic flow equation, can be deduced in traffic jam modeling from a simple parabolic model for a flow-density relation (see [5]). The constants in the following equations result from considerations of maximal possible density and an assumed maximal allowed velocity on the street. In our example a model of a road of 2 km in length and maximal traffic velocity of 100 km/h is chosen. Maximal traffic density is $u_{\max} = 0.2$. The equations are $f(u) = 27.78u - 138.89u^2$ and $u(x, 0) = u_0(x) = 0.1 + 0.1 \cdot \operatorname{Exp}(-((x - 2000)/700)^2)$. The latter was constructed to simulate a slight increase of the initial density u_0 along the street at time $t = 0$. The graph of u_0 can be seen as the front border of the surface.

To determine the cusp curve shown in Fig. 8, consider the following parameterization Φ of the surface S and the projection map $\pi\colon S \to \mathbb{R}^2$:

$$\begin{aligned}\Phi\colon \mathbb{R} \times \mathbb{R}_{\geq 0} &\to S \subset \mathbb{R}^3 \\ (x, t) &\mapsto (x + a(u_0(x)) \cdot t, t, u_0(x))\,.\end{aligned}$$

The Jacobian matrix of $\pi \circ \Phi$ is

$$\begin{aligned}J(\pi \circ \Phi)(x,t) &= \begin{pmatrix} \dfrac{\partial(\pi\circ\Phi)_1}{\partial x} & \dfrac{\partial(\pi\circ\Phi)_1}{\partial t} \\ \dfrac{\partial(\pi\circ\Phi)_2}{\partial x} & \dfrac{\partial(\pi\circ\Phi)_2}{\partial t} \end{pmatrix} \\ &= \begin{pmatrix} 1 + a(u_0(x))' \cdot t & a(u_0(x)) \\ 0 & 1 \end{pmatrix}.\end{aligned}$$



Catastrophe Theory, Figure 8
Characteristic surface *S*, initial density, fold curve and cusp curve

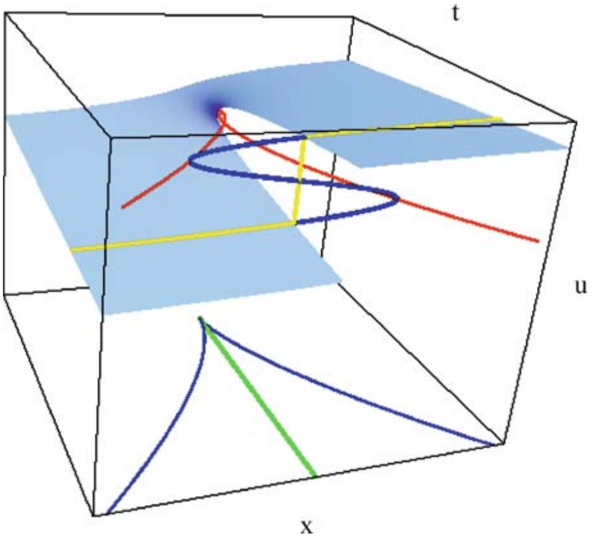
It is singular (that is, its determinate vanishes) if $1 + (a(u_0(x)))' \cdot t = 0$. From this we find the curve $c\colon \mathbb{R} \to \mathbb{R}^2$

$$c(x) = \left(x, -\frac{1}{a'(u_0(x)) \cdot u_0'(x)}\right)$$

which is the pre-image of the cusp curve and the cusp curve in the Fig. 8 is the image of c under $\pi \circ \Phi$.

Whether the surface folds or not depends on both the initial density u_0 and the properties of the flow function $f(u)$. If such folding happens, it would mean that the density would have three values at a given point in xt-space, which is a physical impossibility. Thus, in this case, no continuous solution of the flow equation can exist in the whole plane. "Within" the cusp curve, which is the boundary for the region where the folding happens, there exists a curve with its origin in the cusp point. Along this curve the surface must be "cut" so that a discontinuous solution occurs. The curve within the cusp region is called a shock curve and it is characterized by physical conditions (jump condition, entropy condition). The density makes a jump along the shock.

The cusp, together with the form of the folding of the surface, may give an association to catastrophe theory. One has to find a family of potential functions for this model, that is, functions $F_{(x,t)}(u)$, whose negative gradient, that is, its derivative by u (the inner parameter), describes the characteristic surface S by its zeroes. The family of potentials is given by



Catastrophe Theory, Figure 9
The solution surface is "cut" along the shock curve, running within the cusp region in the *xt*-plane

$$F(x,t,u) = -t\cdot\left(u \cdot a(u) - \int a(u)\mathrm{d}u\right) - \int_0^{x-a(u)\cdot t} u_0(x)\mathrm{d}x$$

(see [6]). Since

$$-\mathrm{grad}F_{x,t}(u) = -\frac{\partial F}{\partial u} = 0 \Leftrightarrow u - u_0(x - a(u) \cdot t) = 0$$

the connection of F and S is evident. One can show that a member f of the family F, the function $f(u) = F(\xi, \tau, u)$, is locally in a traffic jam formation point (ξ, τ) of the form $f(u) = u^4$ (after Taylor expansion and after suitable coordinate transformations, similar to the first example).

The complicated function terms of f and F thus can be replaced in qualitative investigations by simple polynomials. From the theorems of catastrophe theory, more properties of the models under investigation can be discovered.

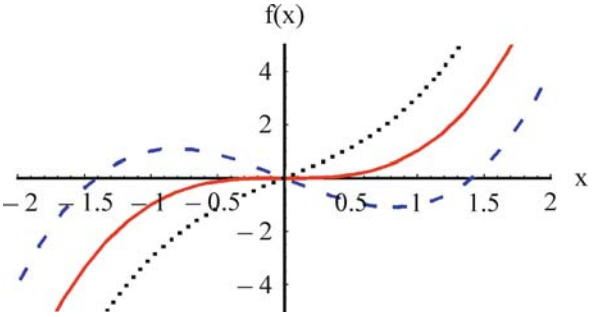
Now we want to use the customary notation of catastrophe theory. We shall rename the variables: u is the inner variables, which in catastrophe theory usually is x, and x, t (the outer parameters) get the names u and v respectively. Thus $F = F(x, u, v)$ is a potential family with one inner variable (the state variable x) and two outer variables (the control variables u and v).

Unfoldings

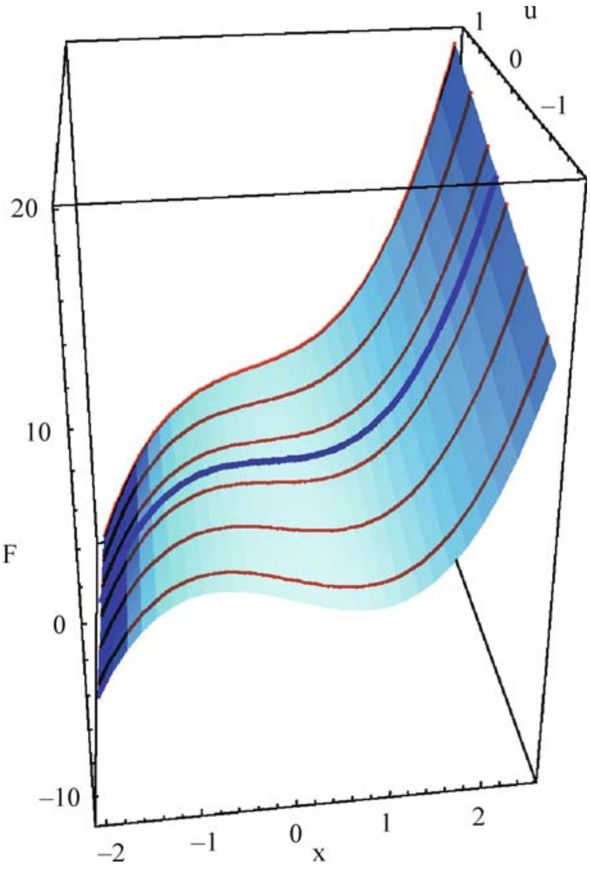
In the first examples above we found a family $F(x, u)$ of potential functions for the eccentric cylinder. For u fixed we found single members of that family. The member which belonged to the degenerated critical point of F, that is, to the point (x, u) where $\partial_x F = \partial_{xx} F = 0$, turned out to be equivalent to x^3. In the second example the corresponding member is equivalent to x^4 (after renaming the variables). These two singularities are members of potential families which also can be transformed into simple forms.

In order to learn how such families, in general, arise from a single function germ, let us look at the singularity $f(x) = x^3$. If we add to it a linear "disturbance" $u \cdot x$, where the factor (parameter) u may assume different values, we qualitatively have one of the function graphs in Fig. 10.

The solid curve in Fig. 10 represents the graph of $f(x) = x^3$. The dashed curve is the graph of $g(x) = x^3 - u \cdot x$; $u > 0$. The dotted line is $h(x) = x^3 + u \cdot x$; $u > 0$.



Catastrophe Theory, Figure 10
Perturbations of x^3

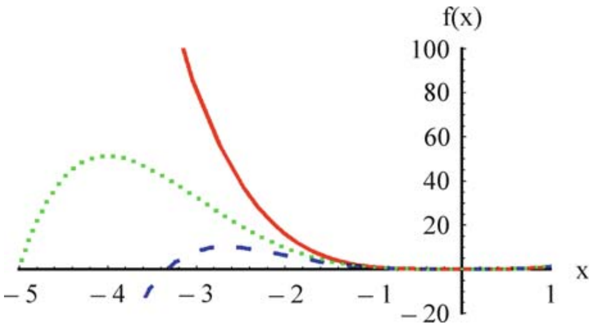


Catastrophe Theory, Figure 11
The function $f(x) = x^3$ embedded in a function family

Please note: While for positive parameter values, the disturbed function has no singularity, in the case of negative u-values a relative maximum and a relative minimum exist. The origin is a singular point only for $u = 0$, that is, for the undisturbed function f.

We can think of $f(x) = x^3$ as a member of a function family, which contains disturbances with linear terms.

The function $f(x) = x^3$ changes the type of singularity with the addition of an appropriate small disturbing function, as we have seen. Therefore f is called "structurally unstable". But we have also learned that $f(x) = x^3$ can be seen as a member of a whole family of functions $F(x, u) = x^3 + u \cdot x$, because $F(x, 0) = f(x)$. This family is stable in the sense that all kinds of singularities of the perturbed function are included by its members. This family is only one of the many possibilities to "unfold" the function germ $f(x) = x^3$ by adding disturbing terms. For example $F(x, u, v) = x^3 + u \cdot x^2 + v \cdot x$ would be another such unfolding (see Sect. "Glossary") of f. But this unfolding has more parameters than the minimum



Catastrophe Theory, Figure 12
The function *f*(*x*) = x4 + *u* · x5, *u* = 0 (*solid*), *u* = 0.2 (*dotted*) and *u* = 0.3 (*dashed*)

needed. The example of a traffic jam leads to an unfolding of $f(x) = x^4$ which is $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$.

How can we find unfoldings, which contain all possible disturbances (all types of singularities), that is, which are stable and moreover have a minimum number of parameters?

Examine the singularity x^4 (Fig. 12). It has a degenerated minimum at the origin. We disturb this function by a neighboring polynomial of higher degree, for example, by $u \cdot x^5$, where the parameter u with small values (here and in what follows "small" means small in amount), ensures that the perturbed function is "near" enough to x^4.

For $f(x) = x^4 + u \cdot x^5$, we find that the equation $f'(x) = 0$ has the solutions $x_1 = x_2 = x_3 = 0$, $x_4 = -4/(5u)$. To the threefold zero of the derivative, that is, the threefold degenerated minimum, another extremum is added which is arbitrarily far away from the origin for values of u that are small enough in amount due to the term $-4/(5u)$. This is so, because the term $-4/(5u)$ increases as u decreases. (This is shown in Fig. 12 with $u = 0$, $u = 0.2$ and $u = 0.3$). The type of the singularity at the origin thus is not influenced by neighboring polynomials of the form $u \cdot x^5$.

If we perturb the polynomial x^4 by a neighboring polynomial of lower degree, for example, $u \cdot x^3$, then for small amounts of u a new minimum is generated arbitrarily close to the existing singularity at the origin.

If disturbed by a linear term, the singularity at the origin can even be eliminated. The only real solution of $h'(x) = 0$, where $h(x) = x^4 + ux$, is $x = -(u/4)^{(1/3)}$.

Note: Only for $u = 0$, that is, for a vanishing disturbance, is h singular at the origin. For each $u \neq 0$ the linear disturbance ensures that no singularity at the origin occurs.

The function $f(x) = x^4$ is structurally unstable. To find a stable unfolding for the singularity $f(x) = x^4$, we must add terms of lower order. But we need not take into account all terms $x^4 + u \cdot x^3 + v \cdot x^2 + w \cdot x + k$, since the absolute term plays no role in the computation of singular points, and each polynomial of degree 4 can be written by a suitable change of coordinates without a cubic term. Similarly, a polynomial of third degree after a coordinate transformation can always be written without a quadratic term. The "Tschirnhaus-transformation": $x \mapsto x - a_1/n$, if a_1 is the coefficient of x^{n-1} in a polynomial $p(x) = x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n$ of degree n, leads to a polynomial $q(x) = x^n + b_1 x^{n-2} + \ldots + b_{n-2}x + b_{n-1}$ without a term of degree $n - 1$. Indeed, the expression $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$ is, as we shall see later, the "universal" unfolding of x^4. Here an unfolding F of a germ f is called universal, if every other unfolding of f can be received by suitable coordinate transformations "morphism of unfoldings" from F, and the number of unfolding parameters of F is minimal.

The minimal number of unfolding parameters needed for the universal unfolding F of the singularity f is called the codimension of f. It can be computed as follows:

$$\operatorname{codim}(f) = \dim_{\mathbb{R}} \mu(n)/\langle \partial_x f \rangle \,.$$

Here $\langle \partial_x f \rangle$ is the Jacobian ideal generated by the partial derivatives of f and $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. $\varepsilon(n)$ is the vector space of function germs at the origin of $\mathbb{R}^n$. The quotient $\mu(n)/\langle \partial_x f \rangle$ is the factor space.

What is the idea behind that factor space? Remember, that the mathematical description of a plane (a two dimensional object) in space (3 dimensional) needs only one equation, while the description of a line (1 dimensional) in space needs two equations. The number of equations that are needed for the description of these geometrical objects is determined by the difference between the dimensions of the surrounding space and the object in it. This number is called the codimension of the object. Thus the codimension of the plane in space is 1 and the codimension of the line in space is 2. From linear algebra it is known that the codimension of a subspace W of a finite dimensional vector space V can be computed as $\operatorname{codim} W = \dim V - \dim W$. It gives the number of linear independent vectors that are needed to complement a basis of W to get a basis of the whole space V. Another well known possibility for the definition is $\operatorname{codim} W = \dim V/W$. This works even for infinite dimensional vector spaces and agrees in finite dimensional cases with the previous definition. In our example $V = \mu(n)$ and W should be the set O of all functions germs, which are r-equivalent to f (O is the orbit of f under the action of the group of local diffeomorphisms preserving the origin of $\mathbb{R}^n$). But this is an infinite dimensional manifold, not a vector space. Instead, we can use its tangent space $T_f O$ in the point f, since spaces tangent to manifolds are vector spaces with the same dimension as the manifold itself and this tangent space is contained in the space tangent to $\mu(n)$ in an obvious way. It turns out, that the tangent space $T_f O$ is $\langle \partial_x f \rangle$. The space tangent to $\mu(n)$ agrees with $\mu(n)$, since it is a vector space. The details for the computations together with some examples can be found in the excellent article by M. Golubitsky (1978).

The Seven Elementary Catastrophes

Elementary catastrophe theory works with families (unfoldings) of potential functions

$$F\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \quad (\text{with } k \leq 4)$$
$$(x, u) \mapsto F(x, u) \, .$$

$\mathbb{R}^n$ is called the state space, $\mathbb{R}^k$ is called the parameter space or control space. Accordingly $x = (x_1, \ldots, x_n)$ is called a state variable or endogenous variable and $u = (u_1, \ldots, u_k)$ is called a control variable (exogenous variable, parameter).

We are interested in the singularities of F with respect to x and the dependence of F with respect to the control variables u. We regard F as the unfolding of a function germ f and want to classify F.

In analogy to the r-equivalence of function germs, we call two unfoldings F and G of the same function germ r-equivalent as unfoldings if there exists an unfolding-morphism (see Sect. "Glossary") from F to G. An unfolding F of f is called versal (or stable), if each other unfolding of f is right-equivalent as unfolding to F, that is, if there exists a right-morphism between the unfoldings. A versal unfolding with minimal unfolding dimensions is called universal unfolding of f.

Thus we have the following theorem (all proofs of the theorems cited in this article can be found in [1,2] or [12].)

Theorem 1 (Theorem on the existence of a universal unfolding) *A singularity f has a universal unfolding iff codim$(f) = k < \infty$.*

Examples (see also the following theorem with its examples):

$$\text{If } \; f(x) = x^3 \quad \text{it follows}$$
$$F(x, u, v) = x^3 + ux^2 + vx \quad \text{is versal,}$$
$$F(x, u) = x^3 + ux \quad \text{is universal,}$$
$$F(x, u) = x^3 + ux^2 \quad \text{is not versal .}$$

The following theorem states how one can find a universal unfolding of a function germ:

Theorem 2 (Theorem on the normal form of universal unfoldings) *Let $f \in \mu(n)$ be a singularity with codim$(f) = k < \infty$. Let $u = (u_1, \ldots, u_k)$ be the parameters of the unfolding. Let $b_i(x)$, $i = 1, \ldots, k$, be the elements of $\mu(n)$, whose cosets modulo $\langle \partial_x f \rangle$ generate the vector space $\mu(n)/\langle \partial_x f \rangle$. Then*

$$F(x, u) = f(x) + \sum_{i=1}^{k} u_i b_i(x)$$

is a universal unfolding of f.

Examples:

1. Consider $f(x) = x^4$. Here $n = 1$ and $\mu(n) = \mu(1) = \langle x \rangle$. The derivative of f is $4x^3$, so $\langle \partial_x f \rangle = \langle x^3 \rangle$ and we get $\langle x \rangle / \langle x^3 \rangle = \langle x, x^2 \rangle$. Thus $F(x, u_1, u_2) = x^4 + u_1 x^2 + u_2 x$ is the universal unfolding of f.
2. Consider $f(x, y) = x^3 + y^3$. Here $n = 2$ and $\mu(n) = \mu(2) = \langle x, y \rangle$. The partial derivatives of f with respect to x and y are $3x^2$ and $3y^2$, thus $\langle \partial_x f \rangle = \langle x^2, y^2 \rangle$ and we get $\langle x, y \rangle / \langle x^2, y^2 \rangle = \langle x, xy, y \rangle$. Therefore $F(x, y, u_1, u_2, u_3) = x^3 + y^3 + u_1 \cdot x \cdot y + u_2 \cdot x + u_3 \cdot y$ is the universal unfolding of f.

Hint: There exists a useful method for the calculation of the quotient space which you will find in the literature as "Siersma's trick" (see for example [7]).

We now present René Thom's famous list of the seven elementary catastrophes.

Classification Theorem (Thom' s List)

Up to addition of a non degenerated quadratic form in other variables and up to multiplication by ±1 a singularity f of codimension k $(1 \leq k \leq 4)$ is right-equivalent to one of the following seven:

f	codim f	universal unfolding	name
x^3	1	$x^3 + ux$	fold
x^4	2	$x^4 + ux^2 + vx$	cusp
x^5	3	$x^5 + ux^3 + vx^2 + wx$	swallowtail
$x^3 + y^3$	3	$x^3 + y^3 + uxy + vx + wy$	hyperbolic umbilic
$x^3 - xy^2$	3	$x^3 - xy^2 + u(x^2 + y^2) + vx + wy$	elliptic umbilic
x^6	4	$x^6 + ux^4 + vx^3 + wx^2 + tx$	butterfly
$x^2y + y^4$	4	$x^2y + y^4 + ux^2 + vy^2 + wx + ty$	parabolic umbilic

In a few words: The seven elementary catastrophes are the seven universal unfoldings of singular function germs of codimension k $(1 \leq k \leq 4)$. The singularities itself are called organization centers of the catastrophes.

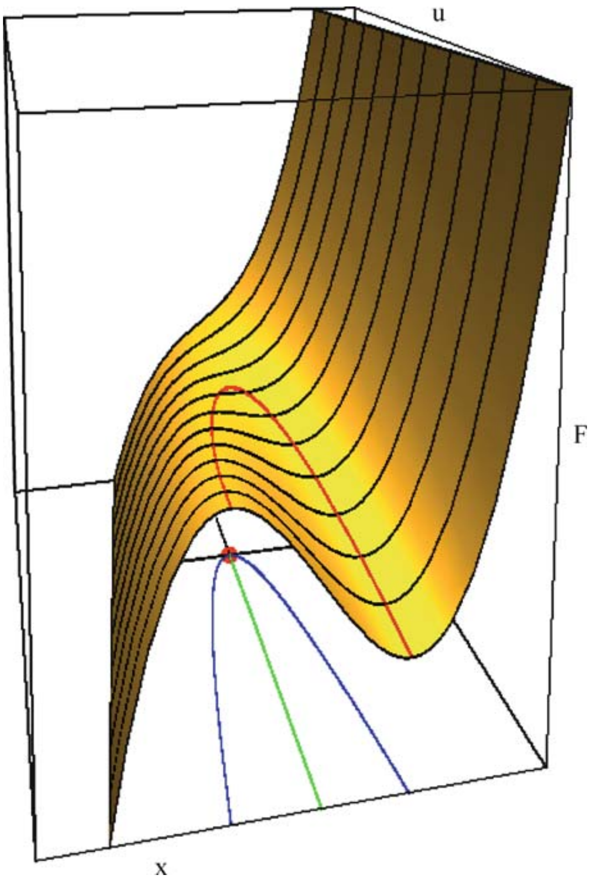
We will take a closer look at the geometry of the simplest of the seven elementary catastrophes: the fold and the cusp. The other examples are discussed in [9].

The Geometry of the Fold and the Cusp

The fold catastrophe is the first in Thom's list of the seven elementary catastrophes. It is the universal unfolding of the singularity $f(x) = x^3$ which is described by the equation $F(x, u) = x^3 + u \cdot x$. We are interested in the set of singular points of F relative to x, that is, for those points where the first partial derivative with respect to x (the system variable) vanishes. This gives the parabola $u = -3x^2$. Inserting this into F gives a space curve.

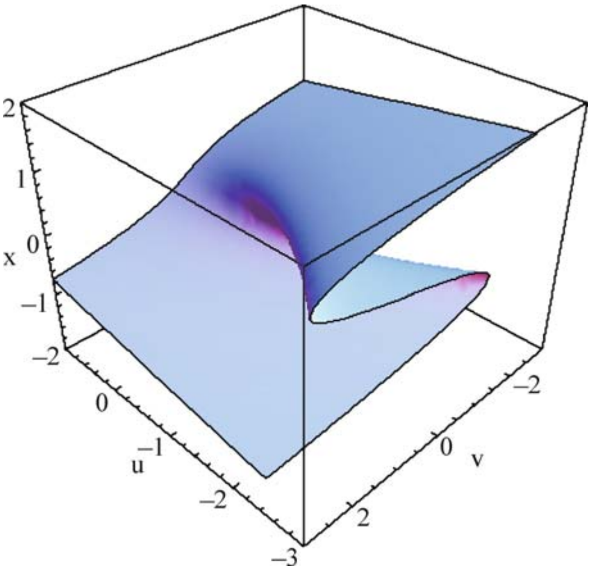
Figure 13 shows the graph of $F(x, u)$. The curve on the surface joins the extrema of F and the projection curve represents their x, u values. There is only one degenerated critical point of F (at the vertex of the Parabola), that is, a critical point where the second derivative by x also vanishes. The two branches of the parabola in the xu-plane give the positions of the maxima and the minima of F, respectively. At the points belonging to a minimum, the system is stable, while in a maximum it is unstable. Projecting x–u space, and with it the parabola, onto parameter space (u-axis), one gets a straight line shown within the interior of the parabola, which represents negative parameter values of u. There, the system has a stable minimum (and an unstable maximum, not observable in nature). For parameter values $u > 0$ the system has no stability points at all. Interpreting the parameter u as time, and "walking along" the u-axis, the point $u = 0$ is the beginning or the end of a stable system behavior. Here the system shows catastrophic behavior. According to the interpretation of the external parameter as time or space, the morphology of the fold is a "beginning," an "end" or a "border", where something new occurs.

What can we say about the fold catastrophe and the modeling of our first example, the eccentric cylinder? We have found that the potential related to the catastrophe point of the eccentric cylinder is equivalent to x^3, and $F(x, u) = x^3 + ux$ is the universal unfolding. This is independent of the size of our 'machine'. The behavior of the machine should be qualitatively the same as that of other machines which are described by the fold catastrophe as a mechanism. We can expect that the begin (or end) of a catastrophe depends on the value of a single outer parameter. Stable states are possible (for $u < 0$ in the standard fold. In the untransformed potential function of our model of the eccentric cylinder, this value is $u = 0.25268$) corresponding to the local minima of $F(., u)$. This means that it is possible for the cylinder to stay at rest on the inclined plane. There are unstable equilibria (local maxima of $F(., u)$ and the saddle point, too). The cylinder can be turned such that the center of gravity lies in the upper position over the supporting point (see Fig. 3). But a tiny disturbance will make the system leave this equilibrium.



Catastrophe Theory, Figure 13
Graph of $F(x, u) = x^3 + u \cdot x$

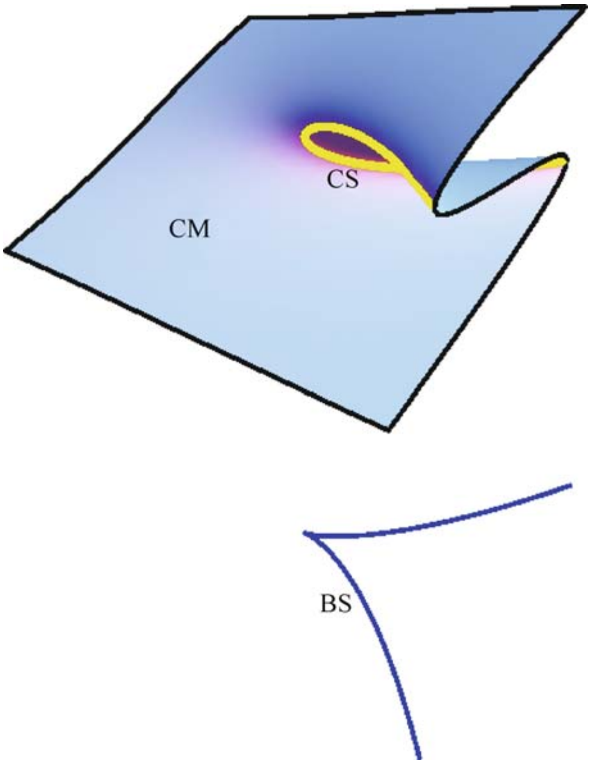
Let us now turn to the cusp catastrophe. The cusp catastrophe is the universal unfolding of the function germ $f(x) = x^4$. Its equation is $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$. In our second example (traffic jam), the function f is equivalent to the organization center of the cusp catastrophe, the second in René Thom's famous list of the seven elementary catastrophes. So the potential family is equivalent to the universal unfolding $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$ of the function x^4. We cannot draw such a function in three variables in three-space, but we can try to draw sections giving u or v certain constant values. We want to look at its catastrophe surface S, that is, the set of those points in (x, u, v)-



Catastrophe Theory, Figure 14
The catastrophe manifold of the cusp catastrophe

space, where the partial derivative of F with respect to the system variable x is zero. It is a surface in three-space called catastrophe manifold, stability surface or equilibrium surface, since the surface S describes the equilibrium points of the system.

If S is not folded over a point in u, v-space there is exactly one minimum of the potential F. If it is folded in three sheets, there are two minima (corresponding to the upper and lower sheet of the fold) and a maximum (corresponding to the middle sheet). Stable states of the system belong to the stable minima, that is, to the upper and lower sheets of the folded surface. Other points in three-space that do not lie on the surface S correspond to the states of the system that are not equilibrium. The system does not rest there. Catastrophes do occur when the system jumps suddenly from one stable state to another stable state, that is, from one sheet to the other. There are two principal possibilities, called "conventions", where this jump can happen. The "perfect delay convention" says that jumps happen at the border of the folded surface. The "Maxwell convention" says that jumps can happen along a curve in uv-space (a "shock-curve") which lays inside the cusp area. This curve consists of those points (u, v) in parameter space, where $F(., u, v)$ has two critical points with the same critical value. In our traffic jam model a shock curve is a curve in xt-space that we can interpret dynamically as the movement of the jam formation front along the street (see Fig. 9). The same figure shows what we call the morphology of the cusp: according to the convention, the



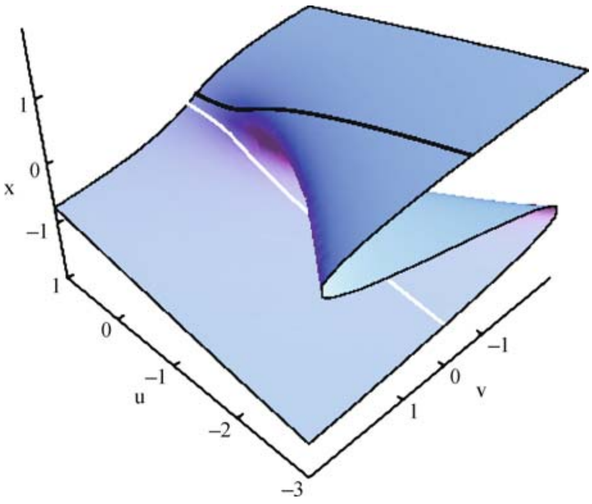
Catastrophe Theory, Figure 15
The catastrophe manifold (CM), catastrophe set (CS) and bifurcation set (BS) of the cusp catastrophe

catastrophe manifold with its fold or its cut can be interpreted as a fault or slip (as in geology) or a separation (if one of the parameter represents time). In our traffic jam model, there is a "line" (actually a region) of separation (shock wave) between the regions of high and low traffic density.

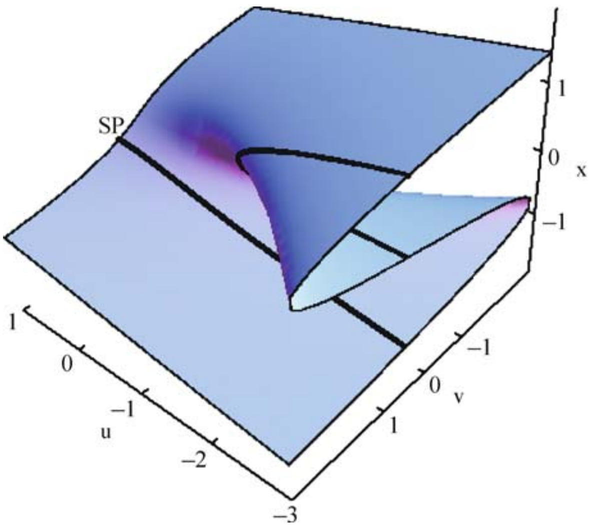
Besides the catastrophe manifold (catastrophe surface) there are two other essential terms: catastrophe set and bifurcation set (see Fig. 15). The catastrophe set can be viewed as the curve which goes along the border of the fold on the catastrophe manifold. It is the set of degenerated critical points and is described mathematically as the set $\{(x, u, v) | \partial_x F = \partial_x^2 F = 0\}$. Its projection into the uv-parameter space is the cusp curve, which is called the bifurcation set of the cusp catastrophe.

The cusp catastrophe has some special properties which we discuss now with the aid of some graphics. If the system under investigation shows one or more of these properties, the experimenter should try to find a possible cusp potential that accompanies the process.

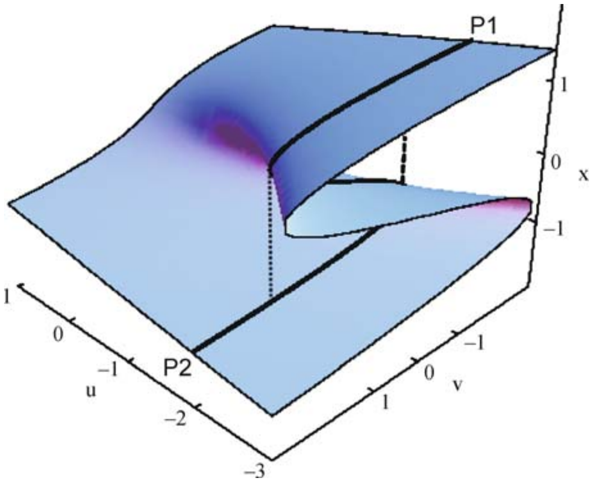
The first property is called divergence. This property can be shown by the two curves (black and white) on the



Catastrophe Theory, Figure 16
The divergence property of the cusp catastrophe



Catastrophe Theory, Figure 18
The bifurcation property of the cusp catastrophe



Catastrophe Theory, Figure 17
The hysteresis property of the cusp catastrophe

catastrophe manifold which describes the system's equilibrium points. Both curves initially start at nearby positions on the surface, that is, they start at nearby stable system states. But the development of the system can proceed quite differently. One of the curves runs on the upper sheet of the surface while the other curve runs on the lower sheet. The system's stability positions are different and thus its behavior is different.

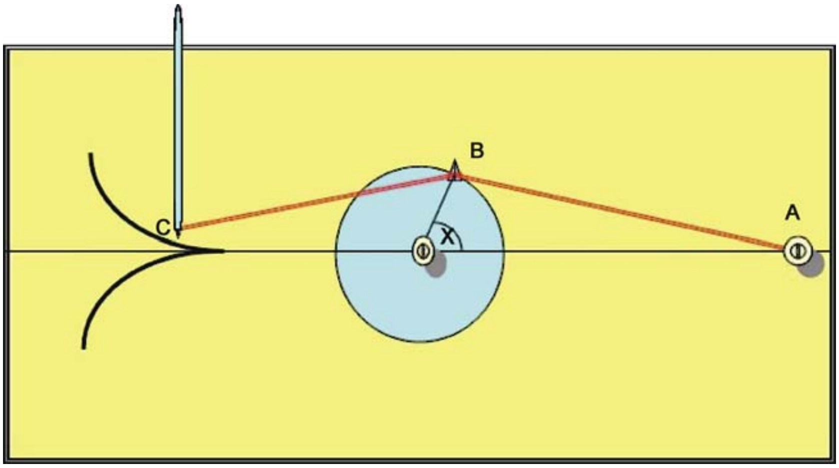
The next property is called hysteresis. If the system development is running along the path shown in Fig. 17 from P1 to P2 or vice versa, the jumps in the system behavior, that is, the sudden changes of internal system variables, occur at different parameter constellations, depending on the direction the path takes, from the upper to the lower sheet or vice versa. Jumps upward or downward happen at the border of the fold along the dotted lines. The name of this property of the cusp comes from the characteristic hysteresis curve in physics, occurring for example at the investigation of the magnetic properties of iron.

Figure 18 shows the bifurcation property of the cusp catastrophe: the number of equilibrium states of the system splits from one to three along the path beginning at the starting point (SP) and running from positive to negative u values as shown. If the upper and the lower sheet of the fold correspond to the local minima of F and the middle sheet corresponds to the local maxima, then along the path shown, one stable state splits into two stable states and one unstable state.

Further Applications

Many applications of catastrophe theory are attributable to Christopher Zeeman. For example, his catastrophe machine, the "Zeeman wheel," is often found in literature. This simple model consists of a wheel mounted flat against a board and able to turn freely. Two elastics are attached at one point (Fig. 19, point B) close to the periphery of the wheel. One elastic is fixed with its second end on the board (point A). The other elastic can be moved with its free end in the plane (point C).

Moving the free end smoothly, the wheel changes its angle of rotation smoothly almost everywhere. But at certain positions of C, which can be marked with a pencil, the wheel suddenly changes this angle dramatically. Join-



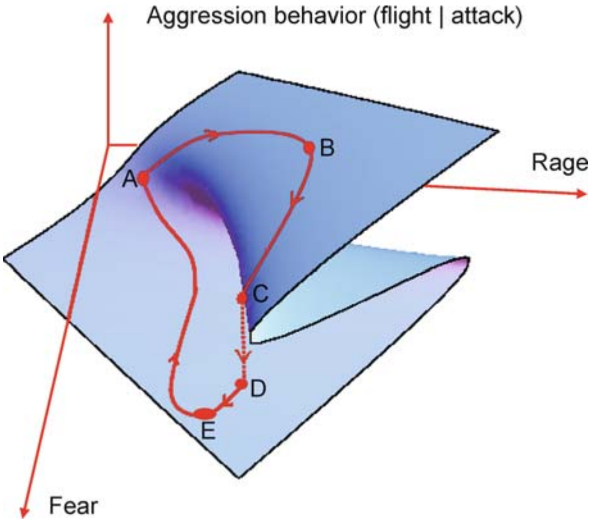
Catastrophe Theory, Figure 19
The Zeeman wheel

ing the marked points where these jumps occur, you will find a cusp curve in the plane. In order to describe this behavior, the two coordinates of the position of point C serve as control parameters in the catastrophe model. The potential function for this model results from Hook's law and simple geometric considerations. Expanding the potential near its degenerated critical point into its Taylor series, and transforming the series with diffeomorphisms similar to the examples we have already calculated, this model leads to the cusp catastrophe. Details can be found in the book of Poston and Stewart [7] or Saunders (1986).

Another example of an application of catastrophe theory is similar to the traffic jam model we used here. Catastrophe theory can describe the formation of shockwaves in hydrodynamics. Again the cusp catastrophe describes this phenomenon. The mathematical frame is given in the works of Lax [6], Golubitsky and Schaeffer[3] and Guckenheimer [4].

In geometrical optics light rays are investigated which are reflected on smooth surfaces. Caustics are the envelopes of the light rays. They are the bright patterns of intensive light, which can be seen, for example, on a cup of coffee when bright sunlight is reflected on the border of the cup. Catastrophe theory can be applied to light caustics, because the light rays obey a variational principle. According to Fermat's principle, light rays travel along geodesics and the role of the potential functions in catastrophe theory is played by geodesic distance functions. Caustics are their bifurcation sets. The calculations are given for example in the work of Sinha [11].

A purely speculative example, given by Zeeman, is an aggression model for dogs. Suppose that fear and rage of a dog can be measured in some way from aspects of its body language (its attitude, ears, tail, mouth, etc.). Fear and rage are two conflicting parameters, so that a catastrophe in the dog's behavior can happen. Look at Fig. 20.



Catastrophe Theory, Figure 20
A dog's aggression behavior depending on its rage and fear

The vertical axis shows the dog's aggression potential, the other axes show its fear and rage. Out of the many different ways the dog can behave, we choose one example. Suppose we start at point A on the catastrophe manifold (behavioral surface). Maybe the dog is dozing in the sun without any thought of aggression. The dog's aggression behavior is neutral (point A). Another dog is coming closer and our dog, now awakened, is becoming more and more angry by this breach of his territory. Moving along the path on the behavior surface, the dog is on a trajectory to attack his opponent. But it suddenly notices that the enemy is very big (point B). So its fear is growing, its rage diminishes. At point C the catastrophe happens: Attack suddenly changes to flight (point D). This flight continues until the dog notices that the enemy does not follow (point E). Fear and rage both get smaller until the dog becomes calm again.

The example can be modified a little bit to apply to conflicts between two states. The conflicting factors can be the costs and the expected gain of wars, and so on.

Among the variety of possible examples, those worthy of mention include investigations of the stability of ships, the gravitational collapse of stars, the buckling beam (Euler beam), the breaking of ocean waves, and the development of cells in biology. The latter two problems are examples for the occurrence of higher catastrophes, for example the hyperbolic umbilic (see Thom's List of the seven elementary catastrophes).

Future Directions

When catastrophe theory came up in the nineteen-sixties, much enthusiasm spread among mathematicians and other scientists about the new tool that was expected to explain many of the catastrophes in natural systems. It seemed to be the key for the explanation of discontinuous phenomena in all sciences. Since many applications were purely qualitative and speculative, a decade later this enthusiasm ebbed away and some scientists took this theory for dead. But it is far from that. Many publications in our day show that it is still alive. The theory has gradually become a 'number producing' theory, so that it is no longer perceived as purely qualitative. It seems that it will be applied the sciences increasingly, producing numerical results. Thom's original ideas concerning mathematical biology seem to have given the basis for the trends in modern biology. This is probably the most promising field of application for catastrophe theory. New attempts are also made, for example, in the realm of decision theory or in statistics, where catastrophe surfaces may help to clarify statistical data.

From the point of view of teaching and learning mathematics, it seems that catastrophe theory is becoming and increasingly popular part of analysis courses at our universities. Thus, students of mathematics meet the basic ideas of catastrophe theory within their first two years of undergraduate studies. Perhaps in the near future its basic principles will be taught not only to the mathematical specialists but to students of other natural sciences, as well.

Bibliography

Primary Literature

1. Bröcker T (1975) Differentiable germs and catastrophes. London Mathem. Soc. Lecture Notes Series. Cambridge Univ Press, Cambridge
2. Golubitsky M, Guillemin V (1973) Stable mappings and their singularities. Springer, New York, GTM 14
3. Golubitsky M, Schaeffer DG (1975) Stability of shock waves for a single conservation law. Adv Math 16(1):65–71
4. Guckenheimer J (1975) Solving a single conservation law. Lecture Notes, vol 468. Springer, New York, pp 108–134
5. Haberman R (1977) Mathematical models – mechanical vibrations, population dynamics, and traffic flow. Prentice Hall, New Jersey
6. Lax P (1972) The formation and decay of shockwaves. Am Math Monthly 79:227–241
7. Poston T, Stewart J (1978) Catastrophe theory and its applications. Pitman, London
8. Sanns W (2000) Catastrophe theory with mathematica – a geometric approach. DAV, Germany
9. Sanns W (2005) Genetisches Lehren von Mathematik an Fachhochschulen am Beispiel von Lehrveranstaltungen zur Katastrophentheorie. DAV, Germany
10. Saunders PT (1980) An introduction to catastrophe theory. Cambridge University Press, Cambridge. Also available in German: Katastrophentheorie. Vieweg (1986)
11. Sinha DK (1981) Catastrophe theory and applications. Wiley, New York
12. Wassermann G (1974) Stability of unfoldings. Lect Notes Math, vol 393. Springer, New York

Books and Reviews

Arnold VI (1984) Catastrophe theory. Springer, New York

Arnold VI, Afrajmovich VS, Il'yashenko YS, Shil'nikov LP (1999) Bifurcation theory and catastrophe theory. Springer

Bruce JW, Gibblin PJ (1992) Curves and singularities. Cambridge Univ Press, Cambridge

Castrigiano D, Hayes SA (1993) Catastrophe theory. Addison Wesley, Reading

Chillingworth DRJ (1976) Differential topology with a view to applications. Pitman, London

Demazure M (2000) Bifurcations and catastrophes. Springer, Berlin

Fischer EO (1985) Katastrophentheorie und ihre Anwendung in der Wirtschaftswissenschaft. Jahrb f Nationalök u Stat 200(1):3–26

Förster W (1974) Katastrophentheorie. Acta Phys Austr 39(3):201–211

Gilmore R (1981) Catastrophe theory for scientists and engineers. Wiley, New York

Golubistky M (1978) An introduction to catastrophe theory. SIAM 20(2):352–387

Golubitsky M, Schaeffer DG (1985) Singularities and groups in bifurcation theory. Springer, New York

Guckenheimer J (1973) Catastrophes and partial differential equations. Ann Inst Fourier Grenoble 23:31–59

Gundermann E (1985) Untersuchungen über die Anwendbarkeit der elementaren Katastrophentheorie auf das Phänomen "Waldsterben". Forstarchiv 56:211–215

Jänich K (1974) Caustics and catastrophes. Math Ann 209:161–180

Lu JC (1976) Singularity theory with an introduction to catastrophe theory. Springer, Berlin

Majthay A (1985) Foundations of catastrophe theory. Pitman, Boston

Mather JN (1969) Stability of C^{∞}-mappings. I: Annals Math 87:89–104; II: Annals Math 89(2):254–291

Poston T, Stewart J (1976) Taylor expansions and catastrophes. Pitman, London

Stewart I (1977) Catastrophe theory. Math Chronicle 5:140–165

Thom R (1975) Structural stability and morphogenesis. Benjamin Inc., Reading

Thompson JMT (1982) Instabilities and catastrophes in science and engineering. Wiley, Chichester

Triebel H (1989) Analysis und mathematische Physik. Birkhäuser, Basel

Ursprung HW (1982) Die elementare Katastrophentheorie: Eine Darstellung aus der Sicht der Ökonomie. Lecture Notes in Economics and Mathematical Systems, vol 195. Springer, Berlin

Woodcock A, Davis M (1978) Catastrophe theory. Dutton, New York

Zeeman C (1976) Catastrophe theory. Sci Am 234(4):65–83

Cell Biology: Networks, Regulation and Pathways

GAŠPER TKAČIK[1], WILLIAM BIALEK[1,2]

[1] Joseph Henry Laboratories of Physics,
Lewis–Sigler Institute for Integrative Genomics,
Princeton University, Princeton, USA

[2] Princeton Center for Theoretical Physics,
Princeton University, Princeton, USA

Article Outline

Glossary

Dynamical system is a set of components the properties of which (e. g. their quantity, activity level etc.) change in time because the components interact among themselves and are also influenced by external forces.

Network node is a constituent component of the network, in biological networks most often identified with a molecular species.

Interaction is a connection between network nodes; in biological networks an interaction means that two nodes chemically react, regulate each other, or effectively influence each other's activities. Interactions are mostly pairwise, but can be higher-order as well; they can be directed or undirected, and are usually characterized by an interaction strength.

Network is a system of interacting nodes. A network can be represented mathematically as a graph, where vertices denote the nodes and edges denote the interactions. Biological networks often are understood to be dynamical systems as well, because the activities of network nodes evolve in time due to the graph of interactions.

Network state is the vector of activities of all nodes that fully characterizes the network at any point in time; since a biological network is a dynamical system, this state generally changes through time according to a set of dynamical equations.

Biological function refers to the role that a specific network plays in the life of the organism; the network can be viewed as existing to perform a task that enables the cell to survive and reproduce, such as the detection or transduction of a specific chemical signal.

Pathway is a subset of nodes and interactions in a network along which information or energy and matter flow in a directed fashion; pathways can be coupled through interactions or unwanted cross-talk.

Curse of dimensionality is the rapid increase of complexity encountered when analyzing or experimentally observing network states, as more and more network nodes are added. If there are N network nodes each of which only has two states (for example *on* and *off*), the number of states that the network can be in grows as 2^N.

Design principle is an (assumed) constraint on the network architecture, stating that a biological network, in addition to performing a certain function, implements that function in a particular way, usually to maximize or minimize some further objective measure, for instance robustness, information transmission, or designability.

Definition of the Subject

In cell biology, networks are systems of interacting molecules that implement cellular functions, such as the regulation of gene expression, metabolism or intracellular signaling. While on a molecular level a biological network is a mesh of chemical reactions between, for example, enzymes and their substrates, or DNA-binding proteins and the genes that they regulate, the collective effect of these reactions can often be thought of as the enabling and regulating the flow of matter and energy (in metabolic networks), or of information (in signaling and transcriptional regulatory networks). The field is concerned primarily with the description and properties of such flows and with their emergence from network constituent parts – the molecules and their physical interactions. An important focus is also the question of how network function and operating principles can be inferred despite the limited experimental access to network states and building blocks.

Introduction

Biological network has come to mean a system of interacting molecules that jointly perform cellular tasks such as the regulation of gene expression, information transmission, or metabolism [28]. Specific instances of biological networks include, for example, the DNA and DNA binding proteins comprising the transcriptional regulatory network; signaling proteins and small molecules comprising various signaling networks; or enzymes and metabolites comprising the metabolic network. Two important assumptions shape our current understanding of such systems: first, that the biological networks have been under selective evolutionary pressure to perform specific cellular functions in a way that furthers the overall reproductive success of the individual; and second, that these functions often are not implemented on a microscopic level by single molecules, but are rather a collective property of the whole interaction network. The question of how complex behavior emerges in a network of (simple) nodes under a functional constraint is thus central [144].

To start off with a concrete example, consider chemotaxis in the bacterium *Escherichia coli* [16,40], one of the paradigmatic examples of signal transduction. This system is dedicated to steering the bacteria towards areas high in nutrient substances and away from repellents. Chemoeffector molecules in the solution outside the bacterium bind to receptor molecules on the cell surface, and the resulting structural changes in the receptors are relayed in turn by the activities of the intracellular signaling proteins to generate a control signal for molecular motors that drive the bacterial flagella. The chemotactic network consists of about 10 nodes (here, signaling proteins), and the interactions between the nodes are the chemical reactions of methylation or phosphorylation. Notable features of this system include its extreme sensitivity, down to the limits set by counting individual molecules as they arrive at the cell surface [17], and the maintenance of this sensitivity across a huge dynamic range, through an adaptation mechanism that provides nearly perfect compensation of background concentrations [27]. More recently it has been appreciated that aspects of this functionality, such as perfect adaptation, are also robust against large variations in the concentrations of the network components [6].

Abstractly, different kinds of signaling proteins, such those in chemotaxis, can be thought of as the building blocks of a network, with their biochemical interactions forming the wiring diagram of the system, much like the components and wiring diagram of, for instance, a radio receiver. In principle, these wiring diagrams are hugely complex; for a network composed of N species, there are $\sim C_k^N$ possible connections among any set of k components, and typically we don't have direct experimental guidance about the numbers associated with each 'wire.' One approach is to view this as giant fitting problem: once we draw a network, there is a direct translation of this graph into dynamical equations, with many parameters, and we should test the predictions of these dynamics against whatever data are available to best determine the underlying parameters. Another approach is to ask whether this large collection of parameters is special in any way other than that it happens to fit the data – are there principles that allow us to predict how these systems *should* work? In the context of chemotaxis, we might imagine that network parameters have been selected to optimize the average progress of bacteria up the chemical gradients of nutrients, or to maximize the robustness of certain functions against extreme parameter variations. These ideas of design principles clearly are not limited to bacterial chemotaxis.

An important aspect of biological networks is that the same components (or components that have an easily identifiable evolutionary relationship) can be (re)used in different modules or used for the same function in a different way across species, as discussed for example by Rao et al. [118] for the case of bacterial chemotaxis. Furthermore, because evolutionary selection depends on function and not directly on microscopic details, different wiring diagrams or even changes in components themselves can result in the same performance; evolutionary process can gradually change the structure of the network as long as its function is preserved; as an example see the discussion of transcriptional regulation in yeast by Tanay et al. [148]. On the other hand, one can also expect that signal processing problems like gain control, noise reduction, ensuring (bi)stability etc, have appeared and were solved repeatedly, perhaps even in similar ways across various cellular functions, and we might be able to detect the traces of their commonality in the network structure, as for example in the discussion of local connectivity in bacterial transcriptional regulation by Shen–Orr et al. [136]. Thus there are reasons to believe that in addition to design principles at the network level, there might also be local organizing principles, similar to common wiring motifs in electronic circuitry, yet still independent of the identity of the molecules that implement these principles.

Biological networks have been approached at many different levels, often by investigators from different disciplines. The basic wiring diagram of a network – the fact that a kinase phosphorylates these particular proteins, and not all others, or that a transcription factor binds to the promoter regions of particular genes – is determined by classical biochemical and structural concepts such as binding specificity. At the opposite extreme, trying to understand the collective behavior of the network as a whole suggests approaches from statistical physics, often looking at simplified models that leave out many molecular details. Analyses that start with design principles are yet a different approach, more in the 'top–down' spirit of statistical physics but leaving perhaps more room for details to emerge as the analysis is refined. Eventually, all of these different views need to converge: networks really are built out of molecules, their functions emerge as collective behaviors, and these functions must really be functions of use to the organism. At the moment, however, we seldom know enough to bridge the different levels of description, so the different approaches are pursued more or less independently, and we follow this convention here. We will start with the molecular building blocks, then look at models for networks as a whole, and finally consider design principles. We hope that this sequence doesn't leave the impression that we actually know how to build up from molecules to function!

Before exploring our subject in more detail, we take a moment to consider its boundaries. Our assignment from the editors was to focus on phenomena at the level of molecular and cellular biology. A very different approach attempts to create a 'science of networks' that searches for common properties in biological, social, economic and computer networks [104]. Even within the biological world, there is a significant divide between work on networks in cell biology and networks in the brain. As far as we can see this division is an artifact of history, since there are many issues which cut across these different fields. Thus, some of the most beautiful work on signaling comes from photoreceptors, where the combination of optical inputs and electrical outputs allowed, already in the 1970s, for experiments with a degree of quantitative analysis that even today is hard to match in systems which take chemical inputs and give outputs that modulate the expression levels of genes [14,121]. Similarly, problems of noise in the control of gene expression have parallels in the long history of work on noise in ion channels, as we have discussed elsewhere [156], and the problems of robustness have also been extensively explored in the network of interactions among the multiple species of ion channels in the membrane [51,88]. Finally, the ideas of collective behavior are much better developed in the context of neural networks than in cellular networks, and it is an open question how much can be learned by studying these different systems in the same language [151].

Biological Networks and Their Building Blocks

Genetic Regulatory Networks

Cells constantly adjust their levels of gene expression. One central mechanism in this regulatory process involves the control of transcription by proteins known as transcription factors (TFs), which locate and bind short DNA sequences in the regulated genes' promoter or enhancer regions. A given transcription factor can regulate either a few or a sizable proportion of the genes in a genome, and a single gene may be regulated by more than one transcription factor; different transcription factors can also regulate each other [166].

In the simplest case of a gene regulated by a single TF, the gene might be expressed whenever the factor – in this case called an activator – is bound to the cognate sequence in the promoter (which corresponds to the situation when the TF concentration in the nucleus is high), whereas the binding of a repressor would shut a normally active gene down. The outlines of these basic control principles were established long ago, well before the individual transcription factors could be isolated, in elegant experiments on the *lactose* operon of *Escherichia coli* [69] and even simpler model systems such as phage λ [115]. To a great extent the lessons learned from these experiments have provided the framework for understanding transcriptional control more generally, in prokaryotes [114], eukaryotes [75], and even during the development of complex multicellular organisms [8].

The advent of high throughput techniques for probing gene regulation has extended our reach beyond single genes. In particular, microarrays [30] and the related data analysis tools, such as clustering [36], have enabled researchers to find sets of genes, or *modules*, that are *coexpressed*, i. e. up- or down-regulated in a correlated fashion when the organism is exposed to different external conditions, and are thus probably regulated by the same set of transcription factors. Chromatin immunoprecipitation (ChIP) assays have made it possible to directly screen for short segments of DNA that known TFs bind; using microarray technology it is then possible to locate the intergenic regions which these segments belong to, and hence find the regulated genes, as has recently been done for the *Saccharomyces cerevisiae* DNA-TF interaction map [86].

These high throughput experimental approaches, combined with traditional molecular biology and complemented by sequence analysis and related mathematical tools [139], provide a large scale, topological view of the transcriptional regulatory network of a particular organism, where each link between two nodes (genes) in the regulatory graph implies either activation or repression [5]. While useful for describing causal interactions and trying to predict responses to mutations and external perturbations [89], this picture does not explain how the network operates on a physical level: it lacks dynamics and specifies neither the strengths of the interactions nor how all the links converging onto a given node jointly exercise control over it. To address these issues, representative wild-type or simple synthetic regulatory elements and networks consisting of a few nodes have been studied extensively to construct quantitative models of the network building blocks.

For instance, combinatorial regulation of a gene by several transcription factors that bind and interact on the promoter has been considered by Buchler et al. [31] as an example of (binary) biological computation and synthetic networks implementing such computations have been created [56,170]. Building on classical work describing allosteric proteins such as hemoglobin, thermodynamic models have been used with success to account for combinatorial interactions on the operator of the λ phage [2]. More recently Bintu et al. [24,25] have reviewed the equilibrium statistical mechanics of such interactions, Setty et al. [134] have experimentally and systematically mapped out the response surface of the *lac* promoter to combinations of its two regulatory inputs, cAMP and IPTG, and Kuhlman et al. [85] have finally provided a consistent picture of the known experimental results and the thermodynamic model for the combinatorial regulation of the lactose operon. There have also been some successes in eukaryotic regulation, where Schroeder et al. [132] used thermodynamically motivated models to detect clusters of binding sites that regulate the gap genes in morphogenesis of the fruit fly.

Gene regulation is a dynamical process composed of a number of steps, for example the binding of TF to DNA, recruitment of transcription machinery and the production of the messenger RNA, post-transcriptional regulation, splicing and transport of mRNA, translation, maturation and possible localization of proteins. While the extensive palette of such microscopic interactions represents a formidable theoretical and experimental challenge for each detailed study, on a network level it primarily induces three effects. First, each node – usually understood as the amount of gene product – in a graph of regulatory interactions is really not a single dynamical variable, but has a nontrivial internal state representing the configuration on the associated promoter, concentration of the corresponding messenger RNA etc.; the relation of these quantities to the concentration of the output protein is not necessarily straightforward, as emphasized in recent work comparing mRNA and protein levels in yeast [46]. Second, collapsing multiple chemical species onto a single node makes it difficult to include non-transcriptional regulation of gene expression in the same framework. Third, the response of the target gene to changes in the concentrations of its regulators will be delayed and extended in time, as in the example explored by Rosenfeld and Alon [123].

Perhaps the clearest testimonies to the importance of dynamics in addition to network topology are provided by systems that involve regulatory loops, in which the output of a network feeds back on one of the inputs as an activator or repressor. McAdams and Shapiro [99] have argued that the time delays in genetic regulatory elements are essential for the proper functioning of the phage λ switch, while Elowitz and Leibler [38] have created a synthetic circuit made up of three mutually repressing genes (the "repressilator"), that exhibits spontaneous oscillations. Circadian clocks are examples of naturally occurring genetic oscillators [171].

In short, much is known about the skeleton of genetic regulatory interactions for model organisms, and physical models exist for several well studied (mostly prokaryotic) regulatory elements. While homology allows us to bridge the gap between model organisms and their relatives, it is less clear how and at which level of detail the knowledge about regulatory elements must be combined into a network to explain and predict its function.

Protein–Protein Interaction Networks

After having been produced, proteins often assemble into complexes through direct contact interactions, and these complexes are functionally active units participating in signal propagation and other pathways. Proteins also interact through less persistent encounters, as when a protein kinase meets its substrate. It is tempting to define a link in the network of protein–protein interactions by such physical associations, and this is the basis of several experimental methods which aim at a genome-wide survey of these interactions. Although starting out being relatively unreliable (with false positive rates of up to 50%), high throughput techniques like the yeast two hybrid assay [68,161] or mass spectrometry [45,61] are providing data of increasing quality about protein–protein interactions, or the “interactome” [84]. While more reliable methods are being developed [5] and new organisms are being analyzed in this way [49,91,125], the existing interaction data from high throughput experiments and curated databases has already been extensively studied.

Interpretation of the interactions in the protein network is tricky, however, due to the fact that different experimental approaches have various biases – for example, mass spectrometry is biased towards detecting interactions between proteins of high abundance, while two hybrid methods seem to be unbiased in this regard; on the other hand, all methods show some degree of bias towards different cellular localizations and evolutionary novelty of the proteins. Assessing such biases, however, currently depends not on direct calibration of the methods themselves but on comparison of the results with manually curated databases, although the databases surely have their own biases [70]. It is reassuring that the intersection of various experimental results shows significantly improved agreement with the databases, but this comes at the cost of a substantial drop in coverage of the proteome [100].

In contrast to the case of transcriptional regulation, the relationship between two interacting proteins is symmetric: if protein A binds to protein B, B also binds to A, so that the network is described by an undirected graph. Most of the studies have been focused on binary interactions that yeast two hybrid and derived approaches can probe, although spectrometry can detect multiprotein complexes as well. Estimates of number of links in these networks vary widely, even in the yeast *Saccharomyces cerevisiae*: Krogan et al. [84] directly measure around 7100 interactions (between 2700 proteins), while Tucker et al. [158] estimate the total to be around 13 000–17 000, and von Mering et al. [100] would put the lower estimate at about 30 000. Apart from the experimental biases that can influence such estimates and have been discussed already, it is important to realize that each experiment can only detect interactions between proteins that are expressed under the chosen external conditions (e. g. the nutrient medium); moreover, interactions can vary from being transient to permanent, to which various measurement methods respond differently. It will thus become increasingly important to qualify each interaction in a graph by specifying how it depends on context in which the interaction takes place.

Proteins ultimately carry out most of the cellular processes such as transcriptional regulation, signal propagation and metabolism, and these processes can be modeled by their respective network and dynamical system abstractions. In contrast, the interactome is not a dynamical system itself, but instead captures specific reactions (like protein complex assembly) and structural and/or functional relations that are present in all of the above processes. In this respect it has an important practical role of annotating currently unknown proteins through ‘guilt by association,’ by tying them into complexes and processes with a previously known function.

Metabolic Networks

Metabolic networks organize our knowledge about anabolic and catabolic reactions between the enzymes, their substrates and co-factors (such as ATP), by reducing the set of reactions to a graph representation where two substrates are joined by a link if they participate in the same reaction. For model organisms like the bacterium *Escherichia coli* the metabolic networks have been studied in depth and are publicly available [77,78], and an increasing number of analyzed genomes offers sufficient sampling power to make statistical statements about the network properties across different domains of life [72].

Several important features distinguish metabolic from protein–protein interaction and transcriptional regulation networks. First, for well studied systems the coverage of metabolic reactions is high, at least for the central routes of energy metabolism and small molecule synthesis; notice that this is a property of our knowledge, not a property of the networks (!). Second, cellular concentrations of metabolites usually are much higher than those of transcription factors, making the stochasticity in reactions due to small molecular counts irrelevant. Third, knowledge of the stoichiometry of reactions allows one to directly write down a system of first order differential equations for the metabolite fluxes [60], which in steady state reduces to a set of linear constraints on the space of solutions. These chemical constraints go beyond topology and can yield strong and testable predictions; for example, Ibarra et al. [66] have shown how computationally maximizing the growth rate of *Escherichia coli* within the space of allowed solutions given by flux balance constraints can correctly predict measurable relationships between oxygen and substrate uptake, and that bacteria can be evolved towards the predicted optimality for growth conditions in which the response was initially suboptimal.

Signaling Networks

Signaling networks consist of receptor and signaling proteins that integrate, transmit and route information by means of chemical transformations of the network constituents. One class of such transformations, for example, are post–translational modifications, where targets are phosphorylated, methylated, acetylated, … on specific residues, with a resulting change in their enzymatic (and thus signaling) activity. Alternatively, proteins might form stable complexes or dissociate from them, again introducing states of differential activity. The ability of cells to modify or tag proteins (possibly on several residues) can increase considerably the cell's capacity to encode its state and transmit information, assuming that the signaling proteins are highly specific not only for the identity but also the modification state of their targets; for a review see [110].

Despite the seeming overlap between the domains of protein–protein network and signaling networks, the focus of the analysis is substantially different. The interactome is simply a set of possible protein–protein interactions and thus a topological (or connectivity) map; in contrast, signaling networks aim to capture signal transduction and therefore need to establish a causal map, in which the nature of the protein–protein interaction, its direction and timescale, and its quantitative effect on the activity of the target protein matter. As an example, see the discussion by Kolch et al. [83] on the role of protein–protein interactions in MAPK signaling cascade.

Experiments on some signaling systems, such as the *Escherichia coli* chemotactic module, have generated enough experimental data to require detailed models in the form of dynamical equations. Molecular processes in a signaling cascade extend over different time scales, from milliseconds required for kinase and phosphatase reactions and protein conformational changes, to minutes or more required for gene expression control, cell movement and receptor trafficking; this fact, along with the (often essential) spatial effects such as the localization of signaling machinery and diffusion of chemical messengers, can considerably complicate analyses and simulations.

Signaling networks are often factored into pathways that have specific inputs, such as the ligands of the G protein coupled receptors on the cell surface, and specific outputs, as with pathways that couple to the transcriptional regulation apparatus or to changes in the intracellular concentration of messengers such as calcium or cyclic nucleotides. Nodes in signaling networks can participate in several pathways simultaneously, thus enabling signal integration or potentially inducing damaging "crosstalk" between pathways; how junctions and nodes process signals is an area of active research [74].

The components of signaling networks have long been the focus of biochemical research, and genetic methods allow experiments to assess the impact of knocking out or over-expressing particular components. In addition, several experimental approaches are being designed specifically for elucidating signaling networks. Ab-chips localize various signaling proteins on chips reminiscent of DNA microarrays, and stain them with appropriate fluorescent antibodies [105]. Multicolor flow cytometry is performed on cells immuno-stained for signaling protein modifications and hundreds of single cell simultaneous measurements of the modification state of pathway nodes are collected [113]. Indirect inference of signaling pathways is also possible from genomic or proteomic data.

One well studied signal transduction system is the mitogen activated protein kinase (MAPK) cascade that controls, among other functions, cell proliferation and differentiation [32]. Because this system is present in all eukaryotes and its structural components are used in multiple pathways, it has been chosen as a paradigm for the study of specificity and crosstalk. Similarly, the TOR system, identified initially in yeast, is responsible for integrating the information on nutrient availability, growth factors and energy status of the cell and correspondingly regulating the cell growth [95]. Another interesting example of signal integration and both intra- and inter-cellular communication is observed in the quorum sensing circuit of the bacterium *Vibrio harveyi*, where different kinds of species- and genus-specific signaling molecules are detected by their cognate receptors on the cell surface, and the information is fed into a common *Lux* phosphorelay pathway which ultimately regulates the quorum sensing genes [165].

Models of Biological Networks

Topological Models

The structural features of a network are captured by its connectivity graph, where interactions (reactions, structural relations) are depicted as the links between the interacting nodes (genes, proteins, metabolites). Information about connectivity clearly cannot and does not describe the network behavior, but it might influence and constrain it in revealing ways, similar to effect that the topology of the lattice has on the statistical mechanics of systems living on it.

Theorists have studied extensively the properties of regular networks and random graphs starting with Erdös and Rényi in 1960s. The first ones are characterized by high symmetry inherent in a square, triangular, or all-to-all (mean field) lattice; the random graphs were without such regularity, constructed simply by distributing K links at random between N nodes. The simple one–point statistical characterization that distinguishes random from regular networks looks at the node degree, that is the probability $P(k)$ that any node has k incoming and/or outgoing links. For random graphs this distribution is Poisson, meaning that most of the nodes have degrees very close to the mean, $\langle k \rangle = \sum_k k\, P(k)$, although there are fluctuations; for regular lattices every node has the same connectivity to its neighbors.

The first analyses of the early reconstructions of large metabolic networks revealed a surprising "scale free" node degree distribution, that is $P(k) \sim k^{-\gamma}$, with γ between 2 and 3 for most networks. For the physics community, which had seen the impact of such scale invariance on our understanding of phase transitions, these observations were extremely suggestive. It should be emphasized that for many problems in areas as diverse as quantum field theory, statistical mechanics and dynamical systems, such scaling relations are much more than curiosities. Power laws relating various experimentally observable quantities are exact (at least in some limit), and the exponents (here, γ) really contain everything one might want to know about the nature of order in the system. Further, some of the first thoughts on scaling emerged from phenomenological analyses of real data. Thus, the large body of work on scaling ideas in theoretical physics set the stage for people to be excited by the experimental observation of power laws in much more complex systems, although it is not clear to us whether the implied promise of connection to a deeper theoretical structure has been fulfilled. For divergent views on these matters see Barabási et al. [10] and Keller et al. [81].

The most immediate practical consequence of a scale free degree distribution is that – relative to expectations based on random graphs – there will be an over-representation of nodes with very large numbers of links, as with pyruvate or co-enzyme A in metabolic networks [72,163]. These are sometimes called hubs, although another consequence of a scale free distribution is that there is no 'critical degree of connection' that distinguishes hubs from nonhubs. In the protein–protein interaction network of *Saccharomyces cerevisiae*, nodes with higher degree are more likely to represent essential proteins [73], suggesting that node degree does have some biological meaning. On the theoretical side, removal of a sizable fraction of nodes from a scale free network will neither increase the network diameter much, nor partition the network into equally sized parts [3], and it is tempting to think that this robustness is also biologically significant. The scale free property has been observed in many non-biological contexts, such as the topology of social interactions, World Wide Web links, electrical power grid connectivity ... [144]. A number of models have been proposed for how such scaling might arise, and some of these ideas, such as growth by preferential attachment, have a vaguely biological flavor [11,12]. Finding the properties of networks that actually discriminate among different mechanisms of evolution or growth turns out to be surprisingly subtle [173].

Two other revealing measures are regularly computed for biological networks. The mean path length, $\langle l \rangle$, is the shortest path between a pair of nodes, averaged over all pairs in the graph, and measures the network's overall 'navigability.' Intuitively, short path lengths correspond to, for example, efficient or fast flow of information and energy in signaling or metabolic networks, quick spread of diseases in a social network and so on. The clustering coefficient of a node i is defined as $C_i = 2n_i/k_i(k_i - 1)$, where n_i is the number of links connecting the k_i neighbors of node i to each other; equivalently, C_i is the ratio between the number of triangles passing through two neighbors of i and node i itself, divided by the maximum possible number of such triangles. Random networks have low path lengths and low clustering coefficients, whereas regular lattices have long path lengths and are locally clustered. Watts and Strogatz [167] have constructed an intermediate regime of "small world" networks, where the regular lattice has been perturbed by a small number of random links connecting distant parts of the network together. These networks, although not necessarily scale free, have short path lengths and high clustering coefficients, a property that was subsequently observed in metabolic and other biological networks as well [163].

A high clustering coefficient suggests the existence of densely connected groups of nodes within a network, which seems contradictory to the idea of scale invariance, in which there is no inherent group or cluster size; Ravasz et al. [120] addressed this problem by introducing hierarchical networks and providing a simple construction for synthetic hierarchical networks exhibiting both scale free and clustering behaviors. Although there is no unique scale for the clusters, clusters will appear at any scale one chooses to look at, and this is revealed by the scaling of clustering coefficient $C(k)$ with the node degree k, $C(k) \sim k^{-1}$, on both synthetic as well as natural metabolic networks of organisms from different domains of life [120]. Another interesting property of some biological networks is an anti-correlation of node degree of connected nodes [96], which we can think of as a 'dissociative' structure; in contrast, for example, with the associative character of social networks, where well connected people usually know one another.

As we look more finely at the structure of the graph representing a network, there is of course a much greater variety of things to look at. For example, Spirin and Mirny [142] have focused on high clustering coefficients as a starting point and devised algorithms to search for modules, or densely connected subgraphs within the yeast protein–protein interaction network. Although the problem has combinatorial complexity in general, the authors found about 50 modules (of 5–10 proteins in size, some of which were unknown at the time) that come in two types: the first represents dynamic functional units (e. g. signaling cascades), and the second protein complexes. A similar conclusion was reached by Han et al. [57], after having analyzed the interactome in combination with the temporal gene expression profiles and protein localization data; the authors argue that nodes of high degree can sit either at the centers of modules, which are simultaneously expressed ("party hubs"), or they can be involved in various pathways and modules at different times ("date hubs"). The former kind is at a lower level of organization, whereas the latter tie the network into one large connected component.

Focusing on even a smaller scale, Shen-Orr et al. [136] have explored motifs, or patterns of connectivity of small sets of nodes that are over-represented in a given network compared to the randomized networks of the same degree distribution $P(k)$. In the transcriptional network of the bacterium *E. coli*, three such motifs were found: feed forward loops (in which gene X regulates Y that regulates Z, but X directly regulates Z as well), single input modules (where gene X regulates a large number of other genes in the same way and usually auto-regulates itself), and dense overlapping regulons (layers of overlapping interactions between genes and a group of transcription factors, much denser than in randomized networks). The motif approach has been extended to combined network of transcriptional regulation and protein–protein interactions [169] in yeast, as well as to other systems [101].

At the risk of being overly pessimistic, we should conclude this section with a note of caution. It would be attractive to think that a decade of work on network topology has resulted in a coherent picture, perhaps of the following form: on the smallest scale, the nodes of biological networks are assembled into motifs, these in turn are linked into modules, and this continues in a hierarchical fashion until the entire network is scale free. As we will discuss again in the context of design principles, the notion of such discrete substructure – motifs and modules – is intuitively appealing, and some discussions suggest that it is essential either for the function or the evolution of networks. On the other hand, the evidence for such structure usually is gathered with reference to some null model (e. g., a random network with the same $P(k)$), so we don't even have an absolute definition of these structures, much less a measure of their sufficiency as a characterization of the whole system; for attempts at an absolute definition of modularity see Ziv et al. [174] and Hofman and Wiggins [62]. Similarly, while it is appealing to think about scale free networks, the evidence for scaling almost always is confined to less than two decades, and in practice scaling often is not exact. It is then not clear whether the idealization of scale invariance captures the essential structure in these systems.

Boolean Networks

A straightforward extension of the topological picture that also permits the study of network dynamics assumes that the entities at the nodes – for example, genes or signaling proteins – are either 'on' or 'off' at each moment of time, so that for node i the state at time t is $\sigma_i(t) \in \{0, 1\}$. Time is usually discretized, and an additional prescription is needed to implement the evolution of the system: $\sigma_i(t+1) = f_i(\{\sigma_\mu(t)\})$, where f_i is a function that specifies how the states of the nodes μ that are the inputs to node i in the interaction graph combine to determine the next state at node i. For instance, f_A might be a Boolean function for gene A, which needs to have its activator gene B present and repressor gene C absent, so that $\sigma_A(t+1) = \sigma_B(t) \wedge \bar{\sigma}_C(t)$. Alternatively, f might be a function that sums the inputs at state t with some weights, and then sets $\sigma_i = 1(0)$ if the result is above (below) a threshold, as in classical models of neural networks.

Boolean networks are amenable both to analytical treatment and to efficient simulation. Early on, Kauffman [80] considered the family of random boolean networks. In these models, each node is connected at random to K other nodes on average, and it computes a random Boolean function of its inputs in which a fraction ρ of the 2^K possible input combinations leads to $\sigma_i(t+1) = 1$. In the limit that the network is large, the dynamics are either regular (settling into a stable fixed cycle) or chaotic, and these two phases are separated by a separatrix $2\rho(1-\rho)K = 1$ in the phase space (ρ, K).

Aldana and Cluzel [4] have shown that for connectivities of $K \sim 20$ that could reasonably be expected in e. g. transcriptional regulatory networks, the chaotic regime dominates the phase space. They point out, however, that if the network is scale free, there is no 'typical' K as the distribution $P(k) \sim k^{-\gamma}$ does not have a well-defined mean for $\gamma \leq 3$ and the phase transition criterion must be restated. It turns out, surprisingly, that regular behavior is possible for values of γ between 2 and 2.5, observed in most biological networks, and this is exactly the region where the separatrix lies. Scale free architecture, at least for Boolean networks, seems to prevent chaos.

Several groups have used Boolean models to look at specific biological systems. Thomas [150] has established a theoretical framework in which current states of the genes (as well as the states in the immediate past) and the environmental inputs are represented by Boolean variables that evolve through the application of Boolean functions. This work has been continued by, for example, Sanchez and Thieffry [128] who analyzed the gap-gene system of the fruit fly *Drosophila* by building a Boolean network that generates the correct levels of gene expression for 4 gap genes in response to input levels of 3 maternal morphogens with spatially varying profiles stretched along the anterior-posterior axis of the fly embryo. Interestingly, to reproduce the observed results and correctly predict the known *Drosophila* segmentation mutants, the authors had to introduce generalized Boolean variables that can take more than two states, and have identified the smallest necessary number of such states for each gene.

In a similar spirit, Li et al. [91] studied the skeleton of the budding yeast cell cycle, composed of 11 nodes, and a thresholding update rule. They found that the topology of this small network generates a robust sequence of transitions corresponding to known progression through yeast cell-cycle phases G1 (growth), S (DNA duplication), G2 (pause) and M (division), triggered by a known 'cell-size checkpoint.' This progression is robust, in the sense that the correct trajectory is the biggest dynamical attractor of the system, with respect to various choices of update rules and parameters, small changes in network topology, and choice of triggering checkpoints.

The usefulness of Boolean networks stems from the relative ease of implementation and simple parametrization of network topology and dynamics, making them suitable for studying medium or large networks. In addition to simplifying the states at the nodes to two (or more) discrete levels, which is an assumption that has not been clearly explored, one should be cautious that the discrete and usually synchronous dynamics in time can induce unwanted artifacts.

Probabilistic Models

Suppose one is able to observe simultaneously the activity levels of several proteins comprising a signaling network, or the expression levels of a set of genes belonging to the same regulatory module. Because they are part of a functional whole, the activity levels of the components will be correlated. Naively, one could build a network model by simply computing pairwise correlation coefficients between pairs of nodes, and postulating an interaction, and therefore a link, between the two nodes whenever their correlation is above some threshold. However, in a test case where A → B → C (gene A induces B which induces C), one expects to see high positive correlation among all three elements, even though there is no (physical) interaction between A and C. Correlation therefore is not equal to interaction or causation. Constructing a network from the correlations in this naive way also does not lead to a generative model that would predict the probabilities for observing different states of the network as a whole. Another approach is clearly needed; see Markowetz and Spang [94] for a review.

In the simple case where the activity of a protein/gene i can either be 'on' ($\sigma_i = 1$) or 'off' ($\sigma_i = 0$), the state of a network with N nodes will be characterized by a binary word of N bits, and because of interaction between nodes, not all these words will be equally likely. For example, if node A represses node B, then combinations such as $1_A 0_B \ldots$ or $0_A 1_B \ldots$ will be more likely than $1_A 1_B \ldots$. In the case of deterministic Boolean networks, having node A be 'on' would imply that node B is 'off' with certainty, but in probabilistic models it only means that there is a positive *bias* for node B to be 'off,' quantified by the probability that node B is 'off' given that the state of node A is known. Having this additional probabilistic degree of freedom is advantageous, both because the network itself might be noisy, and because the experiment can induce errors in the signal readout, making the inference of deterministic rules from observed binary patterns an ill-posed problem.

Once we agree to make a probabilistic model, the goal is to find the distribution over all network states, which we can also think of as the joint distribution of all the N variable that live on the nodes of the network, $P(\sigma_1, \ldots, \sigma_N | C)$, perhaps conditioned on some fixed set of environmental or experimental factors C. The activities of the nodes, σ_i, can be binary, can take on a discrete set of states, or be continuous, depending on our prior knowledge about the system and experimental and numerical constraints. Even for a modest N, experiments of realistic scale will not be enough to directly estimate the probability distribution, since even with binary variable the number of possible states, and hence the number of parameters required to specify the general probability distribution, grows as $\sim 2^N$. Progress thus depends in an essential way on simplifying assumptions.

Returning to the three gene example A → B → C, we realize that C depends on A only through B, or in other words, C is *conditionally independent* of A and hence no interaction should be assigned between nodes A and C. Thus, the joint distribution of three variables can be factorized,

$$P(\sigma_\mathrm{A}, \sigma_\mathrm{B}, \sigma_\mathrm{C}) = P(\sigma_\mathrm{C} | \sigma_\mathrm{B}) P(\sigma_\mathrm{B} | \sigma_\mathrm{A}) P(\sigma_\mathrm{A}) .$$

One might hope that, even in a large network, these sorts of conditional independence relations could be used to simplify our model of the probability distribution. In general this doesn't work, because of feedback loops which, in our simple example, would include the possibility that C affects the state of A, either directly or through some more circuitous path. Nonetheless one can try to make an approximation in which loops either are neglected or (more sensibly) taken into account in some sort of average way; in statistical mechanics, this approximation goes back at least to the work of Bethe [19].

In the computer science and bioinformatics literature, the exploitation of Bethe-like approximations has come to be known as 'Bayesian network modeling' [43]. In practice what this approach does is to search among possible network topologies, excluding loops, and then for fixed topology one uses the conditional probability relationships to factorize the probability distribution and fit the tables of conditional probabilities at each node that will best reproduce some set of data. Networks with more links have more parameters, so one must introduce a trade-off between the quality of the fit to the data and this increasing complexity. In this framework there is thus an explicit simplification based on conditional independence, and an implicit simplification based on a preference for models with fewer links or sparse connectivity.

The best known application of this approach to a biological network is the analysis of the MAPK signaling pathway in T cells from the human immune system [127]. The data for this analysis comes from experiments in which the phosophorylated states of 11 proteins in the pathway are sampled simultaneously by immunostaining [113], with hundreds of cells sampled for each set of external conditions. By combining experiments from multiple conditions, the Bayesian network analysis was able to find a network of interactions among the 11 proteins that has high overlap with those known to occur experimentally.

A very different approach to simplification of probabilistic models is based on the maximum entropy principle [71]. In this approach one views a set of experiments as providing an estimate of some set of correlations, for example the $\sim N^2$ correlations among all pairs of elements in the network. One then tries to construct a probability distribution which matches these correlations but otherwise has as little structure – as much entropy – as possible. We recall that the Boltzmann distribution for systems in thermal equilibrium can be derived as the distribution which has maximum entropy consistent with a given average energy, and maximum entropy modeling generalizes this to take account of other average properties. In fact one can construct a hierarchy of maximum entropy distributions which are consistent with higher and higher orders of correlation [130]. Maximum entropy models for binary variables that are consistent with pairwise correlations are exactly the Ising models of statistical physics, which opens a wealth of analytic tools and intuition about collective behavior in these systems.

In the context of biological networks (broadly construed), recent work has shown that maximum entropy models consistent with pairwise correlations are surprisingly successful at describing the patterns of activity among populations of neurons in the vertebrate retina as it responds to natural movies [131,153]. Similar results are obtained for very different retinas under different conditions [137], and these successes have touched a flurry of interest in the analysis of neural populations more generally. The connection to the Ising model has a special resonance in the context of neural networks, where the collective behavior of the Ising model has been used for some time as a prototype for thinking about the dynamics of computation and memory storage [64]; in the maximum entropy approach the Ising model emerges directly as the least structured model consistent with the experimentally measured patterns of correlation among pairs of cells. A particularly striking result of this analysis is that the Ising models which emerge seem to be poised near a critical point [153]. Returning to cell biology, the maximum entropy approach has also been used to analyze patterns of gene expression in yeast [90] as well as to revisit the MAPK cascade [151].

Dynamical Systems

If the information about a biological system is detailed enough to encompass all relevant interacting molecules along with the associated reactions and estimated reaction rates, and the molecular noise is expected to play a negligible role, it is possible to describe the system with rate equations of chemical kinetics. An obvious benefit is the immediate availability of mathematical tools, such as steady state and stability analyses, insight provided by nonlinear dynamics and chaos theory, well developed numerical algorithms for integration in time and convenient visualization with phase portraits or bifurcation diagrams. Moreover, analytical approximations can be often exploited productively when warranted by some prior knowledge, for example, in separately treating 'fast' and 'slow' reactions. In practice, however, reaction rates and other important parameters are often unknown or known only up to order-of-magnitude estimations; in this case the problem usually reduces to the identification of phase space regions where the behavior of the system is qualitatively the same, for example, regions where the system exhibits limit-cycle oscillations, bistability, convergence into a single steady state etc.; see Tyson et al. [159] for a review. Despite the difficulties, deterministic chemical kinetic models have been very powerful tools in analyzing specific network motifs or regulatory elements, as in the protein signaling circuits that achieve perfect adaptation, homeostasis, switching and so on, described by Tyson et al. [160], and more generally in the analysis of transcriptional regulatory networks as reviewed by Hasty et al. [59].

In the world of bacteria, some of the first detailed computer simulations of the chemotaxis module of *Escherichia coli* were undertaken by Bray et al. [29]. The signaling cascade from the Tar receptor at the cell surface to the modifications in the phosphorylation state of the molecular motor were captured by Michaelis–Menten kinetic reactions (and equilibrium binding conditions for the receptor), and the system of equations was numerically integrated in time. While slow adaptation kinetics was not studied in this first effort, the model nevertheless qualitatively reproduces about 80 percent of examined chemotactic protein deletion and overexpression mutants, although the extreme sensitivity of the system remained unexplained.

In eukaryotes, Novak and Tyson [107] have, for instance, constructed an extensive model of cell cycle control in fission yeast. Despite its complexity (~ 10 proteins and ~ 30 rate constants), Novak and colleagues have provided an interpretation of the system in terms of three interlocking modules that regulate the transitions from G1 (growth) into S (DNA synthesis) phase, from G2 into M (division) phase, and the exit from mitosis, respectively. The modules are coupled through cdc2/cdc13 protein complex and the system is driven by the interaction with the cell size signal (proportional to the number of ribosomes per nucleus). At small size, the control circuit can only support one stable attractor, which is the state with low cdc2 activity corresponding to G1 phase. As the cell grows, new stable state appears and the system makes an irreversible transition into S/G2 at a bifurcation point, and, at an even larger size, the mitotic module becomes unstable and executes limit cycles in cdc2-cdc13 activity until the M phase is completed and the cell returns to its initial size. The basic idea is that the cell, driven by the the size readout, progresses through robust cell states created by bistability in the three modules comprising the cell cycle control – in this way, once it commits to a transition from G2 state into M, small fluctuations will not flip it back into G2. The mathematical model has in this case successfully predicted the behaviors of a number of cell cycle mutants and recapitulated experimental observations collected during 1970s and 1980s by Nurse and collaborators [108].

The circadian clock is a naturally occurring transcriptional module that is particularly amenable to dynamical systems modeling. Leloup and Goldbeter [87] have created a mathematical model of a mammalian clock (with ~ 20 rate equations) that exhibits autonomous sustained oscillations over a sizable range of parameter values, and reproduces the entrainment of the oscillations to the light–dark cycles through light-induced gene expression. The basic mechanism that enables the cyclic behavior is negative feedback transcriptional control, although the actual circuit contains at least two coupled oscillators. Studying circadian clock in mammals, the fruit fly *Drosophila* or *Neurospora* is attractive because of the possibility of connecting a sizable catalog of physiological disorders in circadian rhythms to malfunctions in the clock circuit and direct experimentation with light-dark stimuli [171]. Recent experiments indicate that at least in cyanobacteria the circadian clock can be reconstituted from a surprisingly small set of biochemical reactions, without transcription or translation [102,157], and this opens possibilities for even simpler and highly predictive dynamical models [126].

Dynamical modeling has in addition been applied to many smaller systems. For example, the construction of a synthetic toggle switch [44], and the 'repressilator' – oscillating network of three mutually repressing genes [38] – are examples where mathematical analysis has stimulated the design of synthetic circuits. A successful reaction-diffusion model of how localization and complex formation of Min proteins can lead to spatial limit cycle oscillations (used by *Escherichia coli* to find its division site) was constructed by Huang et al. [65]. It remains a challenge, nevertheless, to navigate in the space of parameters as it becomes ever larger for bigger networks, to correctly account for localization and count various forms of protein modifications, especially when the signaling networks also couple to transcriptional regulation, and to find a proper balance between models that capture all known reactions and interactions and phenomenological models that include coarse-grained variables.

Stochastic Dynamics

Stochastic dynamics is in principle the most detailed level of system description. Here, the (integer) count of every molecular species is tracked and reactions are drawn at random with appropriate probabilities per unit time (proportional to their respective reaction rates) and executed to update the current tally of molecular counts. An algorithm implementing this prescription, called the stochastic simulation algorithm or SSA, was devised by Gillespie [47]; see Gillespie [48] for a review of SSA and a discussion of related methods. Although slow, this approach for simulating chemical reactions can be made exact. In general, when all molecules are present in large numbers and continuous, well-mixed concentrations are good approximations, the (deterministic) rate dynamics equations and stochastic simulation give the same results; however, when molecular counts are low and, consequently, the stochasticity in reaction timing and ordering becomes important, the rate dynamics breaks down and SSA needs to be used. In biological networks and specifically in transcriptional regulation, a gene and its promoter region are only present in one (or perhaps a few) copies, while transcription factors that regulate it can also be at nanomolar concentrations (i. e. from a few to a few hundred molecules per nucleus), making stochastic effects possibly very important [97,98].

One of the pioneering studies of the role of noise in a biological system was a simulation of the phage λ lysis-lysogeny switch by Arkin et al. [7]. The life cycle of the phage is determined by the concentrations of two transcription factors, *cI* (lambda repressor) and *cro*, that compete for binding to the same operator on the DNA. If *cI* is prevalent, the phage DNA is integrated into the host's genome and no phage genes except for *cI* are expressed (the lysogenic state); if *cro* is dominant, the phage is in lytic state, using cell's DNA replication machinery to produce more phages and ultimately lyse the host cell [115]. The switch is bistable and the fate of the phage depends on the temporal and random pattern of gene expression of two mutually antagonistic transcription factors, although the balance can be shifted by subjecting the host cell to stress and thus flipping the toggle into lytic phase. The stochastic simulation correctly reproduces the experimentally observed fraction of lysogenic phages as a function of multiplicity-of-infection. An extension of SSA to spatially extended models is possible.

Although the simulations are exact, they are computationally intensive and do not offer any analytical insight into the behavior of the solutions. As a result, various theoretical techniques have been developed for studying the effects of stochasticity in biological networks. These are often operating in a regime where the deterministic chemical kinetics is a good approximation, and noise (i. e. fluctuation of concentrations around the mean) is added into the system of differential equations as a perturbation; these Langevin methods have been useful for the study of noise propagation in regulatory networks [76,111,149]. The analysis of stochastic dynamics is especially interesting in the context of design principles which consider the reliability of network function, to which we return below.

Network Properties and Operating Principles

Modularity

Biological networks are said to be modular, although the term has several related but nevertheless distinct meanings. Their common denominator is the idea that there exist a partitioning of the network nodes into groups, or modules, that are largely independent of each other and perform separate or autonomous functions. Independence can be achieved through spatial isolation of the module's processes or by chemical specificity of its components. The ability to extract the module from the cell and reconstitute it in vitro, or transplant it to another type of cell is a powerful argument for the existence of modularity [58]. In the absence of such strong and laborious experimental verifications, however, measures of modularity that depend on a particular network model are frequently used.

In topological networks, the focus is on the module's independence: nodes within a module are densely connected to each other, while inter-modular links are sparse [57,120,142] and the tendency to cluster is measured by high clustering coefficients. As a caveat to this view note that despite their sparseness the inter-module links could represent strong dynamical couplings. Modular architecture has been studied in Boolean networks by Kashtan and Alon [79], who have shown that modularity can evolve by mutation and selection in a time-varying fitness landscape where changeable goals decompose into a set of fixed subproblems. In the example studied they computationally evolve networks implementing several Boolean formulae and observe the appearance of a module – a circuit of logical gates implementing a particular Boolean operator (like XOR) in a reusable way. This work makes clear that modularity in networks is plausibly connected to modularity in the kinds of problems that these networks were selected to solve, but we really know relatively little about the formal structure of these problems.

There are also ways of inferring a form of modularity directly without assuming any particular network model. Clustering tools partition genes into co-expressed groups, or clusters, that are often identified with particular modules [36,133,140]. Ihmels et al. [67] have noted that each node can belong to more than one module depending on the biological state of the cell, or the context, and have correspondingly reexamined the clustering problem. Elemento et al. [37] have recently presented a general information theoretic approach to inferring regulatory modules and the associated transcription factor binding sites from various kinds of high-throughput data. While clustering methods have been widely applied in the exploration of gene expression, it should be emphasized that merely finding clusters does not by itself provide evidence for modularity. As noted above, the whole discussion would be much more satisfying if we had independent definitions of modularity and, we might add, clearly stated alternative hypotheses about the structure and dynamics of these networks.

Focusing on the functional aspect of the module, we often observe that the majority of the components of a system (for instance, a set of promoter sites or a set of genes regulating motility in bacteria) are conserved together across species. These observations support the hypothesis that the conserved components are part of a very tightly coupled sub-network which we might identify as a module. Bioinformatic tools can then use the combined sequence and expression data to give predictions about modules, as reviewed by Siggia et al. [139]. Purely phylogenetic approaches that infer module components based on inter-species comparisons have also been productive and can extract candidate modules based only on phylogenetic footprinting, that is, studying the presence or absence of homologous genes across organisms and correlating their presence with hand annotated phenotypic traits [141].

Robustness

Robustness refers to a property of the biological network such that some aspect of its function is not sensitive to perturbations of network parameters, environmental variables (e. g. temperature), or initial state; see de Visser et al. [162] for a review of robustness from an evolutionary perspective and Goulian [53] for mechanisms of robustness in bacterial circuits. Robustness encompasses two very different ideas. One idea has to do with a general principle about the nature of explanation in the quantitative sciences: qualitatively striking facts should not depend on the fine tuning of parameters, because such a scenario just shifts the problem to understanding why the parameters are tuned as they are. The second idea is more intrinsic to the function of the system, and entails the hypothesis that cells cannot rely on precisely reproducible parameters or conditions and must nonetheless function reliably and reproducibly.

Robustness has been studied extensively in the chemotactic system of the bacterium *Escherichia coli*. The systematic bias to swim towards chemoattractants and away from repellents can only be sustained if the bacterium is sensitive to the spatial gradients of the concentration and not to its absolute levels. This discriminative ability is ensured by the mechanism of perfect adaptation, with which the proportion of bacterial straight runs and tumbles (random changes in direction) always returns to the same value in the absence of gradients [27]. Naively, however, the ability to adapt perfectly seems to be sensitive to the amounts of intracellular signaling proteins, which can be tuned only approximately by means of transcriptional regulation. Barkai and Leibler [13] argued that there is integral feedback control in the chemotactic circuit which makes it robust against changes in these parameters, and Alon et al. [6] showed experimentally that precision of adaptation truly stays robust, while other properties of the systems (such as the time to adapt and the steady state) show marked variations with intracellular signaling protein concentrations.

One seemingly clear example of robust biological function is embryonic development. We know that the spatial structure of the fully developed organism follows a 'blueprint' laid out early in development as a spatial pattern of gene expression levels. von Dassow et al. [34] studied one part of this process in the fruit fly *Drosophila*, the 'segment polarity network' that generates striped patterns of expression. They considered a dynamical system based on the wiring diagram of interactions among a small group of genes and signaling molecules, with ~ 50 associated constants parametrizing production and degradation rates, saturation response and diffusion, and searched the parameter space for solutions that reproduce the known striped patterns. They found that, with their initial guess at network topology, such solutions do not exist, but adding a particular link – biologically motivated though unconfirmed at the time – allowed them to find solutions by random sampling of parameter space. Although they presented no rigorous measure for the volume of parameter space in which correct solutions exist, it seems that a wide variety of parameter choices and initial conditions indeed produce striped expression patterns, and this was taken to be a signature of robustness.

Robustness in dynamical models is the ability of the biological network to sustain its trajectory through state space despite parameter or state perturbations. In circadian clocks the oscillations have to be robust against both molecular noise inherent in transcriptional regulation, examined in stochastic simulations by Gonze et al. [52], as well as variation in rate parameters [143]; in the latter work the authors introduce integral robustness measures along the trajectory in state space and argue that the clock network architecture tends to concentrate the fragility to perturbations into parameters that are global to the cell (maximum overall translation and protein degradation rates) while increasing the robustness to processes specific to the circadian oscillator. As was mentioned earlier, robustness to state perturbations was demonstrated by Li et al. [91] in the threshold binary network model of the yeast cell cycle, and examined in scale-free random Boolean networks by Aldana and Cluzel [4].

As with modularity, robustness has been somewhat resistant to rigorous definitions. Importantly, robustness has always been used as a relational concept: function X is robust to variations in Y. An alternative to robustness is for the organism to exert precise control over Y, perhaps even using X as a feedback signal. This seems to be how neurons stabilize a functional mix of different ion channels [93], following the original theoretical suggestion of LeMasson et al. [88]. Pattern formation during embryonic development in *Drosophila* begins with spatial gradients of transcription factors, such as Bicoid, which are established by maternal gene expression, and it has been assumed that variations in these expression levels are inevitable, requiring some robust readout mechanism. Recent measurements of Bicoid in live embryos, however, demonstrate that the absolute concentrations are actually reproducible from embryo to embryo with $\sim 10\%$ precision [54]. While there remain many open questions, these results suggest that organisms may be able to exert surprisingly exact control over critical parameters, rather than having compensation schemes for initially sloppy mechanisms. The example of ion channels alerts us to the possibility that cells may even 'know' which combinations of parameters are critical, so that variations in a multidimensional parameter space are large, but confined to a low dimensional manifold.

Noise

A dynamical system with constant reaction rates, starting repeatedly from the same initial condition in a stable environment, always follows a deterministic time evolution. When the concentrations of the reacting species are low enough, however, the description in terms of time (and possibly space) dependent concentration breaks down, and the stochasticity in reactions, driven by random encounters between individual molecules, becomes important: on repeated trials from the same initial conditions, the system will trace out different trajectories in the state space. As has been pointed out in the section on stochastic dynamics, biological networks in this regime need to be simulated with the Gillespie algorithm [47], or analyzed within approximate schemes that treat noise as perturbation of deterministic dynamics. Recent experimental developments have made it possible to observe this noise directly, spurring new research in the field. Noise in biological networks fundamentally limits the organism's ability to sense, process and respond to environmental and internal signals, suggesting that analysis of noise is a crucial component in any attempt to understand the design of these networks. This line of reasoning is well developed in the context of neural function [20], and we draw attention in particular to work on the ability of the visual system to count single photons, which depends upon the precision of the G-protein mediated signaling cascade in photo receptors; see, for example, [117].

Because transcriptional regulation inherently deals with molecules, such as DNA and transcription factors, that are present at low copy numbers, most noise studies were carried out on transcriptional regulatory elements. The availability of fluorescent proteins and their fusions to wild type proteins have been the crucial tools, enabling researchers to image the cells expressing these probes in a controllable manner, and track their number in time and across the population of cells. Elowitz et al. [39] pioneered the idea of observing the output of two identical regulatory elements driving the expression of two fluorescent proteins of different colors, regulated by a common input in a single *Escherichia coli* cell. In this 'two-color experiment,' the correlated fluctuations in both colors must be due to the *extrinsic* fluctuations in the common factors that influence the production of both proteins, such as overall RNA polymerase or transcription factor levels; on the other hand, the remaining, uncorrelated fluctuation is due to the *intrinsic* stochasticity in the transcription of the gene and translation of the messenger RNA into the fluorescent protein from each of the two promoters [147]. Ozbudak et al. [109] have studied the contributions of stochasticity in transcription and translation to the total noise in gene expression in prokaryotes, while Pedraza and van Oudenaarden [112] and Hooshangi et al. [63] have looked at the propagation of noise from transcription factors to their targets in synthetic multi-gene cascades. Rosenfeld et al. [124] have used the statistics of binomial partitioning of proteins during the division of *Escherichia coli* to convert their fluorescence measurements into the corresponding absolute protein concentrations, and also were able to observe the dynamics of these fluctuations, characterizing the correlation times of both intrinsic and extrinsic noise.

Theoretical work has primarily been concerned with disentangling and quantifying the contributions of different steps in transcriptional regulation and gene expression to the total noise in the regulated gene [111,146,149], often by looking for signatures of various noise sources in the behavior of the measured noise as a function of the mean expression level of a gene. For many of the examples studied in prokaryotes, noise seemed to be primarily attributable to the production of proteins in bursts from single messenger RNA molecules, and to pulsatile and random activation of genes and therefore bursty translation into mRNA [50]. In yeast [26,119] and in mammalian cells [116] such stochastic synthesis of mRNA was modeled and observed as well. Simple scaling of noise with the mean was observed in ∼40 yeast proteins under different conditions by Bar-Even et al. [9] and interpreted as originating in variability in mRNA copy numbers or gene activation.

Bialek and Setayeshgar [22] have demonstrated theoretically that at low concentrations of transcriptional regulator, there should be a lower bound on the noise set by the basic physics of diffusion of transcription factor molecules to the DNA binding sites. This limit is independent of (possibly complex, and usually unknown) molecular details of the binding process; as an example, cooperativity enhances the 'sensitivity' to small changes in concentration, but doesn't lower the physical limit to noise performance [23]. This randomness in diffusive flux of factors to their 'detectors' on the DNA must ultimately limit the precision and reliability of transcriptional regulation, much like the randomness in diffusion of chemoattractants to the detectors on the surface of *Escherichia coli* limits its chemotactic performance [17]. Interestingly, one dimensional diffusion of transcription factors along the DNA can have a big impact on the speed with which TFs find their targets, but the change in noise performance that one might expect to accompany these kinetic changes is largely compensated by the extended correlation structure of one dimensional diffusion [152]. Recent measurements of the regulation of the *hunchback* gene by Bicoid during early fruit fly development by Gregor et al. [54] have provided evidence for the dominant role of such input noise, which coexists with previously studied output noise in production of mRNA and protein [156]. These results raise the possibility that initial decisions in embryonic development are made with a precision limited by fundamental physical principles.

Dynamics, Attractors, Stability and Large Fluctuations

The behavior of a dynamical system as the time tends to infinity, in response to a particular input, is interesting regardless of the nature of the network model. Both discrete and continuous, or deterministic and noisy, systems can settle into a number of fixed points, exhibit limit-cycle oscillations, or execute chaotic dynamics. In biological networks it is important to ask whether these qualitatively different outcomes correspond to distinct phenotypes or behaviors. If so, then a specific stable gene expression profile in a network of developmental genes, for example, encodes that cell's developmental fate, as the amount of lambda repressor encodes the state of lysis vs lysogeny switch in the phage. The history of the system that led to the establishment of a specific steady state would not matter as long as the system persisted in the same attractor: the dynamics could be regarded as a 'computation' leading to the final result, the identity of the attractor, with the activities of genes in this steady state in turn driving the downstream pathways and other modules; see Kauffman [80] for genetic networks and Hopfield [64] for similar ideas in neural networks for associative memory. Alternatively, such partitioning into transient dynamics and 'meaningful' steady states might not be possible: the system must be analyzed as a whole while it moves in state space, and parts of it do not separately and sequentially settle into their attractors.

It seems, for example, that qualitative behavior of the cell cycle can be understood by progression through well-defined states or checkpoints: after transients die away, the cell cycle proteins are in a 'consistent' state that regulates division or growth related activities, so long as the conditions do not warrant a new transition into the next state [33,103]. In the fruit fly *Drosophila* development it has been suggested that combined processes of diffusion and degradation first establish steady-state spatial profiles of maternal morphogens along the major axis of the embryo, after which this stable 'coordinate system' is read out by gap and other downstream genes to generate the body segments. Recent measurements by Gregor et al. [55] have shown that there is a rich dynamics in the Bicoid morphogens concentration, prompting Bergmann et al. [18] to hypothesize that perhaps downstream genes read out and respond to morphogens even before the steady state has been reached. On another note, an interesting excitable motif, called the "feedback resistor," has been found in HIV Tat system – instead of having a bistable switch like the λ phage, HIV (which lacks negative feedback capability) implements a circuit with a single stable 'off' lysogenic state, that is perturbed in a pulse of trans activation when the virus attacks. The pulse probably triggers a threshold-crossing process that drives downstream events, and subsequently decays away; the feedback resistor is thus again an example of a dynamic, as opposed to the steady-state, readout [168]. Excitable dynamics are of course at the heart of the action potential in neurons, which results from the coupled dynamics of ion channel proteins, and related dynamical ideas are now emerging other cellular networks [145].

If attractors of the dynamical system correspond to distinct biological states of the organism, it is important to examine their stability against noise-induced spontaneous flipping. Bistable elements are akin to the 'flip-flop' switches in computer chips – they form the basis of cellular (epigenetic) memory. While this mechanism for remembering the past is not unique – for example, a very slow, but not bistable, dynamics will also retain 'memory' of the initial condition through protein levels that persist on a generation time scale [138], it has the potential to be the most stable mechanism. The naturally occurring bistable switch of the λ phage was studied using stochastic simulation by Arkin et al. [7], and a synthetic toggle switch was constructed in *Escherichia coli* by Gardner et al. [44]. Theoretical studies of systems where large fluctuations are important are generally difficult and restricted to simple regulatory elements, but Bialek [21] has shown that a bistable switch can be created with as few as tens of molecules yet remain stable for years. A full understanding of such stochastic switching brings in powerful methods from statistical physics and field theory [122,129,164], ultimately with the hope of connecting to quantitative experiments [1].

Optimization Principles

If the function of a pathway or a network module can be quantified by a scalar measure, it is possible to explore the space of networks that perform the given function optimally. An example already given was that of maximizing the growth rate of the bacterium *Escherichia coli*, subject to the constraints imposed by the known metabolic reactions of the cell; the resulting optimal joint usage of oxygen and food could be compared to the experiments [66]. If enough constraints exist for the problem to be well posed, and there is sufficient reason to believe that evolution drove the organism towards optimal behavior, optimization principles allow us to both tune the otherwise unknown parameters to achieve the maximum, and also to compare the wild type and optimal performances.

Dekel and Alon [35] have performed the cost/benefit analysis of expressing *lac* operon in bacteria. On one hand *lac* genes allow *Escherichia coli* to digest lactose, but on the other there is the incurred metabolic cost to the cell for expressing them. That the cost is not negligible to the bacterium is demonstrated best by the fact that it shuts off the operon if no lactose is present in the environment. The cost terms are measured by inducing the *lac* operon with changeable amount of IPTG that provides no energy in return; the benefit is measured by fully inducing *lac* with IPTG and supplying variable amounts of lactose; both cost and benefit are in turn expressed as the change in the growth rate compared to the wild-type grown at fixed conditions. Optimal levels of *lac* expression were then predicted as a function of lactose concentration and bacteria were evolved for several hundred generations to verify that evolved organisms lie close to the predicted optimum.

Zaslaver et al. [172] have considered a cascade of amino-acid biosynthesis reactions in *Escherichia coli*, catalyzed by their corresponding enzymes. They have then optimized the parameters of the model that describes the regulation of enzyme gene expression, such that the total metabolic cost for enzyme production was balanced against the benefit of achieving a desired metabolic flux through the biosynthesis pathway. The resulting optimal on-times and promoter activities for the enzymes were compared to the measured activities of amino-acid biosynthesis promoters exposed to different amino-acids in the medium. The authors conclude that the bacterium implements a 'just-in-time' transcription program, with enzymes catalyzing initial steps in the pathway being produced from strong and low-latency promoters.

In signal transduction networks the definition of an objective function to be maximized is somewhat more tricky. The ability of the cell to sense its environment and make decisions, for instance about which genes to up- or down-regulate, is limited by several factors: scarcity of signals coming from the environment, perhaps because of the limited time that can be dedicated to data collection; noise inherent in the signaling network that degrades the quality of the detected signal; (sub-)optimality of the decision strategy; and noise in the effector systems at the output. A first idea would be to postulate that networks are designed to lower the noise, and intuitively the ubiquity of mechanisms such as negative feedback [15,53] is consistent with such an objective. There are various definitions for noise, however, which in addition are generally a function of the input, raising serious issues about how to formulate a principled optimization criterion.

When we think about energy flow in biological systems, there is no doubt that our thinking must at least be consistent with thermodynamics. More strongly, thermodynamics provides us with notions of efficiency that place the performance of biological systems on an absolute scale, and in many cases this performance really is quite impressive. In contrast, most discussions of information in biological systems leave "information" as a colloquial term, making no reference to the formal apparatus of information theory as developed by Shannon and others more than fifty years ago [135]. Although many aspects of information theory that are especially important for modern technology (e. g., sophisticated error-correcting codes) have no obvious connection to biology, there is something at the core of information theory that is vital: Shannon proved that if we want to quantify the intuitive concept that "x provides information about y," then there is only one way to do this that is guaranteed to work under all conditions and to obey simple intuitive criteria such as the additivity of independent information. This unique measure of "information" is Shannon's mutual information. Further, there are theorems in information theory which, in parallel to results in thermodynamics, provide us with limits to what is possible and with notions of efficiency.

There is a long history of using information theoretic ideas to analyze the flow of information in the nervous system, including the idea that aspects of the brain's coding strategies might be chosen to optimize the efficiency of coding, and these theoretical ideas have led directly to interesting experiments. The use of information to think about cellular signaling and its possible optimization is more recent [154,175]. An important aspect of optimizing information flow is that the input/output relations of signaling devices must be matched to the distribution of inputs, and recent measurements on the control of *hunchback* by Bicoid in the early fruit fly embryo [54] seem remarkably consistent with the (parameter free) predictions from these matching relations [155].

In the context of neuroscience there is a long tradition of forcing the complex dynamics of signal processing into a setting where the subject needs to decide between a small set of alternatives; in this limit there is a well developed theory of optimal Bayesian decision making, which uses prior knowledge of the possible signals to help overcome noise intrinsic to the signaling system; Libby et al. [92] have recently applied this approach to the *lac* operon in *Escherichia coli*. The regulatory element is viewed as an inference module that has to 'decide,' by choosing its induction level, if the environmental lactose concentration is high or low. If the bacterium detects a momentarily high sugar concentration, it has to discriminate between two situations: either the environment really is at low overall concentration but there has been a large fluctuation; or the environment has switched to a high concentration mode. The authors examine how plausible regulatory element architectures (e. g. activator vs repressor, cooperative binding etc.) yield different discrimination performance. Intrinsic noise in the *lac* system can additionally complicate such decision making, but can be included into the theoretical Bayesian framework.

The question of whether biological systems are optimal in any precise mathematical sense is likely to remain controversial for some time. Currently opinions are stronger than the data, with some investigators using 'optimized' rather loosely and others convinced that what we see today is only a historical accident, not organizable around such lofty principles. We emphasize, however, that attempts to formulate optimization principles require us to articulate clearly what we mean by "function" in each context, and this is an important exercise. Exploration of optimization principles also exposes new questions, such as the nature of the distribution of inputs to signaling systems, that one might not have thought to ask otherwise. Many of these questions remain as challenges for a new generation of experiments.

Evolvability and Designability

Kirschner and Gerhart [82] define *evolvability* as an organism's capacity to generate heritable phenotypic variation. This capacity may have two components: first, to reduce the lethality of mutations, and second, to reduce the number of mutations needed to produce phenotypically novel traits. The systematic study of evolvability is hard because the genotype-to-phenotype map is highly nontrivial, but there have been some qualitative observations relevant to biological networks. Emergence of *weak linkage* of processes, such as the co-dependence of transcription factors and their DNA binding sites in metazoan transcriptional regulation, is one example. Metazoan regulation seems to depend on combinatorial control by many transcription factors with weak DNA-binding specificities and the corresponding binding sites (called cis-regulatory modules) can be dispersed and extended on the DNA. This is in stark contrast to the strong linkage between the factors and the DNA in prokaryotic regulation or in metabolism, energy transfer or macromolecular assembly, where steric and complementarity requirements for interacting molecules are high. In protein signaling networks, strongly conserved but flexible proteins, like calmodulin, can bind weakly to many other proteins, with small mutations in their sequence probably affecting such binding and making the establishment of new regulatory links possible and perhaps easy.

Some of the most detailed attempts to follow the evolution of network function have been by Francois and coworkers [41,42]. In their initial work they showed how simple functional circuits, performing logical operations or implementing bistable or oscillatory behavior, can be reliably created by a mutational process with selection by an appropriate fitness function. More recently they have considered fitness functions which favor spatial structure in patterns of gene expression, and shown how the networks that emerge from dynamics in this fitness landscape recapitulate the outlines of the segmentation networks known to be operating during embryonic development.

Instead of asking if there *exists* a network of nodes such that they perform a given computation, and if it can be found by mutation and selection as in the examples above, one can ask how many network topologies perform a given computation. In other words, one is asking whether there is only one (fine tuned?) or many topologies or solutions to a given problem. The question of how many network topologies, proxies for different genotypes, produce the same dynamics, a proxy for phenotype, is a question of designability, a concept originally proposed to study the properties of amino-acid sequences comprising functional proteins, but applicable also to biological regulatory networks [106]. The authors examine three- and four-node binary networks with threshold updating rule and show that all networks with the shared phenotype have a common 'core' set of connections, but can differ in the variable part, similar to protein folding where the essential set of residues is necessary for the fold, with numerous variations in the nonessential part.

Future Directions

The study of biological networks is at an early stage, both on the theoretical as well as on the experimental side. Although high-throughput experiments are generating large data sets, these can suffer from serious biases, lack of temporal or spatial detail, and limited access to the component parts of the interacting system. On a theoretical front, general analytical insights that would link dynamics with network topology are few, although for specific systems with known topology computer simulation can be of great assistance. There can be confusion about which aspects of the dynamical model have biological significance and interpretation, and which aspects are just 'temporary variables' and the 'envelope' of the proverbial back-of-the-envelope calculations that cells use to perform their biological computations on; which parts of the trajectory are functionally constrained and which ones could fluctuate considerably with no ill-effects; how much noise is tolerable in the nodes of the network and what is its correlation structure; or how the unobserved, or 'hidden,' nodes (or their modification/activity states) influence the network dynamics.

Despite these caveats, cellular networks have some advantages over biological systems of comparable complexity, such as neural networks. Due to technological developments, we are considerably closer to the complete census of the interacting molecules in a cell than we are generally to the picture of connectivity of the neural tissue. Components of the regulatory networks are simpler than neurons, which are capable of a range of complicated behaviors on different timescales. Modules and pathways often comprise smaller number of interacting elements than in neural networks, making it possible to design small but interesting synthetic circuits. Last but not least, sequence and homology can provide strong insights or be powerful tools for network inference in their own right.

Those of us who come from the traditionally quantitative sciences, such as physics, were raised with experiments in which crucial elements are isolated and controlled. In biological systems, attempts at such isolation may break the regulatory mechanisms that are essential for normal operation of the system, leaving us with a system which is in fact more variable and less controlled than we would have if we faced the full complexity of the organism. It is only recently that we have seen the development of experimental techniques that allow fully quantitative, real time measurements of the molecular events inside individual cells, and the theoretical framework into which such measurements will be fit still is being constructed. The range of theoretical approaches being explored is diverse, and it behooves us to search for those approaches which have the chance to organize our understanding of many different systems rather than being satisfied with models of particular systems. Again, there is a balance between the search for generality and the need to connect with experiments on specific networks. We have tried to give some examples of all these developments, hopefully conveying the correct combination of enthusiasm and skepticism.

Acknowledgments

We thank our colleagues and collaborators who have helped us learn about these issues: MJ Berry, CG Callan, T Gregor, JB Kinney, P Mehta, SE Palmer, E Schneidman, JJ Hopfield, T Mora, S Setayeshgar, N Slonim, GJ Stephens, DW Tank, N Tishby, A Walczak, EF Wieschaus, CH Wiggins and NS Wingreen. Our work was supported in part by NIH grants P50 GM071508 and R01 GM077599, by NSF Grants IIS–0613435 and PHY–0650617, by the Swartz Foundation, and by the Burroughs Wellcome Fund.

Bibliography

1. Acar M, Becksei A, van Oudenaarden A (2005) Enhancement of cellular memory by reducing stochastic transitions. Nature 435:228–232
2. Ackers GK, Johnson AD, Shea MA (1982) Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci (USA) 79(4):1129–33
3. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–82
4. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci (USA) 100:8710–4
5. Alm E, Arkin AP (2003) Biological networks. Curr Opin Struct Biol 13(2):193–202

6. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397(6715):168–71
7. Arkin A, Ross J, McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. Genetics 149(4):1633–48
8. Arnosti DN, Kulkarni MM (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94(5):890–8
9. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N (2006) Noise in protein expression scales with natural protein abundance. Nat Genet 38(6):636–43
10. Barabási AL (2002) Linked: The New Science of Networks. Perseus Publishing, Cambridge
11. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–12
12. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2): 101–13
13. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387(6636):913–7
14. Baylor DA, Lamb TD, Yau KW (1979) Responses of retinal rods to single photons. J Physiol (Lond) 288:613–634
15. Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405(6786):590–3
16. Berg HC (1975) Chemotaxis in bacteria. Annu Rev Biophys Bioeng 4(00):119–36
17. Berg HC, Purcell EM (1977) Physics of chemoreception. Biophys J 20(2):193–219
18. Bergmann S, Sandler O, Sberro H, Shnider S, Schejter E, Shilo BZ, Barkai N (2007) Pre-steady-state decoding of the bicoid morphogen gradient. PLoS Biol 5(2):e46
19. Bethe HA (1935) Statistical theory of superlattices. Proc R Soc London Ser A 150:552–575
20. Bialek W (1987) Physical limits to sensation and perception. Annu Rev Biophys Biophys Chem 16:455–78
21. Bialek W (2001) Stability and noise in biochemical switches. Adv Neur Info Proc Syst 13:103
22. Bialek W, Setayeshgar S (2005) Physical limits to biochemical signaling. Proc Natl Acad Sci (USA) 102(29):10040–5
23. Bialek W, Setayeshgar S (2006) Cooperativity, sensitivity and noise in biochemical signaling. arXiv.org:q-bio.MN/0601001
24. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R (2005a) Transcriptional regulation by the numbers: applications. Curr Opin Genet Dev 15(2): 125–35
25. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R (2005b) Transcriptional regulation by the numbers: models. Curr Opin Genet Dev 15(2):116–24
26. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. Nature 422(6932):633–7
27. Block SM, Segall JE, Berg HC (1983) Adaptation kinetics in bacterial chemotaxis. J Bacteriol 154(1):312–23
28. Bray D (1995) Protein molecules as computational elements in living cells. Nature 376(6538):307–12
29. Bray D, Bourret RB, Simon MI (1993) Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis. Mol Biol Cell 4(5):469–82
30. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21(1 Suppl): 33–7
31. Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. Proc Natl Acad Sci (USA) 100(9): 5136–41
32. Chang L, Karin M (2001) Mammalian map kinase signalling cascades. Nature 410(6824):37–40
33. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol Biol Cell 11(1):369–91
34. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406(6792):188–92
35. Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. Nature 436(7050):588–92
36. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci (USA) 95(25):14863–8
37. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28(2):337–50
38. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403(6767):335–8
39. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297(5584):1183–6
40. Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA (1997) The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. Annu Rev Cell Dev Biol 13:457–512
41. Francois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. Proc Natl Acad Sci (USA) 101(2):580–5
42. Francois P, Hakim V, Siggia ED (2007) Deriving structure from evolution: metazoan segmentation. Mol Syst Bio 3: Article 154
43. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303(5659):799–805
44. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature 403(6767): 339–42
45. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–7
46. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425(6959):737–41
47. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361
48. Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55
49. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley J R L, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets

RA, McKenna MP, Chant J, Rothb erg JM (2003) A protein interaction map of *Drosophila melanogaster*. Science (5651): 1727–36

50. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. Cell 123(6): 1025–36
51. Goldman MS, Golowasch J, Marder E, Abbott LF (2001) Global structure robustness and modulation of neural models. J Neurosci 21:5229–5238
52. Gonze D, Halloy J, Goldbeter A (2002) Robustness of circadian rhythms with respect to molecular noise. Proc Natl Acad Sci (USA) 99(2):673–8
53. Goulian M (2004) Robust control in bacterial regulatory circuits. Curr Opin Microbiol 7(2):198–202
54. Gregor T, Tank DW, Wieschaus EF, Bialek W (2007a) Probing the limits to positional information. Cell 130(1):153–64
55. Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW (2007b) Stability and nuclear dynamics of the bicoid morphogen gradient. Cell 130(1):141–52
56. Guet CC, Elowitz MB, Hsing W, Leibler S (2002) Combinatorial synthesis of genetic networks. Science 296(5572):1466–70
57. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995): 88–93
58. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(6761 Suppl):C47–52
59. Hasty J, McMillen D, Isaacs F, Collins JJ (2001) Computational studies of gene regulatory networks: in numero molecular biology. Nat Rev Genet 2(4):268–79
60. Heinrich R, Schuster S (1996) The Regulation of Cellular Systems. Chapman and Hall, New York
61. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415(6868):180–3
62. Hofman J, Wiggins C (2007) A bayesian approach to network modularity. arXiv.org:07093512
63. Hooshangi S, Thiberge S, Weiss R (2005) Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. Proc Natl Acad Sci (USA) 102(10):3581–6
64. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci (USA) 79(8):2554–8
65. Huang KC, Meir Y, Wingreen NS (2003) Dynamic structures in *Escherichia coli:* spontaneous formation of mine rings and mind polar zones. Proc Natl Acad Sci (USA) 100(22):12724–8
66. Ibarra RU, Edwards JS, Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420(6912):186–9
67. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N (2002) Revealing modular organization in the yeast transcriptional network. Nat Genet 31(4):370–7
68. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci (USA) 98(8):4569–74
69. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–56
70. Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 7(5):535–45
71. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:62–79
72. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407(6804):651–4
73. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–2
74. Jordan JD, Landau EM, Iyengar R (2000) Signaling networks: the origins of cellular multitasking. Cell 103(2):193–200
75. Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116(2):247–57
76. van Kampen NG (2007) Stochastic Processes in Physics and Chemistry. Elsevier, Amsterdam
77. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at genomenet. Nucleic Acids Res 30(1):42–6
78. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S (2002) The ecocyc database. Nucleic Acids Res 30(1):56–8
79. Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci (USA) 102(39):13773–8
80. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22(3):437–67
81. Keller EF (2005) Revisiting "scale-free" networks. Bioessays 27(10):1060–8
82. Kirschner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci (USA) 95(15):8420–7
83. Kolch W (2000) Meaningful relationships: the regulation of the ras/raf/mek/erk pathway by protein interactions. Biochem J 351 Pt 2:289–305
84. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440(7084):637–43
85. Kuhlman T, Zhang Z, Saier J M H, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. Proc Natl Acad Sci (USA) 104(14):6043–8
86. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young

RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799–804

87. Leloup JC, Goldbeter A (2003) Toward a detailed computational model for the mammalian circadian clock. Proc Natl Acad Sci (USA) 100(12):7051–6
88. LeMasson G, Marder E, Abbott LF (1993) Activity-dependent regulation of conductances in model neurons. Science 259:1915–1917
89. Levine M, Davidson EH (2005) Gene regulatory networks for development. Proc Natl Acad Sci (USA) 102(14):4936–42
90. Lezon TR, Banavar JR, Cieplak M, Maritan A, Federoff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proc Natl Acad Sci (USA) 103:19033–19038
91. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci (USA) 101(14):4781–6
92. Libby E, Perkins TJ, Swain PS (2007) Noisy information processing through transcriptional regulation. Proc Natl Acad Sci (USA) 104(17):7151–6
93. Marder E, Bucher D (2006) Variability, compensation and homeostasis in neuron and network function. Nature Rev Neurosci 7:563–574
94. Markowetz F, Spang R (2007) Inferring cellular networks – a review. BMC Bioinformatics 8: 6-S5
95. Martin DE, Hall MN (2005) The expanding tor signaling network. Curr Opin Cell Biol 17(2):158–66
96. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296(5569):910–3
97. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci (USA) 94(3):814–9
98. McAdams HH, Arkin A (1999) It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet 15(2):65–9
99. McAdams HH, Shapiro L (1995) Circuit simulation of genetic networks. Science 269(5224):650–6
100. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887):399–403
101. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of evolved and designed networks. Science 303(5663):1538–42
102. Nakajima M, Imai K, Ito H, Nishiwaki T, Murayama Y, Iwasaki H, Oyama T, Kondo T (2005) Reconstitution of circadian oscillation of cyanobacterial kaic phophorylation in vitro. Science 308:414–415
103. Nasmyth K (1996) At the heart of the budding yeast cell cycle. Trends Genet 12(10):405–12
104. Newman M, Watts D, Barabási AL (2006) The Structure and Dynamics of Networks. Princeton University Press, Princeton
105. Nielsen UB, Cardone MH, Sinskey AJ, MacBeath G, Sorger PK (2003) Profiling receptor tyrosine kinase activation by using ab microarrays. Proc Natl Acad Sci (USA) 100(16):9330–5
106. Nochomovitz YD, Li H (2006) Highly designable phenotypes and mutational buffers emerge from a systematic mapping between network topology and dynamic output. Proc Natl Acad Sci (USA) 103(11):4180–5
107. Novak B, Tyson JJ (1997) Modeling the control of DNA replication in fission yeast. Proc Natl Acad Sci (USA) 94(17): 9147–52
108. Nurse P (2001) Cyclin dependent kinases and cell cycle control. Les Prix Nobel
109. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. Nat Genet 31(1):69–73
110. Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6(2):99–111
111. Paulsson J (2004) Summing up the noise in gene networks. Nature 427(6973):415–8
112. Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. Science 307(5717):1965–9
113. Perez OD, Nolan GP (2002) Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. Nat Biotechnol 20(2):155–62
114. Ptashne M (2001) Genes and Signals. CSHL Press, Cold Spring Harbor, USA
115. Ptashne M (2004) A Genetic Switch: Phage lambda Revisited. CSHL Press
116. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4(10):e309
117. Ramanathan S, Detwiler PB, Sengupta AM, Shraiman BI (2005) G-protein-coupled enzyme cascades have intrinsic properties that improve signal localization and fidelity. Biophys J 88(5):3063–71
118. Rao CV, Kirby JR, Arkin AP (2004) Design and diversity in bacterial chemotaxis: a comparative study in *Escherichia coli* and *Bacillus subtilis*. PLoS Biol 2(2):E49
119. Raser JM, O'Shea EK (2005) Noise in gene expression: origins, consequences, and control. Science 309(5743):2010–3
120. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–5
121. Rieke F, Baylor DA (1998) Single photon detection by rod cells of the retina. Rev Mod Phys 70:1027–1036
122. Roma DM, O'Flanagan R, Ruckenstein AE, Sengupta AM (2005) Optimal path to epigenetic swithcing. Phys Rev E 71: 011902
123. Rosenfeld N, Alon U (2003) Response delays and the structure of transcription networks. J Mol Biol 329(4):645–54
124. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. Science 307(5717):1962–5
125. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062): 1173–8
126. Rust MJ, Markson JS, Lane WS, Fisher DS, O'Shea EK (2007) Ordered phosphorylation giverns oscillation of a three-protein circadian clock. Science 318:809–812
127. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721):523–9

128. Sanchez L, Thieffry D (2001) A logical analysis of the *Drosophila* gap-gene system. J Theor Biol 211(2):115–41
129. Sasai M, Wolynes PG (2003) Stochastic gene expression as a many-body problem. Proc Natl Acad Sci (USA) 100(5): 2374–9
130. Schneidman E, Still S, Berry II MJ, Bialek W (2003) Network information and connected correlations. Phys Rev Lett 91(23):238701
131. Schneidman E, Berry II MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440(7087):1007–12
132. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U (2004) Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biol 2(9):E271
133. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34(2):166–76
134. Setty Y, Mayo AE, Surette MG, Alon U (2003) Detailed map of a cis-regulatory input function. Proc Natl Acad Sci (USA) 100(13):7702–7
135. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423 & 623–656
136. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31(1):64–8
137. Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, Litke AM, Chichilnisky EJ (2006) The structure of multi-neuron firing patterns in primate retina. J Neurosci 26(32):8254–66
138. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Liron Y, Rosenfeld N, Danon T, Perzov N, Alon U (2006) Variability and memory of protein levels in human cells. Nature 444(7119):643–6
139. Siggia ED (2005) Computational methods for transcriptional regulation. Curr Opin Genet Dev 15(2):214–21
140. Slonim N, Atwal GS, Tkačik G, Bialek W (2005) Information-based clustering. Proc Natl Acad Sci (USA) 102(51):18297–302
141. Slonim N, Elemento O, Tavazoie S (2006) Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. Mol Syst Biol 2 (2006) 0005
142. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci (USA) 100(21):12123–8
143. Stelling J, Gilles ED, Doyle 3rd FJ (2004) Robustness properties of circadian clock architectures. Proc Natl Acad Sci (USA) 101(36):13210–5
144. Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–76
145. Süel GM, Garcia-Ojalvo J, Liberman L, Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. Nature 440:545–550
146. Swain PS (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. J Mol Biol 344(4):965–76
147. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci (USA) 99(20):12795–800
148. Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci (USA) 102(20):7203–8
149. Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci (USA) 98(15):8614–9
150. Thomas R (1973) Boolean formalization of genetic control circuits. J Theor Biol 42(3):563–85
151. Tkačik G (2007) Information Flow in Biological Networks. Dissertation, Princeton University, Princeton
152. Tkačik G, Bialek W (2007) Diffusion, dimensionality and noise in transcriptional regulation. arXiv.org:07121852 [q-bio.MN]
153. Tkačik G, Schneidman E, Berry II MJ, Bialek W (2006) Ising models for networks of real neurons. arXiv.org: q-bio.NC/0611072
154. Tkačik G, Callan Jr CG, Bialek W (2008) Information capacity of genetic regulatory elements. Phys Rev E 78:011910. arXiv.org:0709.4209. [q-bioMN]
155. Tkačik G, Callan Jr CG, Bialek W (2008) Information flow and optimization in transcriptional regulation. Proc Natl Acad Sci 105(34):12265–12270. arXiv.org:0705.0313. [q-bio.MN]
156. Tkačik G, Gregor T, Bialek W (2008) The role of input noise in transcriptional regulation. PLoS One 3, e2774 arXiv.org:q-bioMN/0701002
157. Tomita J, Nakajima M, Kondo T, Iwasaki H (2005) No transcription-translation feedback in circadian rhythm of kaic phosphorylation. Science 307:251–254
158. Tucker CL, Gera JF, Uetz P (2001) Towards an understanding of complex protein networks. Trends Cell Biol 11(3):102–6
159. Tyson JJ, Chen K, Novak B (2001) Network dynamics and cell physiology. Nat Rev Mol Cell Biol 2(12):908–16
160. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15(2):221–31
161. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403(6770):623–7
162. de Visser JA, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, Gibson G, Hansen TF, Krakauer D, Lewontin RC, Ofria C, Rice SH, von Dassow G, Wagner A, Whitlock MC (2003) Perspective: Evolution and detection of genetic robustness. Evolution Int J Org Evolution 57(9):1959–72
163. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268(1478):1803–10
164. Walczak AM, Sasai M, Wolynes PG (2005) Self-consistent proteomic field theory of stochastic gene switches. Biophys J 88(2):828–50
165. Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. Annu Rev Cell Dev Biol 21: 319–46
166. Watson JD, Baker TA, Beli SP, Gann A, Levine M, Losick R (2003) Molecular Biology of the Gene: 5th edn. Benjamin Cummings, Menlo Park
167. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–2
168. Weinberger LS, Shenk T (2007) An hiv feedback resistor: autoregulatory circuit deactivator and noise buffer. PLoS Biol 5(1):e9
169. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-

protein interaction. Proc Natl Acad Sci (USA) 101(16):5934–5939
170. Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. Proc Nat Acad Sci (USA) 99(26):16587–91
171. Young MW, Kay SA (2001) Time zones: a comparative genetics of circadian clocks. Nat Rev Genet 2(9):702–15
172. Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, Alon U (2004) Just-in-time transcription program in metabolic pathways. Nat Genet 36(5):486–91
173. Ziv E, Koytcheff R, Middendorf M, Wiggins C (2005a) Systematic identification of statistically significant network measures. Phys Rev E 71:016110
174. Ziv E, Middendorf M, Wiggins C (2005b) Information theoretic approach to network modularity. Phys Rev E 71:046117
175. Ziv E, Nemenman I, Wiggins CH (2006) Optimal signal processing in small stochastic biochemical networks. arXiv.org:q-bio/0612041

Cellular Automata as Models of Parallel Computation

THOMAS WORSCH
Lehrstuhl Informatik für Ingenieure
und Naturwissenschaftler, Universität Karlsruhe,
Karlsruhe, Germany

Article Outline

Glossary

Cellular automaton The classical fine-grained parallel model introduced by John von Neumann.

Hyperbolic cellular automaton A cellular automaton resulting from a tessellation of the hyperbolic plane.

Parallel Turing machine A generalization of Turing's classical model where several control units work cooperatively on the same tape (or set of tapes).

Time complexity Number of steps needed for computing a result. Usually a function $t\colon \mathbb{N}_+ \to \mathbb{N}_+$, $t(n)$ being the maximum ("worst case") for any input of size n.

Space complexity Number of cells needed for computing a result. Usually a function $s\colon \mathbb{N}_+ \to \mathbb{N}_+$, $s(n)$ being the maximum for any input of size n.

State change complexity Number of proper state changes of cells during a computation. Usually a function $sc\colon \mathbb{N}_+ \to \mathbb{N}_+$, $sc(n)$ being the maximum for any input of size n.

Processor complexity Maximum number of control units of a parallel Turing machine which are simultaneously active during a computation. Usually a function $sc\colon \mathbb{N}_+ \to \mathbb{N}_+$, $sc(n)$ being the maximum for any input of size n.

$\mathbb{N}_+$ The set $\{1, 2, 3, \dots\}$ of positive natural numbers.

$\mathbb{Z}$ The set $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ of integers.

Q^G The set of all (total) functions from a set G to a set Q.

Definition of the Subject

This article will explore the properties of cellular automata (CA) as a parallel model.

The Main Theme

We will first look at the standard model of CA and compare it with Turing machines as the standard sequential model, mainly from a computational complexity point of view. From there we will proceed in two directions: by removing computational power and by adding computational power in different ways in order to gain insight into the importance of some ingredients of the definition of CA.

What Is Left Out

There are topics which we will not cover although they would have fit under the title.

One such topic is *parallel algorithms* for CA. There are algorithmic problems which make sense only for parallel models. Probably the most famous for CA is the so-called *Firing Squad Synchronization Problem*. This is the topic of Umeo's article (▶ Firing Squad Synchronization Problem in Cellular Automata), which can also be found in this encyclopedia.

Another such topic in this area is the Leader election problem. For CA it has received increased attention in recent years. See the paper by Stratmann and Worsch [29] and the references therein for more details.

And we do want to mention the most exciting (in our opinion) CA algorithm: Tougne has designed a CA which, starting from a single point, after t steps has generated the discretized circle of radius t, for all t; see [5] for this gem.

There are also models which generalize standard CA by making the cells more powerful. Kutrib has introduced *push-down cellular automata* [14]. As the name indicates, in this model each cell does not have a finite memory but can make use of a potentially unbounded stack of symbols.

The area of nondeterministic CA is also not covered here. For results concerning formal language recognition with these devices refer to ▶ Cellular Automata and Language Theory.

All these topics are, unfortunately, beyond the scope of this article.

Structure of the Paper

The core of this article consists of four sections:

Introduction: The main point is the standard definition of Euclidean deterministic synchronous cellular automata. Furthermore, some general aspects of parallel models and typical questions and problems are discussed.

Time and space complexity: After defining the standard computational complexity measures, we compare CA with different resource bounds. The comparison of CA with the Turing Machine (TM) gives basic insights into their computational power.

Measuring and controlling activities: There are two approaches to measure the "amount of parallelism" in CA. One is an additional complexity measured directly for CA, the other via the definition of so-called parallel Turing machines. Both are discussed.

Communication: Here we have a look at "variants" of CA with communication structures other than the one-dimensional line. We sketch the proofs that some of these variants are in the second machine class.

Introduction

In this section we will first formalize the classical model of cellular automata, basically introduced by von Neumann [21]. Afterwards we will recap some general facts about parallel models.

Definition of Cellular Automata

There are several equivalent formalizations of CA and of course one chooses the one most appropriate for the topics to be investigated. Our point of view will be that each CA consists of a regular arrangement of basic processing elements working in parallel while exchanging data.

Below, for each of the words *regular*, *basic*, *processing*, *parallel* and *exchanging*, we first give the standard definition for clarification. Then we briefly point out possible alternatives which will be discussed in more detail in later sections.

Underlying Grid A Cellular Automaton (CA) consists of a set G of cells, where each cell has at least one *neighbor* with which it can exchange data. Informally speaking one usually assumes a "regular" arrangement of cells and, in particular, identically shaped neighborhoods.

For a *d-dimensional CA*, $d \in \mathbb{N}_+$, one can think of $G = \mathbb{Z}^d$. Neighbors are specified by a finite set N of coordinate differences called the *neighborhood*. The cell $i \in G$ has as its neighbors the cells $i + n$ for all $n \in N$. Usually one assumes that $0 \in N$. (Here we write 0 for a vector of d zeros.)

As long as one is not specifically interested in the precise role N is playing, one may assume some standard neighborhood: The *von Neumann neighborhood* of radius r is $N^{(r)} = \{(k_1, \ldots, k_d) \mid \sum_j |k_j| \leq r\}$ and the *Moore neighborhood* of radius r is $M^{(r)} = \{(k_1, \ldots, k_d) \mid \max_j |k_j| \leq r\}$.

The choices of G and N determine the structure of what in a real parallel computer would be called the "communication network". We will usually consider the case $G = \mathbb{Z}^d$ and assume that the neighborhood is $N = N^{(1)}$.

Discussion The structure of connections between cells is sometimes defined using the concept of *Cayley graphs*. Refer to the article by Ceccherini–Silberstein (▶ Cellular Automata and Groups), also in this encyclopedia, for details.

Another approach is via regular tessellations. For example the 2-dimensional Euclidean space can be tiled with copies of a square. These can be considered as cells, and cells sharing an edge are neighbors. Similarly one can tile, e. g., the *hyperbolic plane* with copies of a regular k-gon. This will be considered to some extent in Sect. "Communication in CA". A more thorough exposition can be found in the article by Margenstern (▶ Cellular Automata in Hyperbolic Spaces) also in this encyclopedia.

CA resulting, for example, from tessellations of the 2-dimensional Euclidean plane with triangles or hexagons are considered in the article by Bays (▶ Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations) also in this encyclopedia.

Global and Local Configurations The basic processing capabilities of each cell are those of a finite automaton. The set of possible states of each cell, denoted by Q, is finite. As inputs to be processed each cell gets the states of all the cells in its neighborhood.

We will write Q^G for the set of all functions from G to Q. Thus each $c \in Q^G$ describes a possible global state of the whole CA. We will call these c *(global) configurations*.

On the other hand, functions $\ell\colon N \to Q$ are called *local configurations*. We say that in a configuration c cell i *observes* the local configuration $c_{i+N}\colon N \to Q$ where $c_{i+N}(n) = c(i+n)$. A cell gets its currently observed local configuration as input. It remains to be defined how they are processed.

Dynamics The dynamics of a CA are defined by specifying the local dynamics of a single cell and how cells "operate in parallel" (if at all). In the first sections we will consider the classical case:

- A *local transition function* is a function $f\colon Q^N \to Q$ prescribing for each local configuration $\ell \in Q^N$ the next state $f(\ell)$ of a cell which currently observes ℓ in its neighborhood. In particular, this means that we are considering *deterministic* behavior of cells.
- Furthermore, we will first concentrate on CA where all cells are working *synchronously*: The possible transitions from one configuration to the next one in one global step of the CA can be described by a function $F\colon Q^G \to Q^G$ requiring that *all* cells make one state transition: $\forall i \in G\colon F(c)(i) = f(c_{i+N})$.

For alternative definitions of the dynamic behavior of CA see Sect. "Measuring and Controlling the Activities".

Discussion Basically, the above definition of CA is the standard one going back to von Neumann [21]; he used $G = \mathbb{Z}^2$ and $N = \{(-1,0),(1,0),(0,-1),(0,1)\}$ for his construction. But for all of the aspects just defined there are other possibilities, some of which will be discussed in later sections.

Finite Computations on CA

In this article we are interested in using CA as devices for computing, given some finite input, in a finite number of steps a finite output.

Inputs As the prototypical examples of problems to be solved by CA and other models we will consider the recognition of formal languages. This has the advantages that the inputs have a simple structure and, more importantly, the output is only one bit (accept or reject) and can be formalized easily.

A detailed discussion of CA as formal language recognizers can be found in the article by Kutrib (▶ Cellular Automata and Language Theory).

The *input alphabet* will be denoted by A. We assume that $A \subset Q$. In addition there has to be a special state $q \in Q$ which is called a *quiescent state* because it has the property that for the quiescent local configuration $\ell_q\colon N \to Q\colon n \mapsto q$ the local transition function must specify $f(\ell_q) = q$.

In the literature two input modes are usually considered.

- Parallel input mode: For an input $w = x_1 \cdots x_n \in A^n$ the initial configuration c_w is defined as

$$c_w(i) = \begin{cases} x_j & \text{iff } i = (j, 0, \ldots, 0) \\ q & \text{otherwise .} \end{cases}$$

- Sequential input mode: In this case, all cells are in state q in the initial configuration. But cell $(0, \ldots, 0)$ is a designated *input cell* and acts differently from the others. It works according to a function $g\colon Q^N \times (A {\sim} \cup \{q\})$. During the first n steps the input cell gets input symbol x_j in step j; after the last input symbol it always gets q. CA using this type of input are often called *iterative arrays (IA)*.

Unless otherwise noted we will always assume parallel input mode.

Conceptually the difference between the two input modes has the same consequences as for TM. If input is provided sequentially it is meaningful to have a look at computations which "use" less than n cells (see the definition of space complexity later on).

Technically, some results occur only for the parallel input mode but not for the sequential one, or vice versa. This is the case, for example, when one looks at devices with small time bounds like n or $n + \sqrt{n}$ steps. But as soon as one considers more generally $\Theta(n)$ or more steps and a space complexity of at least $\Theta(n)$, both CA and IA can simulate each other in linear time:

- An IA can first read the complete input word, storing it in successive cells, and then start simulating a CA, and
- A CA can shift the whole word to the cell holding the first input symbol and have it act as the designated input cell of an IA.

Outputs Concerning output, one usually defines that a CA has finished its work whenever it has reached a *stable configuration* c, i. e., $F(c) = c$. In such a case we will also say that the CA *halts* (although formally one can continue to apply F). An input word $w \in A^+$ is *accepted*, iff the cell $(1, 0, \ldots, 0)$ which got the first input symbol, is in an *accepting* state from a designated finite subset $F_+ \subset Q$ of states. We write $L(C)$ for the set of all words $w \in A^+$ which are accepted by a CA C.

For the sake of simplicity we will assume that all deterministic machines under consideration halt for all inputs. E. g., the CA as defined above always reaches a stable configuration.

The sequence of all configurations from the initial one for some input w to the stable one is called the *computation* for input w.

Discussion For 1-dimensional CA the definition of parallel input is the obvious one. For higher-dimensional CA, say $G = \mathbb{Z}^2$, one could also think of more compact forms of input for one-dimensional words, e. g., inscribing the input symbols row by row into a square with side length $\lceil\sqrt{n}\rceil$. But this requires extra care. Depending on the formal language to be accepted, the special way in which the symbols are input might provide additional information which is useful for the language recognition task at hand.

Since a CA performs work on an infinite number of bits in each step, it would also be possible to consider inputs and outputs of infinite length, e. g., as representations of all real numbers in the interval $[0, \dots, 1]$. There is much less literature about this aspect; see for example chapter 11 of [8].

It is not too surprising that this area is also related to the view of CA as dynamical systems (instead of computers); see the contributions by Formenti (▶ Chaotic Behavior of Cellular Automata) and Kůrka (▶ Topological Dynamics of Cellular Automata).

Example: Recognition of Palindromes As an example that will also be useful later, consider the formal language L_{pal} of palindromes of odd length:

$$L_{\text{pal}} = \{vxv^R \mid v \in A^* \wedge x \in A\} \, .$$

(Here v^R is the mirror image of v.) For example, if A contains all Latin letters `saippuakauppias` belongs to L_{pal} (the Finnish word for a soap dealer).

It is known that each $\mathbb{T}$ − TM with only one tape and only one head on it (see Subsect. "Turing Machines" for a quick introduction) needs time $\Omega(n^2)$ for the recognition of at least some inputs of length n belonging to L_{pal} [10].

We will sketch a CA recognizing L_{pal} in time $\Theta(n)$. As the set of states for a single cell we use $Q = A \cup \{\sqcup\} \cup Q_l \times Q_r \times Q_v \times Q_{lr}$, basically subdividing each cell in 4 "registers", each containing a "substate". The substates from $Q_l = \{\texttt{<a}, \texttt{<b}, \sqcup\}$ and $Q_r = \{\texttt{a>}, \texttt{b>}, \sqcup\}$ are used to shift input symbols to the left and to the right respectively. In the third register a substate from $Q_v = \{+, -\}$ indicates the results of comparisons. In the fourth register substates from $Q_{lr} = \{\texttt{>}, \texttt{<}, \texttt{<+>}, \texttt{<->}, \sqcup\}$ are used to realize "signals" > and < which identify the middle cell and distribute the relevant overall comparison result to all cells.

As accepting states one chooses those, whose last component is <+>: $F_+ = Q_l \times Q_r \times Q_v \times \{\texttt{<+>}\}$.

There is a total of $3 + 3 \cdot 3 \cdot 2 \cdot 5 = 93$ states and for a complete definition of the local transition function one would have to specify $f(x, y, z)$ for $93^3 = 804357$ triples of states. We will not do that, but we will sketch some important parts. In the first step the registers are initialized. For all $x, y, z \in A$:

$\ell(-1)$	$\ell(0)$	$\ell(1)$	$f(\ell)$
$\sqcup$	y	z	$(\texttt{<}y, y\texttt{>}, +, \texttt{>})$
x	y	z	$(\texttt{<}y, y\texttt{>}, +, \sqcup)$
x	y	$\sqcup$	$(\texttt{<}y, y\texttt{>}, +, \texttt{<})$
$\sqcup$	y	$\sqcup$	$(\texttt{<}y, y\texttt{>}, +, \texttt{<+>})$

In all later steps, if

$$\ell(-1) = \boxed{\begin{array}{r} < x_l \\ x_r > \\ v_l \\ d_l \end{array}} \qquad \ell(0) = \boxed{\begin{array}{r} < y_l \\ y_r > \\ v_m \\ d_m \end{array}} \qquad \ell(1) = \boxed{\begin{array}{r} < z_l \\ z_r > \\ v_r \\ d_r \end{array}}$$

then

$$f(\ell) = \boxed{\begin{array}{r} < z_l \\ x_r > \\ v'_m \\ d'_m \end{array}}$$

Here, of course, the new value of the third register is computed as

$$v'_m = \begin{cases} + & \text{if } v_m = + \wedge z_l = x_r \\ - & \text{otherwise} \, . \end{cases}$$

We do not describe the computation of d'_m in detail. Figure 1 shows the computation for the input `babbbab`, which is a palindrome. Horizontal double lines separate configurations at subsequent time steps. Registers in state $\sqcup$ are simply left empty.

As can be seen there is a triangle in the space time diagram consisting of the n input cells at time $t = 0$ and shrinking at both ends by one cell in each subsequent step where "a lot of activity" can happen due to the shifting of the input symbols in both directions.

Clearly a two-head TM can also recognize L_{pal} in linear time, by first moving one head to the last symbol and then synchronously shifting both heads towards each other comparing the symbols read.

Informally speaking in this case the ability of multihead TM to transport a small amount of information over

...		b	a	b	b	b	a	b		...
...		<b	<a	<b	<b	<b	<a	<b		...
...		b>	a>	b>	b>	b>	a>	b>		...
...		+	+	+	+	+	+	+		...
...		>						<		...
...		<a	<b	<b	<b	<a	<b			...
...			b>	a>	b>	b>	b>	a>		...
...		-	+	-	+	-	+	-		...
...			>				<			...
...		<b	<b	<b	<a	<b				...
...				b>	a>	b>	b>	b>		...
...		-	-	-	+	-	-	-		...
...				>		<				...
...		<b	<b	<a	<b					...
...					b>	a>	b>	b>		...
...		-	-	-	+	-	-	-		...
...					<+>					...
...		<b	<a	<b						...
...						b>	a>	b>		...
...		-	-	-	+	-	-	-		...
...				<+>	<+>	<+>				...
...		<a	<b							...
...							b>	a>		...
...		-	-	-	-	-	-	-		...
...			<+>	<+>	<+>	<+>	<+>			...
...		<b								...
...								b>		...
...		-	-	-	-	-	-	-		...
...		<+>	<+>	<+>	<+>	<+>	<+>	<+>		...
...										...
...										...
...		-	-	-	-	-	-	-		...
...		<+>	<+>	<+>	<+>	<+>	<+>	<+>		...

Cellular Automata as Models of Parallel Computation, Figure 1
Recognition of a palindrome; the last configuration is stable and the cell which initially stored the first input symbol is in an accepting state

a long distance in one step can be "compensated" by CA by shifting a large amount of information over a short distance. We will see in Theorem 3 that this observation can be generalized.

Complexity Measures: Time and Space For one-dimensional CA it is straightforward to define their time and space complexity. We will consider only worst-case complexity. Remember that we assume that all CA halt for all inputs (reach a stable configuration).

For $w \in A^+$ let $\text{time}'(w)$ denote the smallest number τ of steps such that the CA reaches a stable configuration after τ steps when started from the initial configuration for input w. Then

$$\text{time}\colon \mathbb{N}_+ \to \mathbb{N}_+ : n \mapsto \max\{\text{time}'(w) \mid w \in A^n\}$$

is called the *time complexity* of the CA.

Similarly, let $\text{space}'(w)$ denote the total number of cells which are not quiescent in at least one configuration occurring during the computation for input w. Then

$$\text{space}\colon \mathbb{N}_+ \to \mathbb{N}_+ : n \mapsto \max\{\text{space}'(w) \mid w \in A^n\}$$

is called the *space complexity* of the CA. If we want to mention a specific CA C, we indicate it as an index, e. g., time_C.

If s and t are functions $\mathbb{N}_+ \to \mathbb{N}_+$, we write $\text{CA} - \text{Spc}(s) - \text{Time}(t)$ for the set of formal languages which can be accepted by some CA C with $\text{space}_C \leq s$ and $\text{time}_C \leq t$, and analogously $\text{CA} - \text{Spc}(s)$ and $\text{CA} - \text{Time}(t)$ if only one complexity measure is bounded. Thus we only look at upper bounds. For a whole set of functions $\mathcal{T}$, we will use the abbreviation

$$\text{CA} - \text{Time}(\mathcal{T}) = \bigcup_{t \in \mathcal{T}} \text{CA} - \text{Time}(f)\,.$$

Typical examples will be $\mathcal{T} = \text{O}(n)$ or $\mathcal{T} = \text{Pol}(n)$, where in general $\text{Pol}(f) = \bigcup_{k \in \mathbb{N}_+} \text{O}(f^k)$.

Resource bounded complexity classes for other computational models will be noted similarly. If we want to make the dimension of the CA explicit, we write $\mathbb{Z}^d - \text{CA} - \ldots$; if the prefix $\mathbb{Z}^d$ is missing, $d = 1$ is to be assumed.

Throughout this article n will always denote the length of input words. Thus a time complexity of $\text{Pol}(n)$ simply means polynomial time, and similarly for space, so that $\text{TM} - \text{Time}(\text{Pol}(n)) = \mathbf{P}$ and $\text{TM} - \text{Spc}(\text{Pol}(n)) = \mathbf{PSPACE}$.

Discussion For higher-dimensional CA the definition of space complexity requires more consideration. One possibility is to count the number of cells used during the computation. A different, but sometimes more convenient approach is to count the number of cells in the smallest hyper-rectangle comprising all used cells.

Turing Machines

For reference, and because we will consider a parallel variant, we set forth some definitions of Turing machines.

In general we allow Turing machines with k work tapes and h heads on each of them. Each square carries a symbol from the *tape alphabet* B which includes the blank symbol $\square$. The control unit (CU) is a finite automaton with *set of states* S. The possible actions of a deterministic TM are described by a function $f\colon S \times B^{kh} \to S \times B^{kh} \times D^{kh}$, where $D = \{-1, 0, +1\}$ is used for indicating the direction of movement of a head.

If the machine reaches a situation in which $f(s, b_1, \ldots, b_{kh}) = (s, b_1, \ldots, b_{kh}, 0, \ldots, 0)$ for the current state s and the currently scanned symbols $b_1, \ldots, b_{kh}$, we say that it halts.

Initially a word w of length n over the input alphabet $A \subset B$ is written on the first tape on squares $1, \ldots, n$, all other tape squares are empty, i. e., carry the $\square$. An input is *accepted* if the CU halts in an accepting state from a designated subset $F_+ \subset S$. $L(T)$ denotes the formal language of all words accepted by a TM T.

We write $k\mathbb{T}h - \mathrm{TM} - \mathrm{SPC}(s) - \mathrm{TIME}(t)$ for the class of all formal languages which can be recognized by TM T with k work tapes and h heads on each of them which have a space complexity $\mathrm{space}_T \leq s$ and $\mathrm{time}_T \leq t$. If k and/or h is missing, 1 is assumed instead. If the whole prefix $k\mathbb{T}h$ is missing, $\mathbb{T}$ is assumed. If arbitrary k and h are allowed, we write $*\mathbb{T}*$.

Sequential Versus Parallel Models

Today quite a number of different computational models are known which intuitively look as if they are parallel. Several years ago van Emde Boas [6] observed that many of these models have one property in common. The problems that can be solved in polynomial *time* on such a model $\mathcal{P}$, coincide with the problems that can be solved in polynomial *space* on Turing machines:

$$\mathcal{P} - \mathrm{TIME}(\mathrm{Pol}(n)) = \mathrm{TM} - \mathrm{SPC}(\mathrm{Pol}(n)) = \mathbf{PSPACE} .$$

Here we have chosen $\mathcal{P}$ as an abbreviation for "parallel" model. Models $\mathcal{P}$ satisfying this equality are by definition the members of the so-called *second machine class*.

On the other hand, the *first machine class* is formed by all models S satisfying the relation

$$S - \mathrm{SPC}(s) - \mathrm{TIME}(t) = \mathrm{TM} - \mathrm{SPC}(\Theta(s)) - \mathrm{TIME}(\mathrm{Pol}(t))$$

at least for some reasonable functions s and t. We deliberately avoid making this more precise. In general there is consensus on which models are in the these machine classes.

We do want to point out that the naming of the two machine classes does *not* mean that they are different or even disjoint. This is not known. For example, if $\mathbf{P} = \mathbf{PSPACE}$, it might be that the classes coincide. Furthermore there are models, e. g., Savitch's NLPRAM [25], which might be in neither machine class.

Another observation is that in order to possibly classify a machine model it obviously has to have something like "time complexity" and/or "space complexity". This may sound trivial, but we will see in Subsect. "Parallel Turing Machines" that, for example, for so-called parallel Turing machines with several work tapes it is in fact not.

Time and Space Complexity

Comparison of Resource Bounded One-Dimensional CA

It is clear that time and space complexity for CA are Blum measures [2] and hence infinite hierarchies of complexity classes exist. It follows from the more general Theorem 9 for parallel Turing machines that the following holds:

Theorem 1 *Let s and t be two functions such that s is fully CA space constructable in time t and t is CA computable in space s and time t. Then:*

$$\bigcup_{\gamma \notin \mathrm{O}(1)} \mathrm{CA} - \mathrm{SPC}(\Theta(s/\gamma)) - \mathrm{TIME}(\Theta(t/\gamma)) \subsetneqq \mathrm{CA} - \mathrm{SPC}(\mathrm{O}(s)) - \mathrm{TIME}(\mathrm{O}(t))$$

$$\mathrm{CA} - \mathrm{SPC}(o(s)) - \mathrm{TIME}(o(t)) \subsetneqq \mathrm{CA} - \mathrm{SPC}(\mathrm{O}(s)) - \mathrm{TIME}(\mathrm{O}(t))$$

$$\mathrm{CA} - \mathrm{TIME}(o(t)) \subsetneqq \mathrm{CA} - \mathrm{TIME}(\mathrm{O}(t)) .$$

The second and third inclusion are simple corollaries of the first one. We do not go into the details of the definition of CA constructibility, but note that for hierarchy results for TM one sometimes needs analogous additional conditions. For details, interested readers are referred to [3,12,17].

We not that for CA the situation is better than for deterministic TM: there one needs $f(n)\log f(n) \in o(g(n))$ in order to prove $\mathrm{TM} - \mathrm{TIME}(f) \subsetneqq \mathrm{TM} - \mathrm{TIME}(g)$.

Open Problem 2 For the proper inclusions in Theorem 1 the construction used in [34] really needs to increase the space used in order to get the time hierarchy. It is an open problem whether there also exists a time hierarchy if the space complexity is *fixed*, e. g., as $s(n) = n$. It is even an open problem to prove or disprove that the inclusion

$$\mathrm{CA} - \mathrm{SPC}(n) - \mathrm{TIME}(n) \subseteq \mathrm{CA} - \mathrm{SPC}(n) - \mathrm{TIME}(2^{\mathrm{O}(n)}) .$$

is proper or not. We will come back to this topic in Subsect. "Parallel Turing Machines" on parallel Turing machines.

Comparison with Turing Machines

It is well-known that a TM with one tape and one head on that tape can be simulated by a one-dimensional CA. See for example the paper by Smith [27]. But even multi-tape TM can be simulated by a one-dimensional CA without any significant loss of time.

Theorem 3 *For all space bounds $s(n) \geq n$ and all time bounds $t(n) \geq n$ the following holds for one-dimensional CA and TM with an arbitrary number of heads on its tapes:*

$$* \mathbb{T} * - \mathrm{TM} - \mathrm{SPC}(s) - \mathrm{TIME}(t) \subseteq \mathrm{CA} - \mathrm{SPC}(s) - \mathrm{TIME}(\mathrm{O}(t)) .$$

Sketch of the Simulation We first describe a simulation for $1\mathbb{T}1 - \mathrm{TM}$.

In this case the actions of the TM are of the form $s, b \to s', b', d$ where $s, s' \in S$ are old and new state, $b, b' \in B$ old and new tape symbol and $d \in \{-1, 0, +1\}$ the direction of *head* movement.

The simulating CA uses three substates in each cell, one for a TM state, one for a tape symbol, and an additional one for shifting tape symbols: $Q = Q_S \times Q_T \times Q_M$. We use $Q_S = S \cup \{\text{␣}\}$ and a substate of ␣ means, that the cell does not store a state. Similarly $Q_T = B \cup \{ < \circ, \circ > \}$ and a substate of $< \circ$ or $\circ >$ means that there is no symbol stored but a "hole" to be filled with an adjacent symbol. Substates from $Q_S = B \times \{ <, > \}$, like $< b$ and $b >$, are used for shifting symbols from one cell to the adjacent one to the left or right.

Instead of moving the state one cell to the right or left whenever the TM moves its head, the tape contents as stored in the CA are shifted in the opposite direction. Assume for example that the TM performs the following actions:

- s0,d $\to$ s1, d$'$, +1
- s1,e $\to$ s2, e$'$, −1
- s2,d$'$ $\to$ s3, d$''$, −1
- s3,c $\to$ s4, c$'$, +1

Figure 2 shows how shifting the tape in direction d can be achieved by sending the current symbol in that direction and sending a "hole" $\circ$ in the opposite direction $-d$. It should be clear that the required state changes of each cell depend only on information available *in its neighborhood*.

A consequence of this approach to incrementally shift the tape contents is that it takes an arbitrary large number of steps until all symbols have been shifted. On the other hand, after only *two* steps the cell simulating the TM control unit has information about the next symbol visited and can simulate the next TM step and initialize the next tape shift.

Clearly the same approach can be used if one wants to simulate a TM with several tapes, each having one head. For each additional tape the CA would use two additional registers analogously to the middle and bottom row used in Fig. 2 for one tape. Stoß [28] has proved that $k\mathbb{T}\,h - \text{TM}$ (h heads on each tape) can be simulated by $(kh)\mathbb{T} - \text{TM}$ (only one head on each tape) in linear time. Hence there is nothing left to prove.

Discussion As one can see in Fig. 2, in every second step one signal is sent to the left and one to the right. Thus, if the TM moves its head a lot and if the tape segment which has to be shifted is already long, many signals are traveling simultaneously.

… … …		a	b	c	s0 d	e	f	g		… … …
… … …		a	b	c	s1 o> <d’	e	f	g		… … …
… … …		a	b	d’ <c	s1 e	o>	f	g		… … …
… … …		a	c <c	d’	s2 <o e’>	f	o>	g		… … …
… … …		b <a	c	<o	s2 d’	e’ f>	g	o>		… … …
… … …	a	b	<o	c	s3 <o d">	e’	f g>			… … …
… … …	a	<o	b	<o	s3 c	d" e’>	f	g		… … …
… … …	<o	a	<o	b	s4 o> <c’	d"	e’ f>	g		… … …

Cellular Automata as Models of Parallel Computation, Figure 2
Shifting tape contents step by step

In other words, the CA transports "a large amount of information over a short distance in one step". Theorem 3 says that this ability is at least as powerful as the ability of multi-head TM to transport "a small amount of information over a long distance in one step".

Open Problem 4 The question remains whether some kind of converse also holds and in Theorem 3 an $=$ sign would be correct instead of the $\subseteq$, or whether CA are more powerful, i. e., a $\subsetneq$ sign would be correct. This is not known.

The best simulation of CA by TM that is known is the obvious one: states of neighboring cells are stored on adjacent tape squares. For the simulation of one CA step the TM basically makes one sweep across the complete tape segment containing the states of all non-quiescent cells updating them one after the other. As a consequence one gets

Theorem 5 *For all space bounds $s(n) \geq n$ and all time bounds $t(n) \geq n$ holds:*

$$\begin{aligned} \text{CA} - \textsc{Spc}(s) - \textsc{Time}(t) &\subseteq \text{TM} - \textsc{Spc}(s) - \textsc{Time}(\mathrm{O}(s \cdot t)) \\ &\subseteq \text{TM} - \textsc{Spc}(s) - \textsc{Time}\left(\mathrm{O}\left(t^2\right)\right) . \end{aligned}$$

The construction proving the first inclusion needs only a one-head TM, and no possibility is known to take advantage of more heads. The second inclusion follows from the observation that in order to use an initially blank tape square, a TM must move one of its heads there, which requires time. Thus $s \in \mathrm{O}(t)$.

Taking Theorems 3 and 5 together, one immediately gets

Corollary 6 *Cellular automata are in the first machine class.*

And it is not known that CA are in the second machine class. In this regard they are "more like" sequential models. The reason for this is the fact, the number of active processing units only grows polynomially with the number of steps in a computation. In Sect. "Communication in CA" variations of the standard CA model will be considered, where this is different.

Measuring and Controlling the Activities

Parallel Turing Machines

One possible way to make a parallel model from Turing machines is to allow several control units (CU), but with all of them working on the same tape (or tapes). This model can be traced back at least to a paper by Hemmerling [9] who called it *Systems of Turing automata*. A few years later Wiedermann [32] coined the term *Parallel Turing machine* (PTM).

We consider only the case where there is only one tape and each of the control units has only one head on that tape. As for sequential TM, we usually drop the prefix $1\mathbb{T}1$ for PTM, too. Readers interested in the case of PTM with multi-head CUs are referred to [33].

PTM with One-Head Control Units The specification of a PTM includes a tape alphabet B with a blank Cybil $\square$ and a set S of possible states for each CU. A PTM starts with one CU on the first input symbol as a sequential $1\mathbb{T}1-\text{TM}$. During the computation the number of control units may increase and decrease, but all CUs always work cooperatively on one common tape.

The idea is to have the CUs act independently unless they are "close" to each other, retaining the idea of only local interactions, as in CA.

A *configuration* of a PTM is a pair $c = (p, b)$. The mapping $b\colon \mathbb{Z} \to B$ describes the contents of the tape. Let 2^S denote the power set of S. The mapping $p\colon \mathbb{Z} \to 2^S$ describes for each tape square i the set of states of the finite automata currently visiting it. In particular, this formalization means that it is not possible to distinguish two automata on the same square and in the same state: the idea is that because of that they will always behave identically and hence need not be distinguished.

The mode of operation of a PTM is determined by the *transition function* $f\colon 2^Q \times B \to 2^{Q\times D} \times B$ where D is the set $\{-1, 0, 1\}$ of possible movements of a control unit. In order to compute the successor configuration $c' = (p', b')$ of a configuration $c = (p, b)$, f is simultaneously computed for all tape positions $i \in \mathbb{Z}$. The arguments used are the set of states of the finite automata currently visiting square i and its tape symbol. Let $(M'_i, b'_i) = f(p(i), b(i))$. Then the new symbol on square i in configuration c' is $b'(i) = b'_i$. The set of finite automata on square i is replaced by a new set of finite automata (defined by $M'_i \subseteq Q \times D$) each of which changes the tape square according to the indicated direction of movement. Therefore $p'(i) = \{q \mid (q, 1) \in M'_{i-1} \vee (q, 0) \in M'_i \vee (q, -1) \in M'_{i+1}\}$. Thus f induces a global transition function F mapping global configurations to global configurations.

In order to make the model useful (and to come up to some intuitive expectations) it is required, that CUs cannot arise "out of nothing" and that the symbol on a tape square can change only if it is visited by at least one CU. In other words we require that $\forall\, b \in B\colon\ f(\emptyset, b) = (\emptyset, b)$.

Observe that the number of finite automata on the tape may change during a computation. Automata may vanish, for example if $f(\{s\}, b) = (\emptyset, b)$ and new automata may be generated, for example if $f(\{s\}, b) = (\{(q, 1), (q', 0)\}, b)$.

For the recognition of formal languages we define the *initial configuration* c_w for an input word $w \in A^+$ as the one in which w is written on the otherwise blank tape on squares $1, 2, \dots, |w|$, and in which there exists exactly one finite automaton in an initial state q_0 on square 1.

A configuration (p, b) of aPTM is called *accepting* iff it is stable (i. e. $F((p, b)) = (p, b)$) and $p(1) \subseteq F_+$. The *language $L(P)$ recognized by a* PTM P is the set of input words, for which it reaches an accepting configuration.

Complexity Measures for PTM Time complexity of a PTM can be defined in the obvious way.

For space complexity, one counts the total number of tape squares which are used in at least one configuration. Here we call a tape square i *unused* in a configuration $c = (p, b)$ if $p(i) = \emptyset$ and $b(i) = \square$; otherwise it is used.

What makes PTM interesting is the definition of its *processor complexity*. Let $\text{proc}'(w)$ denote the maximum number of CU which exist simultaneously in a configuration occurring during the computation for input w and define $\text{proc}\colon \mathbb{N}_+ \to \mathbb{N}_+\colon n \mapsto \max\{\text{proc}'(w) \mid w \in A^n\}$. For complexity classes we use the notation $\text{PTM}-\textsc{Spc}(s)-\textsc{Time}(t)-\textsc{Proc}(p)$ etc.

The processor complexity is one way to measure (an upper bound on) "how many activities" happened simultaneously. It should be clear that at the lower end one has the case of constant $\text{proc}(n) = 1$, which means that the PTM is in fact (equivalent to) a sequential TM. The other extreme is to have CUs "everywhere". In that case

$\text{proc}(n) \in \Theta(\text{space}(n))$, and one basically has a CA. In other words, processor complexity measures the mount of parallelism of a PTM.

Theorem 7 *For all space bounds s and time bounds t:*

$$\begin{aligned}\text{PTM} - \text{Spc}(s) - \text{Time}(t) - \text{Proc}(1) \\ &= \text{TM} - \text{Spc}(s) - \text{Time}(t) \\ \text{PTM} - \text{Spc}(s) - \text{Time}(t) - \text{Proc}(s) \\ &= \text{CA} - \text{Spc}(s) - \text{Time}(\text{O}(t))\,.\end{aligned}$$

Under additional constructibility conditions it is even possible to get a generalization of Theorem 5:

Theorem 8 *For all functions $s(n) \geq n$, $t(n) \geq n$, and $h(n) \geq 1$, where h is fully* PTM *processor constructable in space s, time t, and with h processors, holds:*

$$\begin{aligned}&\text{CA} - \text{Spc}(\text{O}(s)) - \text{Time}(\text{O}(t)) \\ &\subseteq \text{PTM}-\text{Spc}(\text{O}(s))-\text{Time}(\text{O}(st/h))-\text{Proc}(\text{O}(h)).\end{aligned}$$

Decreasing the processor complexity indeed leads to the expected slowdown.

Relations Between PTM Complexity Classes (part 1) The interesting question now is whether different upper bounds on the processor complexity result in different computational power. In general that is not the case, as PTM with only one CU are TM and hence computationally universal. (As a side remark we note that therefore processor complexity cannot be a Blum measure. In fact that should be more or less clear since, e. g., deciding whether a second CU will ever be generated might require finding out whether the first CU, i. e., a TM, ever reaches a specific state.)

In this first part we consider the case where two complexity measures are allowed to grow in order to get a hierarchy. Results which only need one growing measure are the topic of the second part.

First of all it turns out that for fixed processor complexity between $\log n$ and $s(n)$ there is a space/time hierarchy:

Theorem 9 *Let s and t be two functions such that s is fully* PTM *space constructable in time t and t is* PTM *computable in space s and time t and let $h \geq \log$. Then:*

$$\begin{aligned}\bigcup_{\gamma \notin \text{O}(1)} \text{PTM} - \text{Spc}(\Theta(s/\gamma)) - \text{Time}(\Theta(t/\gamma)) \\ - \text{Proc}(\text{O}(h)) \\ \subsetneqq \text{PTM} - \text{Spc}(\text{O}(s)) - \text{Time}(\text{O}(t)) - \text{Proc}(\text{O}(h))\,.\end{aligned}$$

The proof of this theorem applies the usual idea of diagonalization. Technical details can be found in [34].

Instead of keeping processor complexity fixed and letting space complexity grow, one can also do the opposite. As for analogous results for TM, one needs the additional restriction to one *fixed tape alphabet*. One gets the following result, where the complexity classes carry the additional information about the size of the tape alphabet.

Theorem 10 *Let s, t and h be three functions such that s is fully* PTM *space constructable in time t and with h processors, and that t and h are* PTM *computable in space s and time t and with h processors such that in all cases the tape is not written. Let $b \geq 2$ be the size of the tape alphabet. Then:*

$$\begin{aligned}\bigcup_{\gamma \notin \text{O}(1)} \text{PTM} - \text{Spc}(s) - \text{Time}\Theta(t/\gamma) - \text{Proc}(h/\gamma) \\ - \text{Alph}(b) \\ \subsetneqq \text{PTM} - \text{Spc}(s) - \text{Time}(\Theta(st)) - \text{Proc}(\Theta(h)) \\ - \text{Alph}(b)\,.\end{aligned}$$

Again we do not go into the details of the constructibility definitions which can be found in [34]. The important point here is that one can prove that increasing time and processor complexity by a non-constant factor does increase the (language recognition) capabilities of PTM, even if the space complexity is fixed, provided that one does not allow any changes to the tape alphabet. In particular the theorem holds for the case $\text{space}(n) = n$.

It is now interesting to reconsider Open Problem 2. Let's *assume* that

$$\text{CA} - \text{Spc}(n) - \text{Time}(n) = \text{CA} - \text{Spc}(n) - \text{Time}(2^{\text{O}(n)})\,.$$

One may choose $\gamma(n) = \log n$, $t(n) = 2^{n/\log n}$ and $h(n) = n$ in Theorem 10. Using that together with Theorem 7 the assumption would give rise to

$$\begin{aligned}\text{PTM} - \text{Spc}(n) - \text{Time}\left(\frac{2^{n/\log n}}{\log n}\right) - \text{Proc}\left(\frac{n}{\log n}\right) \\ - \text{Alph}(b) \\ \subsetneqq \text{PTM} - \text{Spc}(n) - \text{Time}(n2^{n/\log n}) - \text{Proc}(n) \\ - \text{Alph}(b) \\ = \text{PTM} - \text{Spc}(n) - \text{Time}(n) - \text{Proc}(n) \\ - \text{Alph}(b)\,.\end{aligned}$$

If the polynomial time hierarchy for n-space bounded CA collapses, then there are languages which cannot be recognized by PTM in almost exponential time with $n/\log n$ processors but which can be recognized by PTM with n processors in linear time, *if the tape alphabet is fixed*.

Relations Between PTM Complexity Classes (part 2) One can get rid of the fixed alphabet condition by using a combinatorial argument for a specific formal language (instead of diagonalization) and even have not only the space but also the processor complexity fixed and still get a time hierarchy. The price to pay is that the range of time bounds is more restricted than in Theorem 10.

Consider the formal language

$$L_{\mathrm{vv}} = \left\{ v\mathtt{c}^{|v|}v \mid v \in \{\mathtt{a},\mathtt{b}\}^{+} \right\} .$$

It contains all words which can be divided into three segments of equal length such that the first and third are identical. Intuitively whatever type of machine is used for recognition, it is unavoidable to "move" the complete information from one end to the other. L_{vv} shares this feature with L_{pal}.

Using a counting argument inspired by Hennie's concept of crossing sequences [10] applied to L_{pal}, one can show:

Lemma 11 ([34]) *If P is a PTM recognizing L_{vv}, then* $\mathrm{time}_P^2 \cdot \mathrm{proc}_P \in \Omega(n^3/\log^2 n)$.

On the other hand, one can construct a PTM recognizing L_{vv} with processor complexity n^a for sufficiently nice a:

Lemma 12 ([34]) *For each $a \in \mathbb{Q}$ with $0 < a < 1$ holds:*

$$L_{\mathrm{vv}} \in \mathrm{PTM{-}S}{\small PC}(n)\mathrm{{-}T}{\small IME}\left(\Theta\left(n^{2-a}\right)\right)\mathrm{{-}P}{\small ROC}\left(\Theta\left(n^{a}\right)\right) .$$

Putting these lemmas together yields another hierarchy theorem:

Theorem 13 *For rational numbers $0 < a < 1$ and $0 < \varepsilon < 3/2 - a/2$ holds:*

$$\begin{aligned} &\mathrm{PTM} - \mathrm{S}{\small PC}(n) - \mathrm{T}{\small IME}\left(\Theta\left(n^{3/2-a/2-\varepsilon}\right)\right) \\ &\qquad\qquad - \mathrm{P}{\small ROC}\left(\Theta\left(n^{a}\right)\right) \\ &\subsetneqq \mathrm{PTM} - \mathrm{S}{\small PC}(n) - \mathrm{T}{\small IME}\left(\Theta\left(n^{2-a}\right)\right) \\ &\qquad\qquad - \mathrm{P}{\small ROC}\left(\Theta\left(n^{a}\right)\right) . \end{aligned}$$

Hence, for a close to 1 a "small" increase in time by some $n^{\varepsilon'}$ suffices to increase the recognition power of PTM while the processor complexity is fixed at n^a and the space complexity is fixed at n as well.

Open Problem 14 For the recognition of L_{vv} there is a gap between the lower bound of $\mathrm{time}_P^2 \cdot \mathrm{proc}_P \in \Omega(n^3/\log^2 n)$ in Lemma 11 and the upper bound of $\mathrm{time}_P^2 \cdot \mathrm{proc}_P \in \mathrm{O}(n^{4-a})$ in Lemma 12. It is not known whether the upper or the lower bound or both can be improved.

An even more difficult problem is to prove a similar result for the case $a = 1$, i. e., cellular automata, as mentioned in Open Problem 2.

State Change Complexity

In CMOS technology, what costs most of the energy is to make a *proper state change*, from zero to one or from one to zero. Motivated by this fact Vollmar [30] introduced the state change complexity for CA.

There are two variants based on the same idea: Given a halting CA computation for an input w and a cell i one can count the number of time points τ, $1 \leq \tau \leq \mathrm{time}'(w)$, such that cell i is in different states at times $\tau - 1$ and τ. Denote that number by $\mathrm{change}'(w, i)$. Define

$$\mathrm{maxchg}'(w) = \max_{i \in G} \mathrm{change}'(w, i) \quad \text{and}$$
$$\mathrm{sumchg}'(w) = \sum_{i \in G} \mathrm{change}'(w, i)$$

and

$$\mathrm{maxchg}(n) = \max\{\mathrm{maxchg}'(w) \mid w \in A^n\}$$
$$\mathrm{sumchg}(n) = \max\{\mathrm{sumchg}'(w) \mid w \in A^n\} .$$

For the language L_{vv} which already played a role in the previous subsection one can show:

Lemma 15 *Let $f(n)$ be a non-decreasing function which is not in $\mathrm{O}(\log n)$, i. e., $\lim_{n\to\infty} \log n / f(n) = 0$. Then any CA C recognizing L_{vv} makes a total of at least $\Omega(n^2/f(n))$ state changes in the segment containing the n input cells and n cells to the left and to the right of them.*

In particular, if $\mathrm{time}_C \in \Theta(n)$, then $\mathrm{sumchg}_C \in \Omega(n^2/f(n))$.

Furthermore $\mathrm{maxchg}_C \in \Omega(n/f(n))$.

In the paper by Sanders et al. [24] a generalization of this lemma to d-dimensional CA is proved.

Open Problem 16 While the processor complexity of PTM measures how many activities happen simultaneously "across space", state change complexity measures how many activities happen over time. For both cases we have made use of the same formal language in proofs. That might be an indication that there are connections between the two complexity measures. But no non-trivial results are known until now.

Asynchronous CA

Until now we have only considered one global mode of operation: the so-called *synchronous* case, where in each global step of the CA all cells must update their states synchronously. Several models have been considered where this requirement has been relaxed.

Generally speaking, asynchronous CA are characterized by the fact that in one global step of the CA *some* cells are active and do update their states (all according to the same local transition function) while others do nothing, i. e., remain in the same state as before. There are then different approaches to specify some restrictions on which cells may be active or not.

Asynchronous update mode. The simplest possibility is to not quantify anything and to say that a configuration c' is a legal successor of configuration c, denoted $c \vdash c'$, iff for all $i \in G$ one has $c'(i) = c(i)$ or $c'(i) = f(c_{i+N})$.

Unordered sequential update mode. In this special case it is required that there is only one active cell in each global step, i. e., $\text{card}(\{i | c'(i) \neq c(i)\}) \leq 1$.

Since CA with an asynchronous update mode are no longer deterministic from a global (configuration) point of view, it is not completely clear how to define, e. g., formal language recognition and time complexity. Of course one could follow the way it is done for nondeterministic TM. To the best of our knowledge this has not considered for asynchronous CA. (There are results for general nondeterministic CA; see for example ► Cellular Automata and Language Theory.)

It should be noted that Nakamura [20] has provided a very elegant construction for simulating a CA C_s with synchronous update mode on a CA C_a with one of the above asynchronous update modes. Each cell stores the "current" and the "previous" state of a C_s-cell before its last activation and a counter value T modulo 3 ($Q_a = Q_s \times Q_s \times \{0, 1, 2\}$). The local transition function f_a is defined in such a way that an activated cell does the following:

- T always indicates how often a cell has already been updated modulo 3.
- If the counters of all neighbors have value T or $T + 1$, the current C_s-state of the cell is remembered as previous state and a new current state is computed according to f_s from the current and previous C_s-states of the neighbors; the selection between current and previous state depends on the counter value of that cell. In this case the counter is incremented.
- If the counter of at least one neighboring cell is at $T - 1$, the activated cell keeps its complete state as it is.

Therefore, if one does want to gain something using asynchronous CA, their local transition functions would have to be designed for that specific usage.

Recently, interest has increased considerably in CA where the "degree of (a-)synchrony" is quantified via probabilities. In these cases one considers CA with only a finite number of cells.

Probabilistic update mode. Let $0 \leq \alpha \leq 1$ be a probability. In probabilistic update mode each legal global step $c \vdash c'$ of the CA is assigned a probability by requiring that each cell i independently has a probability α of updating its state.

Random sequential update mode. This is the case when in each global step one of the cells in G is chosen with even probability and its state updated, while all others do not change their state. CA operating in this mode are called *fully asynchronous* by some authors.

These models can be considered special cases of what is usually called *probabilistic* or *stochastic CA*. For these CA the local transition function is no longer a map from Q^N to Q, but from Q^N to $[0; 1]^Q$. For each $\ell \in Q^N$ the value $f(\ell)$ is a probability distribution for the next state (satisfying $\sum_{q \in Q} f(\ell)(q) = 1$). There are only very few papers about formal language recognition with probabilistic CA; see [18].

On the other hand, probabilistic update modes have received some attention recently. See for example [22] and the references therein. Development of this area is still at its beginning. Until now, specific local rules have mainly been investigated; for an exception see [7].

Communication in CA

Until now we have considered only one-dimensional Euclidean CA where one bit of information can reach $O(t)$ cells in t steps. In this section we will have a look at a few possibilities for changing the way cells communicate in a CA.

First we have a quick look at CA where the underlying grid is $\mathbb{Z}^d$. The topic of the second subsection is CA where the cells are connected to form a tree.

Different Dimensionality

In $\mathbb{Z}^d$ − CA with, e. g., von Neumann neighborhood of radius 1, a cell has the potential to influence $O(t^d)$ cells in t steps. This is a polynomial number of cells. It comes as no surprise that

$$\mathbb{Z}^d - \text{CA} - \textsc{Spc}(s) - \textsc{Time}(t) \subseteq \text{TM} - \textsc{Spc}(\text{Pol}(s)) - \textsc{Time}(\text{Pol}(t))$$

and hence $\mathbb{Z}^d - \text{CA}$ are in the first machine class. One might only wonder why for the TM a space bound of $\Theta(s)$ might not be sufficient. This is due to the fact that the shape of the cells actually used by the CA might have an "irregular" structure and that the TM has to perform some bookkeeping or simulate a whole (hyper-)rectangle of cells encompassing all that are really used by the CA.

Trivially, a d-dimensional CA can be simulated on a d'-dimensional CA, where $d' > d$. The question is how much one loses when *decreasing* the dimensionality. The currently known best result in this direction is by Scheben [26]:

Theorem 17 *It is possible to simulate a d'-dimensional CA with running time t on a d-dimensional CA, $d < d'$ with running time and space* $\mathrm{O}(t^{2^{\mathrm{ld}\lceil d'/d \rceil}})$.

It should be noted that the above result is not directly about language recognition; the redistribution of input symbols needed for the simulation is not taken into account. Readers interested in that as well are referred to [1].

Open Problem 18 Try to find simulations of lower on higher dimensional CA which somehow make use of the "higher connectivity" between cells. It is probably much too difficult or even impossible to hope for general speedups. But efficient use of space (small hypercubes) for computations without losing time might be achievable.

Tree CA and hyperbolic CA

Starting from the root of a full binary tree one can reach an exponential number 2^t of nodes in t steps. If there are some computing capabilities related to the nodes, there is at least the possibility that such a device might exhibit some kind of strong parallelism.

One of the earliest papers in this respect is Wiedermann's article [31] (unfortunately only available in Slovak). The model introduced there one would now call parallel Turing machines, where a tape is not a linear array of cells, but where the cells are connected in such a way as to form a tree. A proof is sketched, showing that these devices can simulate PRAMs in linear time (assuming the so-called logarithmic cost model). PRAMs are in the second machine class.

So, indeed in some sense, trees are powerful. Below we first quickly introduce a **PSPACE**-complete problem which is a useful tool in order to prove the power of computational models involving trees. A few examples of such models are considered afterwards.

Quantified Boolean Formula The instances of the problem *Quantified Boolean Formula* (QBF, sometimes also called QSAT) have the structure

$$Q_1 x_1 Q_2 x_2 \cdots Q_k x_k \colon F(x_1, \ldots, x_k) .$$

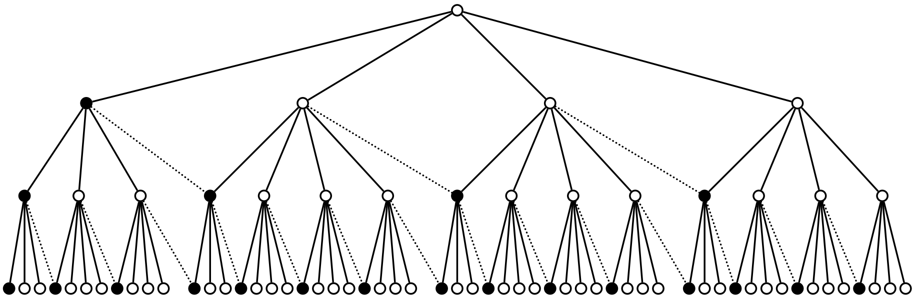
Here $F(x_1, \ldots, x_k)$ is a Boolean formula with variables $x_1, \ldots, x_k$ and connectives $\wedge$, $\vee$ and $\neg$. Each Q_j is one of the quantifiers $\forall$ or $\exists$. The problem is to decide whether the formula is true under the obvious interpretation. This problem is known to be complete for **PSPACE**. All known TM, i. e., all deterministic sequential algorithms, for solving QBF require exponential time.

Thus a proof that QBF can be solved by some model $\mathcal{M}$ in polynomial time (usually) implies that all problems in **PSPACE** can be solved by $\mathcal{M}$ in polynomial time. Often this can be paired with the "opposite" results that problems that can be solved in polynomial time on $\mathcal{M}$ are in **PSPACE**, and hence TM $-$ **PSPACE** $= \mathcal{M} -$ **P**.

This, of course, is not the case only for models "with trees"; see [6] for many alternatives.

Tree CA A tree CA (TCA for short) working on a full d-ary tree can be defined as follows: There is a set of states Q. For the root there is a local transition function $f_0 \colon (A \cup \{\square\}) \times Q \times \mathbb{Q}^d \to Q$, which uses an input symbol (if available), the root cell's own state and those of the d child nodes to compute the next state of the node. And there are d local transition functions $f_i \colon Q \times Q \times Q^d \to Q$, where $1 \leq i \leq d$. The ith child of a node uses f_i to compute its new state depending the state of its parent node, its own state and the states of the d child nodes. For language recognition, input is provided sequentially to the root node during the first n steps and a blank symbol $\square$ afterwards. A word is accepted if the root node enters an accepting state from a designated subset $F_+ \subset Q$.

Mycielski and Niwiński [19] were the first to realize that sequential polynomial reductions can be carried out by TCA and that QBF can be recognized by tree CA as well: A formula to be checked with k variables is copied and distributed to 2^k "evaluation cells". The sequences of left/right choices of the paths to them determine a valuation of the variables with zeros and ones. Each evaluation cell uses the subtree below it to evaluate $F(x_1, \ldots, x_k)$ accordingly. The results are propagated up to the root. Each cell in level i below the root, $1 \leq i \leq k$ above an evalua-



Cellular Automata as Models of Parallel Computation, Figure 3
The first levels of a tree of cells resulting from a tiling of the hyperbolic plane with 6-gons

tion cell combines the results using $\vee$ or $\wedge$, depending on whether the ith quantifier of the formula was $\exists$ or $\forall$.

On the other hand, it is routine work to prove that the result of a TCA running in polynomial time can be computed sequentially in polynomial space by a depth first procedure. Hence one gets:

Theorem 19

$$\mathrm{TCA} - \mathrm{T{\small IME}}(\mathrm{Pol}(n)) = \boldsymbol{PSPACE}$$

Thus tree cellular automata are in the second machine class.

Hyperbolic CA Two-dimensional CA as defined in Sect. "Introduction" can be considered as arising from the tessellation of the Euclidean plane $\mathbb{Z}^2$ with squares. Therefore, more generally, sometimes CA on a grid $G = \mathbb{Z}^d$ are called Euclidean CA. Analogously, some *Hyperbolic CA* arise from tessellation of the hyperbolic plane with some regular polygon. They are covered in depth in a separate article (► Cellular Automata in Hyperbolic Spaces).

Here we just consider one special case: The two-dimensional hyperbolic plane can be tiled with copies of the regular 6-gon with six right angles. If one considers only one quarter and draws a graph with the tiles as nodes and links between those nodes who share a common tile edge, one gets the graph depicted in Fig. 3. Basically it is a tree with two types of nodes, *black* and *white* ones, and some "additional" edges depicted as dotted lines. The root is a white node. The first child of each node is black. All other children are white; a black node has 2 white children, a white node has 3 white children.

For *hyperbolic CA* (HCA) one uses the formalism analogously to that described for tree CA. As one can see, basically HCA are trees with some additional edges. It is therefore not surprising that they can accept the languages from **PSPACE** in polynomial time. The inverse inclusion is also proved similarly to the tree case. This gives:

Theorem 20

$$\mathrm{HCA} - \mathrm{T{\small IME}}(\mathrm{Pol}(n)) = \boldsymbol{PSPACE}$$

It is also interesting to have a look at the analogs of **P**, **PSPACE**, and so on for hyperbolic CA. Somewhat surprisingly Iwamoto et al. [11] have shown:

Theorem 21

$$\begin{aligned}\mathrm{HCA} - \mathrm{T{\small IME}}(\mathrm{Pol}(n)) &= \mathrm{HCA} - \mathrm{S{\small PC}}(\mathrm{Pol}(n)) \\ &= \mathrm{NHCA} - \mathrm{T{\small IME}}(\mathrm{Pol}(n)) = \mathrm{NHCA} - \mathrm{S{\small PC}}(\mathrm{Pol}(n))\end{aligned}$$

where NHCA *denotes nondeterministic hyperbolic CA. The analogous equalities hold for exponential time and space.*

Outlook There is yet another possibility for bringing trees into play: trees of configurations. The concept of *alternation* [4] can be carried over to cellular automata. Since there are several active computational units, the definitions are little bit more involved and it turns out that one has several possibilities which also result in models with slightly different properties. But in all cases one gets models from the second machine class. For results, readers are referred to [13] and [23].

On the other end there is some research on what happens if one restricts the possibilities for communication between neighboring cells: Instead of getting information about the complete states of the neighbors, in the extreme case only *one bit* can be exchanged. See for example [15].

Future Directions

At several points in this paper we have pointed out open problems which deserve further investigation. Here, we want to stress three areas which we consider particularly interesting in the area of "CA as a parallel model".

Proper Inclusions and Denser Hierarchies

It has been pointed out several times, that inclusions of complexity classes are not known to be proper, or that gaps between resource bounds still needed to be "large" in order to prove that the inclusion of the related classes is a proper one. For the foreseeable future it remains a wide area for further research. Most probably new techniques will have to be developed to make significant progress.

Activities

The motivation for considering state change complexity was the energy consumption of CMOS hardware. It is known that irreversible physical computational processes must consume energy. This seems not to be the case for reversible ones. Therefore *reversible* CA are also interesting in this respect. The definition of reversible CA and results for them are the topic of the article by Morita (▶ Reversible Cellular Automata), also in this encyclopedia. Surprisingly, all currently known simulations of irreversible CA on reversible ones (this is possible) exhibit a *large* state change complexity. This deserves further investigation.

Also the examination of CA which are "reversible on the computational core" has been started only recently [16]. There are first surprising results; the impacts on computational complexity are unforeseeable.

Asynchronicity and Randomization

Randomization is an important topic in sequential computing. It is high time that this is also investigated in much more depth for cellular automata. The same holds for cellular automata where not all cells are updating their states synchronously. These areas promise a wealth of new insights into the essence of fine-grained parallel systems.

Bibliography

1. Achilles AC, Kutrib M, Worsch T (1996) On relations between arrays of processing elements of different dimensionality. In: Vollmar R, Erhard W, Jossifov V (eds) Proceedings Parcella '96, no. 96 in Mathematical Research. Akademie, Berlin, pp 13–20
2. Blum M (1967) A machine-independent theory of the complexity of recursive functions. J ACM 14:322–336
3. Buchholz T, Kutrib M (1998) On time computability of functions in one-way cellular automata. Acta Inf 35(4):329–352
4. Chandra AK, Kozen DC, Stockmeyer LJ (1981) Alternation. J ACM 28(1):114–133
5. Delorme M, Mazoyer J, Tougne L (1999) Discrete parabolas and circles on 2D cellular automata. Theor Comput Sci 218(2):347–417
6. van Emde Boas P (1990) Machine models and simulations. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science, vol A. Elsevier Science Publishers and MIT Press, Amsterdam, chap 1, pp 1–66
7. Fatès N, Thierry É, Morvan M, Schabanel N (2006) Fully asynchronous behavior of double-quiescent elementary cellular automata. Theor Comput Sci 362:1–16
8. Garzon M (1991) Models of Massive Parallelism. Texts in Theoretical Computer Science. Springer, Berlin
9. Hemmerling A (1979) Concentration of multidimensional tape-bounded systems of Turing automata and cellular spaces. In: Budach L (ed) International Conference on Fundamentals of Computation Theory (FCT '79). Akademie, Berlin, pp 167–174
10. Hennie FC (1965) One-tape, off-line Turing machine computations. Inf Control 8(6):553–578
11. Iwamoto C, Margenstern M (2004) Time and space complexity classes of hyperbolic cellular automata. IEICE Trans Inf Syst E87-D(3):700–707
12. Iwamoto C, Hatsuyama T, Morita K, Imai K (2002) Constructible functions in cellular automata and their applications to hierarchy results. Theor Comput Sci 270(1–2):797–809
13. Iwamoto C, Tateishi K, Morita K, Imai K (2003) Simulations between multi-dimensional deterministic and alternating cellular automata. Fundamenta Informaticae 58(3/4):261–271
14. Kutrib M (2008) Efficient pushdown cellular automata: Universality, time and space hierarchies. J Cell Autom 3(2):93–114
15. Kutrib M, Malcher A (2006) Fast cellular automata with restricted inter-cell communication: Computational capacity. In: Navarro YKG, Bertossi L (ed) Proceedings Theor Comput Sci (IFIP TCS 2006), pp 151–164
16. Kutrib M, Malcher A (2007) Real-time reversible iterative arrays. In: Csuhaj-Varjú E, Ésik Z (ed) Fundametals of Computation Theory 2007. LNCS vol 4639. Springer, Berlin, pp 376–387
17. Mazoyer J, Terrier V (1999) Signals in one-dimensional cellular automata. Theor Comput Sci 217(1):53–80
18. Merkle D, Worsch T (2002) Formal language recognition by stochastic cellular automata. Fundamenta Informaticae 52(1–3):181–199
19. Mycielski J, Niwiński D (1991) Cellular automata on trees, a model for parallel computation. Fundamenta Informaticae XV:139–144
20. Nakamura K (1981) Synchronous to asynchronous transformation of polyautomata. J Comput Syst Sci 23:22–37
21. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Champaign. Edited and completed by Arthur W. Burks
22. Regnault D, Schabanel N, Thierry É (2007) Progress in the analysis of stochastic 2d cellular automata: a study of asychronous 2d minority. In: Csuhaj-Varjú E, Ésik Z (ed) Fundametals of Computation Theory 2007. LNCS vol 4639. Springer, Berlin, pp 376–387
23. Reischle F, Worsch T (1998) Simulations between alternating CA, alternating TM and circuit families. In: MFCS'98 satellite workshop on cellular automata, pp 105–114
24. Sanders P, Vollmar R, Worsch T (2002) Cellular automata: Energy consumption and physical feasibility. Fundamenta Informaticae 52(1–3):233–248
25. Savitch WJ (1978) Parallel and nondeterministic time complexity classes. In: Proc. 5th ICALP, pp 411–424
26. Scheben C (2006) Simulation of d'-dimensional cellular automata on d-dimensional cellular automata. In: El Yacoubi S, Chopard B, Bandini S (eds) Proceedings ACRI 2006. LNCS, vol 4173. Springer, Berlin, pp 131–140
27. Smith AR (1971) Simple computation-universal cellular spaces. J ACM 18(3):339–353

28. Stoß HJ (1970) *k*-Band-Simulation von *k*-Kopf-Turing-Maschinen. Computing 6:309–317
29. Stratmann M, Worsch T (2002) Leader election in *d*-dimensional CA in time *diam* · log(*diam*). Future Gener Comput Syst 18(7):939–950
30. Vollmar R (1982) Some remarks about the 'efficiency' of polyautomata. Int J Theor Phys 21:1007–1015
31. Wiedermann J (1983) Paralelný Turingov stroj – Model distribuovaného počítača. In: Gruska J (ed) Distribouvané a parallelné systémy. CRC, Bratislava, pp 205–214
32. Wiedermann J (1984) Parallel Turing machines. Tech. Rep. RUU-CS-84-11. University Utrecht, Utrecht
33. Worsch T (1997) On parallel Turing machines with multi-head control units. Parallel Comput 23(11):1683–1697
34. Worsch T (1999) Parallel Turing machines with one-head control units and cellular automata. Theor Comput Sci 217(1):3–30

Cellular Automata, Classification of

Klaus Sutner
Carnegie Mellon University, Pittsburgh, USA

Article Outline

Glossary

Cellular automaton For our purposes, a (one-dimensional) cellular automaton (CA) is given by a local map $\rho : \Sigma^w \to \Sigma$ where Σ is the underlying alphabet of the automaton and w is its width. As a data structure, suitable as input to a decision algorithm, a CA can thus be specified by a simple lookup table. We abuse notation and write $\rho(x)$ for the result of applying the global map of the CA to configuration $x \in \Sigma^{\mathbb{Z}}$.

Wolfram classes Wolfram proposed a heuristic classification of cellular automata based on observations of typical behaviors. The classification comprises four classes: evolution leads to trivial configurations, to periodic configurations, evolution is chaotic, evolution leads to complicated, persistent structures.

Undecidability It was recognized by logicians and mathematicians in the first half of the 20th century that there is an abundance of well-defined problems that cannot be solved by means of an algorithm, a mechanical procedure that is guaranteed to terminate after finitely many steps and produce the appropriate answer. The best known example of an undecidable problem is Turing's Halting Problem: there is no algorithm to determine whether a given Turing machine halts when run on an empty tape.

Semi-decidability A problem is said to be semi-decidable or computably enumerable if it admits an algorithm that returns "yes" after finitely many steps if this is indeed the correct answer. Otherwise the algorithm never terminates. The Halting Problem is the standard example for a semi-decidable problem. A problem is decidable if, and only if, the problem itself and its negation are semi-decidable.

Universality A computational device is universal it is capable of simulating any other computational device. The existence of universal computers was another central insight of the early days of computability theory and is closely related to undecidability.

Reversibility A discrete dynamical system is reversible if the evolution of the system incurs no loss of information: the state at time t can be recovered from the state at time $t + 1$. For CAs this means that the global map is injective.

Surjectivity The global map of a CA is surjective if every configuration appears as the image of another. By contrast, a configuration that fails to have a predecessor is often referred to as a Garden-of-Eden.

Finite configurations One often considers CA with a special quiescent state: the homogeneous configuration where all cells are in the quiescent state is required to be fixed point under the global map. Infinite configurations where all but finitely many cells are in the quiescent state are often called finite configurations. This is somewhat of a misnomer; we prefer to speak about configurations with finite support.

Definition of the Subject

Cellular automata display a large variety of behaviors. This was recognized clearly when extensive simulations of cellular automata, and in particular one-dimensional CA, became computationally feasible around 1980. Surprisingly, even when one considers only elementary CA, which are constrained to a binary alphabet and local maps involving only nearest neighbors, complicated behaviors are observed in some cases. In fact, it appears that most behaviors observed in automata with more states and larger neighborhoods already have qualitative analogues in the realm of elementary CA. Careful empirical studies lead Wolfram to suggest a phenomenological classification of CA based on the long-term evolution of configurations, see [68,71] and Sect. "Introduction". While Wolfram's four classes clearly capture some of the behavior of CA it turns out that any attempt at formalizing this taxonomy meets with considerable difficulties. Even apparently simple questions about the behavior of CA turn out to be algorithmically undecidable and it is highly challenging to provide a detailed mathematical analysis of these systems.

Introduction

In the early 1980's Wolfram published a collection of 20 open problems in the the theory of CA, see [69]. The first problem on his list is "What overall classification of cellular automata behavior can be given?" As Wolfram points out, experimental mathematics provides a first answer to this problem: one performs a large number of explicit simulations and observes the patterns associated with the long term evolution of a configuration, see [67,71]. Wolfram proposed a classification that is based on extensive simulations in particular of one-dimensional cellular automata where the evolution of a configuration can be visualized naturally as a two-dimensional image. The classification involves four classes that can be described as follows:

- *W1*: Evolution leads to homogeneous fixed points.
- *W2*: Evolution leads to periodic configurations.
- *W3*: Evolution leads to chaotic, aperiodic patterns.
- *W4*: Evolution produces persistent, complex patterns of localized structures.

Thus, Wolfram's first three classes follow closely concepts from continuous dynamics: fixed point attractors, periodic attractors and strange attractors, respectively. They correspond roughly to systems with zero temporal and spatial entropy, zero temporal entropy but positive spatial entropy, and positive temporal and spatial entropy, respectively. *W4* is more difficult to associate with a continuous analogue except to say that transients are typically very long. To understand this class it is preferable to consider CA as models of massively parallel computation rather than as particular discrete dynamical systems. It was conjectured by Wolfram that *W4* automata are capable of performing complicated computations and may often be computationally universal. Four examples of elementary CA that are typical of the four classes are shown in Fig. 1. Li and Packard [32,33] proposed a slightly modified version of this hierarchy by refining the low classes and in particular Wolfram's *W2*. Much like Wolfram's classification, the Li–Packard classification is concerned with the asymptotic behavior of the automaton, the structure and behavior of the limiting configurations. Here is one version of the Li–Packard classification, see [33].

- *LP1*: Evolution leads to homogeneous fixed points.
- *LP2*: Evolution leads to non-homogeneous fixed points, perhaps up a to a shift.
- *LP3*: Evolution leads to ultimately periodic configurations. Regions with periodic behavior are separated by domain walls, possibly up to a shift.
- *LP4*: Configurations produce locally chaotic behavior. Regions with chaotic behavior are separated by domain walls, possibly up to a shift.
- *LP5*: Evolution leads to chaotic patterns that are spatially unbounded.
- *LP6*: Evolution is complex. Transients are long and lead to complicated space-time patterns which may be nonmonotonic in their behavior.

By contrast, a classification closer to traditional dynamical systems theory was introduced by Kůrka, see [27,28]. The classification rests on the notions of equicontinuity, sensitivity to initial conditions and expansivity. Suppose x is a point in some metric space and f a map on that space. Then f is *equicontinuous* at x if

$$\forall\, \varepsilon > 0\, \exists\, \delta > 0\, \forall\, y \in B_\delta(x), n \in \mathbb{N}\, (d(f^n(x), f^n(y)) < \varepsilon)$$

where $d(.,.)$ denotes a metric. Thus, all points in a sufficiently small neighborhood of x remain close to the iterates of x for the whole orbit. Global equicontinuity is a fairly strong condition, it implies that the limit set of the automaton is reached after finitely many steps. The map is *sensitive* (to initial conditions) if

$$\forall\, x, \varepsilon > 0\, \exists\, \delta > 0\, \forall\, y \in B_\delta(x) \exists\, n \in \mathbb{N} \\ (d(f^n(x), f^n(y)) \geq \varepsilon)\,.$$

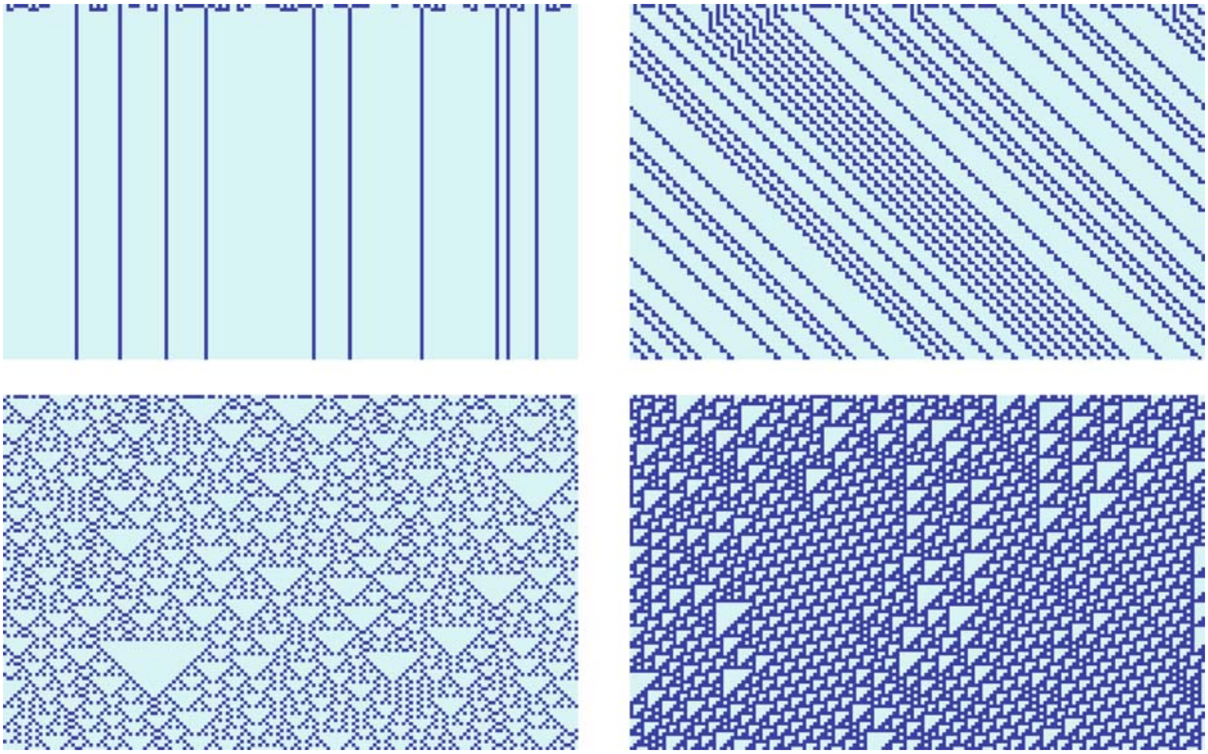
Lastly, the map is *positively expansive* if

$$\exists\, \varepsilon > 0\, \forall\, x \neq y\, \exists\, n \in \mathbb{N}\, (d(f^n(x), f^n(y)) \geq \varepsilon)\,.$$

Kůrka's classification then takes the following form.

- *K1*: All points are equicontinuous under the global map.
- *K2*: Some but not all points are equicontinuous under the global map.
- *K3*: The global map is sensitive but not positively expansive.
- *K4*: The global map is positively expansive.

This type of classification is perfectly suited to the analysis of uncountable spaces such as the Cantor space $\{0,1\}^{\mathbb{N}}$



Cellular Automata, Classification of, Figure 1
Typical examples of the behavior described by Wolfram's classes among elementary cellular automata

or the full shift space $\Sigma^{\mathbb{Z}}$ which carry a natural metric structure. For the most part we will not pursue the analysis of CA by topological and measure theoretic means here and refer to ▶ Topological Dynamics of Cellular Automata in this volume for a discussion of these methods. See Sect. "Definability and Computability" for the connections between topology and computability.

Given the apparent complexity of observable CA behavior one might suspect that it is difficult to pinpoint the location of an arbitrary given CA in any particular classification scheme with any precision. This is in contrast to simple parameterizations of the space of CA rules such as Langton's λ parameter that are inherently easy to compute. Briefly, the λ value of a local map is the fraction of local configurations that map to a non-zero value, see [29,33]. Small λ values result in short transients leading to fixed points or simple periodic configurations. As λ increases the transients grow longer and the orbits become more and more complex until, at last, the dynamics become chaotic. Informally, sweeping the λ value from 0 to 1 will produce CA in *W1*, then *W2*, then *W4* and lastly in *W3*. The last transition appears to be associated with a threshold phenomenon. It is unclear what the connection between Langton's λ-value and computational properties of a CA is, see [37,46]. Other numerical measures that appear to be loosely connected to classifications are the mean field parameters of Gutowitz [20,21], the Z-parameter by Wuensche [72], see also [44]. It seems doubtful that a structured taxonomy along the lines of Wolfram or Li–Packard can be derived from a simple numerical measure such as the λ value alone, or even from a combination of several such values. However, they may be useful as empirical evidence for membership in a particular class.

Classification also becomes significantly easier when one restricts one's attention to a limited class of CA such as additive CA, see ▶ Additive Cellular Automata. In this context, additive means that the local rule of the automaton has the form $\rho(\vec{x}) = \sum_i c_i x_i$ where the coefficients as well as the states are modular numbers. A number of properties starting with injectivity and surjectivity as well as topological properties such as equicontinuity and sensitivity can be expressed in terms of simple arithmetic conditions on the rule coefficients. For example, equicontinuity is equivalent to all prime divisors of the modulus m dividing all coefficients c_i, $i > 1$, see [35] and the references therein. It is also noteworthy that in the linear case methods tend to carry over to arbitrary dimensions; in general there is a significant step in complexity from dimension one to dimension two.

No claim is made that the given classifications are complete; in fact, one should think of them as prototypes rather than definitive taxonomies. For example, one might add the class of *nilpotent* CA at the bottom. A CA is nilpotent if all configurations evolve to a particular fixed point after finitely many steps. Equivalently, by compactness, there is a bound n such that all configurations evolve to the fixed point in no more than n steps. Likewise, we could add the class of *intrinsically universal* CA at the top. A CA is intrinsically universal if it is capable of simulating all other CA of the same dimension in some reasonable sense. For a fairly natural notion of simulation see [45]. At any rate, considerable effort is made in the references to elaborate the characteristics of the various classes. For many concrete CA visual inspection of the orbits of a suitable sample of configurations readily suggests membership in one of the classes.

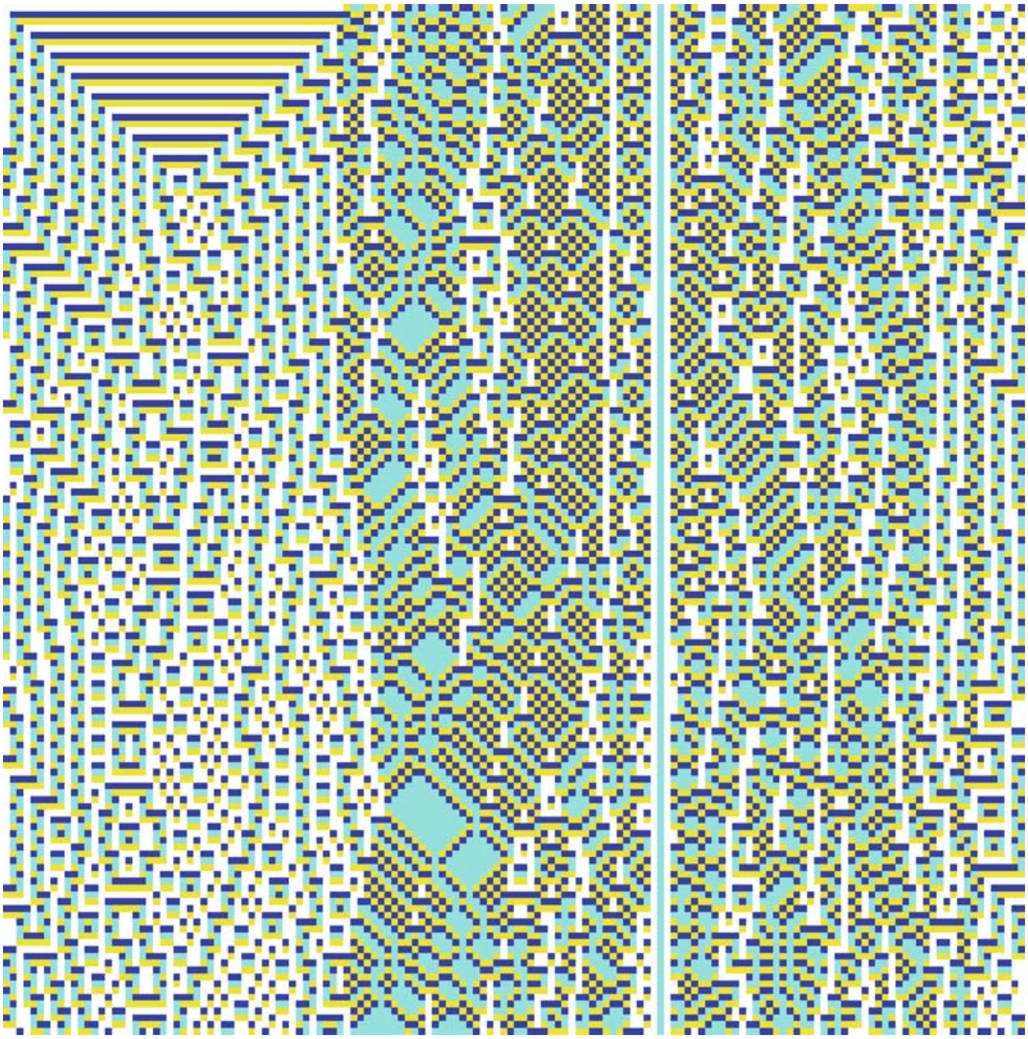
Reversibility and Surjectivity

A first tentative step towards the classification of a dynamical systems is to determine its reversibility or lack thereof. Thus we are trying to determine whether the evolution of the system is associated with loss of information, or whether it is possible to reconstruct the state of the system at time t from its state at time $t + 1$. In terms of the global map of the system we have to decide injectivity. Closely related is the question whether the global map is surjective, i. e., whether there is no Garden-of-Eden: every configuration has a predecessor under the global map. As a consequence, the limit set of the automaton is the whole space. It was shown of Hedlund that for CA the two notions are connected: every reversible CA is also surjective, see [24], ▶ Reversible Cellular Automata. As a matter of fact, reversibility of the global map of a CA implies openness of the global map, and openness implies surjectivity. The converse implications are both false. By a well-known theorem by Hedlund [24] the global maps of CA are precisely the continuous maps that commute with the shift. It follows from basic topology that the inverse global map of a reversible CA is again the global map of a suitable CA. Hence, the predecessor configuration of a given configuration can be reconstructed by another suitably chosen CA. For results concerning reversibility on the limit set of the automaton see [61].

From the perspective of complexity the key result concerning reversible systems is the work by Lecerf [30] and Bennett [7]. They show that reversible Turing machines can compute any partial recursive function, modulo a minor technical problem: In a reversible Turing machine there is no loss of information; on the other hand even simple computable functions are clearly irreversible in the sense that, say, the sum of two natural numbers does not determine these numbers uniquely. To address this issue one has to adjust the notion of computability slightly in the context of reversible computation: given a partial recursive function $f : \mathbb{N} \to \mathbb{N}$ the function $\widehat{f}(x) = \langle x, f(x) \rangle$ can be computed by a reversible Turing machine where $\langle ., . \rangle$ is any effective pairing function. If f itself happens to be injective then there is no need for the coding device and f can be computed by a reversible Turing machine directly. For example, we can compute the product of two primes reversibly. Morita demonstrated that the same holds true for one-dimensional cellular automata [38,40,62], ▶ Tiling Problem and Undecidability in Cellular Automata: reversibility is no obstruction to computational universality. As a matter of fact, any irreversible cellular automaton can be simulated by a reversible one, at least on configurations with finite support. Thus one should expect reversible CA to exhibit fairly complicated behavior in general.

For infinite, one-dimensional CA it was shown by Amoroso and Patt [2] that reversibility is decidable. Moreover, it is decidable if the the global map is surjective. An efficient practical algorithm using concepts of automata theory can be found in [55], see also [10,14,23]. The fast algorithm is based on interpreting a one-dimensional CA as a deterministic transducer, see [6,48] for background. The underlying semi-automaton of the transducer is a de Bruijn automaton $\mathcal{B}$ whose states are words in Σ^{w-1} where Σ is the alphabet of the CA and w is its width. The transitions are given by $ax \xrightarrow{c} xb$ where $a, b, c \in \Sigma$, $x \in \Sigma^{w-2}$ and $c = \rho(axb)$, ρ being the local map of the CA. Since $\mathcal{B}$ is strongly connected, the product automaton of $\mathcal{B}$ will contain a strongly connected component C that contains the diagonal D, an isomorphic copy of $\mathcal{B}$. The global map of the CA is reversible if, and only if, $C = D$ is the only non-trivial component. It was shown by Hedlund [24] that surjectivity of the global map is equivalent with local injectivity: the restriction of the map to configurations with finite support must be injective. The latter property holds if, and only if, $C = D$ and is thus easily decidable. Automata theory does not readily generalize to words of dimensions higher than one. Indeed, reversibility and surjectivity in dimensions higher than one are undecidable, see [26] and ▶ Tiling Problem and Undecidability in Cellular Automata in this volume for the rather intricate argument needed to establish this fact.

While the structure of reversible one-dimensional CA is well-understood, see ▶ Tiling Problem and Undecid-



Cellular Automata, Classification of, Figure 2

A reversible automaton obtained by applying Fredkin's construction to the irreversible elementary CA 77

ability in Cellular Automata, [16], and while there is an efficient algorithm to check reversibility, few methods are known that allow for the construction of interesting reversible CA. There is a noteworthy trick due to Fredkin that exploits the reversibility of the Fibonacci equation $X_{n+1} = X_n + X_{n-1}$. When addition is interpreted as exclusive or this can be used to construct a second-order CA from any given binary CA; the former can then be recoded as a first-order CA over a 4-letter alphabet. For example, for the open but irreversible elementary CA number 90 we obtain the CA shown in Fig. 2.

Another interesting class of reversible one-dimensional CA, the so-called *partitioned cellular automata (PCA)*, is due to Morita and Harao, see [38,39,40]. One can think of a PCA as a cellular automaton whose cells are divided into multiple tracks; specifically Morita uses an alphabet of the form $\Sigma = \Sigma_1 \times \Sigma_2 \times \Sigma_3$. The configurations of the automaton can be written as (X, Y, Z) where $X \in {\Sigma_1}^{\mathbb{Z}}$, $Y \in {\Sigma_2}^{\mathbb{Z}}$ and $Z \in {\Sigma_3}^{\mathbb{Z}}$. Now consider the *shearing map* σ defined by $\sigma(X, Y, Z) = (\mathsf{RS}(X), Y, \mathsf{LS}(Z))$ where RS and LS denote the right and left shift, respectively. Given any function $f : \Sigma \to \Sigma$ we can define a global map $f \circ \sigma$ where f is assumed to be applied point-wise. Since the shearing map is bijective, the CA will be reversible if, and only if, the map f is bijective. It is relatively easy to construct bijections f that cause the CA to perform particular computational tasks, even when a direct construction appears to be entirely intractable.

Definability and Computability

Formalizing Wolfram's Classes

Wolfram's classification is an attempt to categorize the complexity of the CA by studying the patterns observed during the long-term evolution of all configurations. The first two classes are relatively easy to observe, but it is difficult to distinguish between the last two classes. In particular *W4* is closely related to the kind of behavior that would be expected in connection with systems that are capable of performing complicated computations, including the ability to perform universal computation; a property that is notoriously difficult to check, see [52]. The focus on the full configuration space rather than a significant subset thereof corresponds to the worst-case approach well-known in complexity theory and is somewhat inferior to an average case analysis. Indeed, Baldwin and Shelah point out that a product construction can be used to design a CA whose behavior is an amalgamation of the behavior of two given CA, see [3,4]. By combining CA in different classes one obtains striking examples of the weakness of the worst-case approach. A natural example of this mixed type of behavior is elementary CA 184 which displays class II or class III behavior, depending on the initial configuration. Another basic example for this type of behavior is the well-studied elementary CA 30, see Sect. "Conclusion".

Still, for many CA a worst-case classification seems to provide useful information about the structural properties of the automaton. The first attempt at formalizing Wolfram's class was made by Culik and Yu who proposed the following hierarchy, given here in cumulative form, see [11]:

- *CY1*: All configurations evolve to a fixed point.
- *CY2*: All configurations evolve to a periodic configuration.
- *CY3*: The orbits of all configurations are decidable.
- *CY4*: No constraints.

The Culik–Yu classification employs two rather different methods. The first two classes can be defined by a simple formula in a suitable logic whereas the third (and the fourth in the disjoint version of the hierarchy) rely on notions of computability theory. As a general framework for both approaches we consider *discrete dynamical systems*, structures of the form $\mathcal{A} = \langle C, \to \rangle$ where $C \subseteq \Sigma^{\mathbb{Z}}$ is the space of *configurations* of the system and → is the "next configuration" relation on C. We will only consider the deterministic case where for each configuration x there exists precisely one configuration y such that $x \to y$. Hence we are really dealing with algebras with one unary function, but iteration is slightly easier to deal with in the relational setting. The structures most important in this context are the ones arising from a CA. For any local map ρ we consider the structure $\mathcal{A}_\rho = \langle C, \to \rangle$ where the next configuration relation is determined by $x \to \rho(x)$.

Using the standard language of first order logic we can readily express properties of the CA in terms of the system $\mathcal{A}_\rho$. For example, the system is reversible, respectively surjective, if the following assertions are valid over $\mathcal{A}$:

$$\forall\, x, y, z\, (x \to z \text{ and } y \to z \text{ implies } x = y)\,,$$
$$\forall\, x\, \exists\, y\, (y \to x)\,.$$

As we have seen, both properties are easily decidable in the one-dimensional case. In fact, one can express the basic predicate $x \to y$ (as well as equality) in terms of finite state machines on infinite words. These machines are defined like ordinary finite state machines but the acceptance condition requires that certain states are reached infinitely and co-infinitely often, see [8,19]. The emptiness problem for these automata is easily decidable using graph theoretic algorithms. Since regular languages on infinite words are closed under union, complementation and projection, much like their finite counterparts, and all the corresponding operations on automata are effective, it follows that one can decide the validity of first order sentences over $\mathcal{A}_\rho$ such as the two examples above: the model-checking problem for these structures and first order logic is decidable, see [34]. For example, we can decide whether there is a configuration that has a certain number of predecessors. Alternatively, one can translate these sentences into monadic second order logic of one successor, and use well-known automata-based decision algorithms there directly, see [8]. Similar methods can be used to handle configurations with finite support, corresponding to weak monadic second order logic. Since the complexity of the decision procedure is non-elementary one should not expect to be able to handle complicated assertions. On the other hand, at least for weak monadic second order logic practical implementations of the decision method exist, see [17]. There is no hope of generalizing this approach as the undecidability of, say, reversibility in higher dimensions demonstrates.

Write $x \xrightarrow{t} y$ if x evolves to y in exactly t steps, $x \xrightarrow{+} y$ if x evolves to y in any positive number of steps and $x \xrightarrow{*} y$ if x evolves to y in any number of steps. Note that $\xrightarrow{t}$ is definable for each fixed t, but $\xrightarrow{*}$ fails to be so definable in first order logic. This is in analogy to the undefinability of path existence problems in the first order theory of graphs, see [34]. Hence it is natural to extend our language so we can express iterations of the global map, either by adding transitive closures or by moving to some limited system of higher order logic over $\mathcal{A}_\rho$ where $\overset{*}{\rightarrow}$ is definable, see [8].

Arguably the most basic decision problem associated with a system $\mathcal{A}$ that requires iteration of the global map is the *Reachability Problem*: given two configurations x and y, does the evolution of x lead to y? A closely related but different question is the *Confluence Problem*: will two configurations x and y evolve to the same limit cycle? Confluence is an equivalence relation and allows for the decomposition of configuration space into limit cycles together with their basins of attraction. The Reachability and Confluence Problem amount to determining, given configurations x and y, whether

$$x \overset{*}{\rightarrow} y ,$$

$$\exists\, z\, (x \overset{*}{\rightarrow} z \text{ and } y \overset{*}{\rightarrow} z) ,$$

respectively. As another example, the first two Culik–Yu class can be defined like so:

$$\forall\, x\, \exists\, z\, (x \overset{*}{\rightarrow} z \text{ and } z \rightarrow z),$$

$$\forall\, x\, \exists\, z\, (x \overset{*}{\rightarrow} z \text{ and } z \overset{+}{\rightarrow} z) .$$

It is not difficult to give similar definitions for the lower Li–Packard classes if one extends the language by a function symbol denoting the shift operator.

The third Culik–Yu class is somewhat more involved. By definition, a CA lies in the third class if it admits a global decision algorithm to determine whether a given configuration x evolves to another given configuration y in a finite number of steps. In other words, we are looking for automata where the Reachability Problem is algorithmically solvable. While one can agree that *W4* roughly translates into undecidability and is thus properly situated in the hierarchy, it is unclear how chaotic patterns in *W3* relate to decidability. No method is known to translate the apparent lack of tangible, persistent patterns in rules such as elementary CA 30 into decision algorithms for Reachability. There is another, somewhat more technical problem to overcome in formalizing classifications. Recall that the full configuration space is $C = \Sigma^{\mathbb{Z}}$. Intuitively, given $x \in C$ we can effectively determine the next configuration $y = \rho(x)$. However, classical computability theory does not deal with infinitary objects such as arbitrary configuration so a bit of care is needed here. The key insight is that we can determine arbitrary finite segments of $\rho(x)$ using only finite segments of x (and, of course, the lookup table for the local map). There are several ways to model computability on $\Sigma^{\mathbb{Z}}$ based on this idea of finite approximations, we refer to [66] for a particularly appealing model based on so-called type-2 Turing machines; the reference also contains many pointers to the literature as well as a comparison between the different approaches. It is easy to see that for any CA the global map ρ as well as all its iterates ρ^t are computable, the latter uniformly in t. However, due to the finitary nature of all computations, equality is not decidable in type-2 computability: the unequal operator $U_0(x, y) = 0$ if $x \neq y$, $U_0(x, y)$ undefined otherwise, is computable and thus unequality is semi-decidable, but the stronger $U_0(x, y) = 0$ if $x \neq y$, $U_0(x, y) = 1$, otherwise, is not computable. The last result is perhaps somewhat counterintuitive, but it is inevitable if we strictly adhere to the finite approximation principle.

In order to avoid problems of this kind it has become customary to consider certain subspaces of the full configuration space, in particular C_{fin}, the collection of configurations with finite support, C_{per}, the collection of spatially periodic configurations and C_{ap}, the collection of almost periodic configurations of the form $\ldots uuuwvvv \ldots$ where u, v and w are all finite words over the alphabet of the automaton. Thus, an almost periodic configuration differs from a configuration of the form ${}^{\omega}u\, v^{\omega}$ in only finitely many places. Configurations with finite support correspond to the special case where $u = v = 0$ is a special quiescent symbol and spatially periodic configurations correspond to $u = v$, $w = \varepsilon$. The most general type of configuration that admits a finitary description is the class C_{rec} of recursive configurations, where the assignment of state to a cell is given by a computable function.

It is clear that all these subspaces are closed under the application of a global map. Except for C_{fin} there are also closed under inverse maps in the following sense: given a configuration y in some subspace that has a predecessor x in C_{all} there already exists a predecessor in the same subspace, see [55,58]. This is obvious except in the case of recursive configurations. The reference also shows that the recursive predecessor cannot be computed effectively from the target configuration. Thus, for computational purposes the dynamics of the cellular automaton are best reflected in C_{ap}: it includes all configuration with finite support and we can effectively trace an orbit in both directions. It is not hard to see that C_{ap} is the least such class. Alas, it is standard procedure to avoid minor technical difficulties arising from the infinitely repeated spatial patterns and establish classifications over the subspace C_{fin}. There is a arguably not much harm in this simplification since C_{fin} is a dense subspace of C_{all} and compactness can be used to lift properties from C_{fin} to the full configuration space.

The Culik–Yu hierarchy is correspondingly defined over C_{fin}, the class of all configurations of finite support. In this setting, the first three classes of this hierarchy are undecidable and the fourth is undecidable in the disjunctive version: there is no algorithm to test whether a CA admits undecidable orbits. As it turns out, the CA classes are complete in their natural complexity classes within the arithmetical hierarchy [50,52]. Checking membership in the first two classes comes down to performing an infinite number of potentially unbounded searches and can be described logically by a Π_2 expression, a formula of type $\forall\, x\, \exists\, y\, R(x, y)$ where R is a decidable predicate. Indeed, *CY1* and *CY2* are both Π_2-complete. Thus, deciding whether all configurations on a CA evolve to a fixed point is equivalent to the classical problem of determining whether a semi-decidable set is infinite. The third class is even less amenable to algorithmic attack; one can show that *CY3* is Σ_3-complete, see [53]. Thus, deciding whether all orbits are decidable is as difficult as determining whether any given semi-decidable set is decidable. It is not difficult to adjust these undecidability results to similar classes such as the lower levels of the Li–Packard hierarchy that takes into account spatial displacements of patterns.

Effective Dynamical Systems and Universality

The key property of CA that is responsible for all these undecidability results is the fact that CA are capable of performing arbitrary computations. This is unsurprising when one defines computability in terms of Turing machines, the devices introduced by Turing in the 1930's, see [47,63]. Unlike the Gödel–Herbrand approach using general recursive functions or Church's λ-calculus, Turing's devices are naturally closely related to discrete dynamical systems. For example, we can express an instantaneous description of a Turing machine as a finite sequence

$$a_{-l}\, a_{-l+1} \dots a_{-1}\; p\; a_1 a_2 \dots a_r$$

where the a_i are tape symbols and p is a state of the machine, with the understanding that the head is positioned at a_1 and that all unspecified tape cells contain the blank symbol. Needless to say, these Turing machine configurations can also be construed as finite support configurations of a one-dimensional CA. It follows that a one-dimensional CA can be used to simulate an arbitrary Turing machine, hence CA are computational universal: any computable function whatsoever can already be computed by a CA.

Note, though, that the simulation is not entirely trivial. First, we have to rely on input/output conventions. For example, we may insist that objects in the input domain, typically tuples of natural numbers, are translated into a configuration of the CA by a primitive recursive coding function. Second, we need to adopt some convention that determines when the desired output has occurred: we follow the evolution of the input configuration until some "halting" condition applies. Again, this condition must be primitive recursively decidable though there is considerable leeway as to how the end of a computation should be signaled by the CA. For example, we could insist that a particular cell reaches a special state, that an arbitrary cell reaches a special state, that the configuration be a fixed point and so forth. Lastly, if and when a halting configuration is reached, we a apply a primitive recursive decoding function to obtain the desired output.

Restricting the space to configurations that have finite support, that are spatially periodic, and so forth, produces an *effective dynamical system*: the configurations can be coded as integers in some natural way, and the next configuration relation is primitive recursive in the sense that the corresponding relation on code numbers is so primitive recursive. A classical example for an effective dynamical system is given by selecting the instantaneous descriptions of a Turing machine M as configurations, and one-step relation of the Turing machine as the operation of C. Thus we obtain a system $\mathcal{A}_M$ whose orbits represent the computations of the Turing machine. Likewise, given the local map ρ of a CA we obtain a system $\mathcal{A}_\rho$ whose operation is the induced global map. While the full configuration space C_{all} violates the effectiveness condition, any of the spaces C_{per}, C_{fin}, C_{ap} and C_{rec} will give rise to an effective dynamical system. Closure properties as well as recent work on the universality of elementary CA 110, see Sect. "Conclusion", suggests that the class of almost periodic configurations, also known as backgrounds or wallpapers, see [9,58], is perhaps the most natural setting. Both C_{fin} and C_{ap} provide a suitable setting for a CA that simulates a Turing machine: we can interpret $\mathcal{A}_M$ as a subspace of $\mathcal{A}_\rho$ for some suitably constructed one-dimensional CA ρ; the orbits of the subspace encode computations of the Turing machine. It follows from the undecidability of the Halting Problem for Turing machines that the Reachability Problem for these particular CA is undecidable.

Note, though, that orbits in $\mathcal{A}_M$ may well be finite, so some care must be taken in setting up the simulation. For example, one can translate halting configurations into fixed points. Another problem is caused by the worst-case nature of our classification schemes: in Turing machines and their associated systems $\mathcal{A}_M$ it is only behavior on specially prepared initial configurations that matters, whereas the behavior of a CA depends on all configurations. The behavior of a Turing machine on all instantaneous descriptions, rather than just the ones that can occur during a legitimate computation on some actual input, was first studied by Davis, see [12,13], and also Hooper [25].

Call a Turing machine *stable* if it halts on any instantaneous description whatsoever. With some extra care one can then construct a CA that lies in the first Culik–Yu class, yet has the same computational power as the Turing machine. Davis showed that every total recursive function can already be computed by a stable Turing machine, so membership in *CY1* is not an impediment to considerable computational power. The argument rests on a particular decomposition of recursive functions. Alternatively, one directly manipulate Turing machines to obtain a similar result, see [49,53]. On the other hand, unstable Turing machines yield a natural and coding-free definition of universality: a Turing machine is *Davis-universal* if the set of all instantaneous description on which the machine halts is Σ_1-complete.

The mathematical theory of infinite CA is arguably more elegant than the actually observable finite case. As a consequence, classifications are typically concerned with CA operating on infinite grids, so that even a configuration with finite support can carry arbitrarily much information. If we restrict our attention to the space of configurations on a finite grid a more fine-grained analysis is required. For a finite grid of size n the configuration space has the form $C_n = [n] \to \Sigma$ and is itself finite, hence any orbit is ultimately periodic and the Reachability Problem is trivially decidable. However, in practice there is little difference between the finite and infinite case. First, computational complexity issues make it practically impossible to analyze even systems of modest size. The Reachability Problem for finite CA, while decidable, is PSPACE-complete even in the one-dimensional case. Computational hardness appears in many other places. For example, if we try to determine whether a given configuration on a finite grid is a Garden-of-Eden the problem turns out to be NLOG-complete in dimension one and $\mathbb{NP}$-complete in all higher dimensions, see [56].

Second, it stands to reason that the more interesting classification problem in the finite case takes the following parameterized form: given a local map together with boundary conditions, determine the behavior of ρ on all finite grids. Under periodic boundary conditions this comes down to the study of C_{per} and it seems that there is little difference between this and the fixed boundary case. Since all orbits on a finite grid are ultimately periodic one needs to apply a more fine-grained classification that takes into account transient lengths. It is undecidable whether all configurations on all finite grids evolve to a fixed point under a given local map, see [54]. Thus, there is no algorithm to determine whether

$$\langle C_n, \to \rangle \models \forall\, x\, \exists\, z\, (x \overset{*}{\to} z \text{ and } z \to z)$$

for all grid sizes n. The transient lengths are trivially bounded by k^n where k is the size of the alphabet of the automaton. It is undecidable whether the transient lengths grow according to some polynomial bound, even when the polynomial in question is constant.

Restrictions of the configuration space are one way to obtain an effective dynamical system. Another is to interpret the approximation-based notion of computability on the full space in terms of topology. It is well-known that computable maps $C_{\text{all}} \to C_{\text{all}}$ are continuous in the standard product topology. The clopen sets in this topology are the finite unions of cylinder sets where a cylinder set is determined by the values of a configuration in finitely many places. By a celebrated result of Hedlund the global maps of a CA on the full space are characterized by being continuous and shift-invariant. Perhaps somewhat counter-intuitively, the decidable subsets of C_{all} are quite weak, they consist precisely of the clopen sets. Now consider a partition of C_{all} into finitely many clopen sets $C_0, C_2, \dots, C_{n-1}$. Thus, it is decidable which block of the partition a given point in the space belongs to. Moreover, Boolean operations on clopen sets as well as application of the global map and the inverse global map are all computable. The partition affords a natural projection $\pi : C_{\text{all}} \to \Sigma_n$ where $\Sigma_n = \{0, 1, \dots, n-1\}$ and $\pi(x) = i$ iff $x \in C_i$. Hence the projection translates orbits in the full space C_{all} into a class W of ω-words over Σ_n, the symbolic orbits of the system. The Cantor space $\Sigma_n^{\mathbb{Z}}$ together with the shift describes all logically possible orbits with respect to the given partition and W describes the symbolic orbits that actually occur in the given CA. The shift operator corresponds to an application of the global map of the CA. The finite factors of W provide information about possible finite traces of an orbit when filtered through the given partition. Whole orbits, again filtered through the partition, can be described by ω-words. To tackle the classification of the CA in terms of W it was suggested by Delvenne et al., see [15], to refer to the CA as *decidable* if there it is decidable whether W has non-empty intersection with a ω-regular language. Alas, decidability in this sense is very difficult, its complexity being Σ_1^1-complete and thus outside of the arithmetical hierarchy. Likewise it is suggested to call a CA *universal* if the problem of deciding whether the cover of W, the collection of all finite factors, is Σ_1-complete, in analogy to Davis-universality.

Computational Equivalence

In recent work, Wolfram suggests a so-called *Principle of Computational Equivalence*, or PCE for short, see [71],

p. 717. PCE states that most computational processes come in only two flavors: they are either of a very simple kind and avoid undecidability, or they represent a universal computation and are therefore no less complicated than the Halting Problem. Thus, Wolfram proposes a zero-one law: almost all computational systems, and thus in particular all CA, are either as complicated as a universal Turing machine or are computationally simple. As evidence for PCE Wolfram adduces a very large collection of simulations of various effective dynamical systems such as Turing machines, register machines, tag systems, rewrite systems, combinators, and cellular automata. It is pointed out in Chap. 3 of [71], that in all these classes of systems there are surprisingly small examples that exhibit exceedingly complicated behavior – and presumably are capable of universal computation. Thus it is conceivable that universality is a rather common property, a property that is indeed shared by all systems that are not obviously simple. Of course, it is often very difficult to give a complete proof of the computational universality of a natural system, as opposed to carefully constructed one, so it is not entirely clear how many of Wolfram's examples are in fact universal. As a case in point consider the universality proof of Conway's Game of Life, or the argument for elementary CA 110. If Wolfram's PCE can be formally established in some form it stands to reason that it will apply to all effective dynamical systems and in particular to CA. Hence, classifications of CA would be rather straightforward: at the top there would be the class of universal CA, directly preceded by a class similar to the third Culik–Yu class, plus a variety of subclasses along the lines of the lower Li–Packard classes.

The corresponding problem in classical computability theory was first considered in the 1930's by Post and is now known as Post's Problem: is there a semi-decidable set that fails to be decidable, yet is not as complicated as the Halting Set? In terms of Turing degrees the problem thus is to construct a semi-decidable set A such that $\emptyset <_{\mathrm{T}} A <_{\mathrm{T}} \emptyset'$, or to rule out the existence of any such set, see [31,47,52] for background on Turing degrees in general and semi-decidable degrees in particular. Post's Problem resisted all attempts at resolution until Friedberg and Muchnik independently and almost simultaneously discovered a way to construct a set of intermediate complexity, see [18,41]. The construction is based on the idea of a so-called priority argument and is significantly more complicated than any construction of semi-decidable sets previously known [52]. Indeed, priority arguments have since become the hallmark of computability theory and have even engendered some criticism as being so very technical that, occasionally, the proofs seem to attract more attention than the theorems being established, see [65]. Be that as it may, it is striking how much more artificial and ad hoc intermediate sets are, as compared to natural examples such as the theory of the reals (decidable) or of Diophantine equations (equivalent to the Halting Problem). No natural examples of intermediate semi-decidable sets are known to date.

Nonetheless, given an intermediate set A one can construct a one-dimensional CA whose Reachability Problem has the same degree as A. This suggests a degree-based classification: given any computably enumerable degree $\mathbf{d}$, define the class $\mathbb{C}_{\mathbf{d}}$ to consist of all CA whose Reachability Problem has degree exactly $\mathbf{d}$, see [57,59]. The degree classification is non-trivial in the sense that every class is non-empty. Note that the first three Culik–Yu classes are all contained in $\mathbb{C}_{\mathbf{0}}$ whereas $\mathbb{C}_{\mathbf{0}'}$ comprises all computationally universal CA. Unsurprisingly, it is again undecidable whether a CA belongs to any particular class. At the bottom end of the hierarchy it is Σ_3-complete to determine membership in $\mathbb{C}_{\mathbf{0}}$; at the top end it is Σ_4-complete to determine membership in $\mathbb{C}_{\mathbf{0}'}$. Thus, it is easier to determine decidability than universality. In general, deciding membership in $\mathbb{C}_{\mathbf{d}}$ is Σ_3^d-complete for any semi-decidable degree $\mathbf{d}$. Similar results hold for the analogous cumulative classes $\mathbb{C}_{\leq \mathbf{d}} = \bigcup_{\mathbf{e} \leq \mathbf{d}} \mathbb{C}_{\mathbf{e}}$.

Unlike the Culik–Yu classification, the structure of the degree classification between $\mathbb{C}_{\mathbf{0}}$ and $\mathbb{C}_{\mathbf{0}'}$ is exceedingly complicated. For example, the proof of the Friedberg–Muchnik theorem shows that there are incomparable semi-decidable degrees $\mathbf{d}_1$ and $\mathbf{d}_2$. Hence there is are CA whose orbits are undecidable but not as complicated as the Halting Problem. Indeed, complete knowledge of the orbits of one of the two CA will not help in deciding membership in the orbits of the other. Another surprising result in the theory of computably enumerable degrees is Sack's Density Theorem, see [52]: between any two computably enumerable degrees $\mathbf{d}_1 < \mathbf{d}_2$ there lies a third: $\mathbf{d}_1 < \mathbf{d} < \mathbf{d}_2$. Thus, between any two CA of strictly increasing complexity there is an infinite and dense hierarchy of other CA. The computably enumerable degrees form a semi-lattice, so it is natural to try to understand the complexity of the structure by analyzing its first order theory. It is well-known that the Σ_1-theory of this semi-lattice is decidable. However, the reason for this decidability result lies in the fact that any countable partial order can be embedded into the semi-lattice so that the relative computational strength of cellular automata is indeed arbitrarily complicated. On the other hand, the full theory of the semi-lattice of semi-decidable degrees is known to be highly undecidable, see [22]; its degree is $\emptyset^{(\omega)}$. One might hope that restriction to reversible CA would simplify the situation somewhat. Somewhat surprisingly it turns out that each class $\mathbb{C}_{\mathbf{d}}$ already contains an irreversible CA, see [60], so the same difficulties arise in the classification of reversible CA as in the classification of ordinary CA.

While reachability is arguably the most basic relation between configurations, similar difficulties also arise with confluence. As a matter of fact, one can construct a CA whose Reachability Problem has complexity some arbitrarily chosen computably enumerable degree $\mathbf{d}_1$ while the Confluence Problem for the same CA has degree $\mathbf{d}_2$, another arbitrarily chosen computably enumerable degree. Thus, a classification according to reachability is entirely independent of a confluence-based classification.

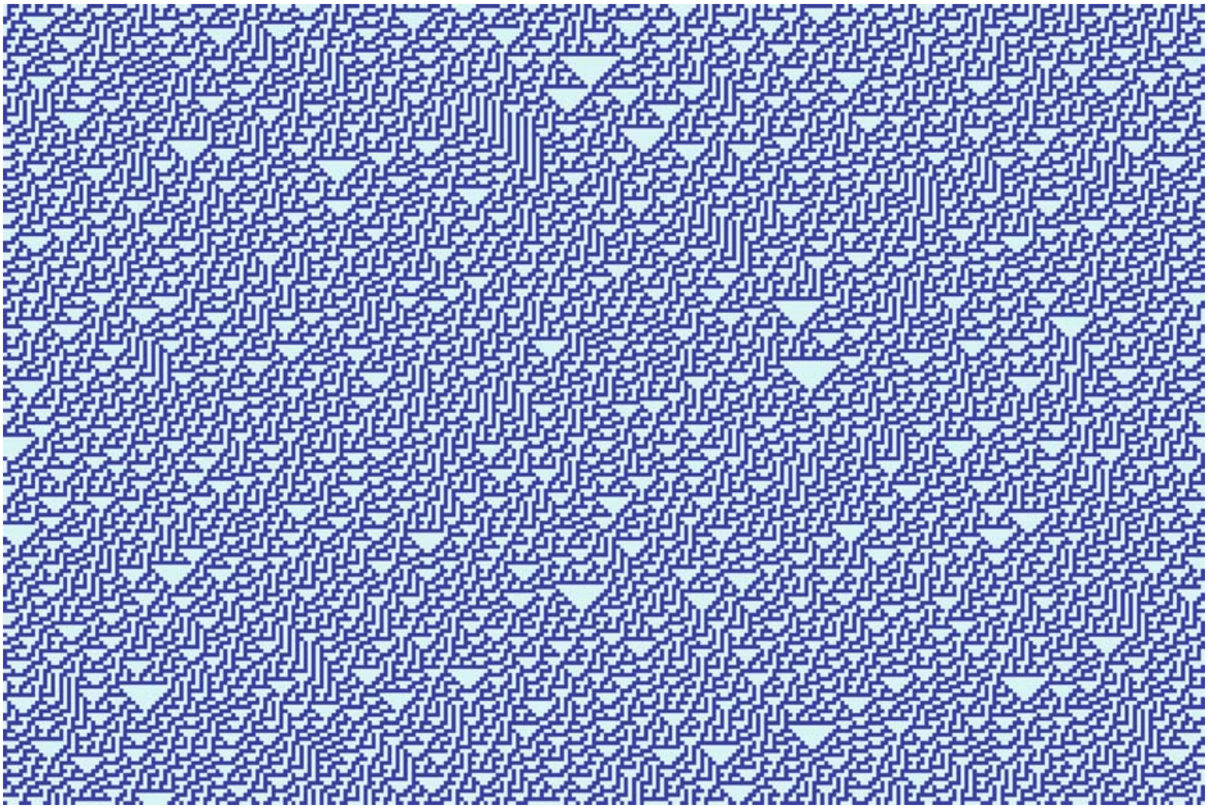
How do these results relate to PCE? Wolfram would not accept any of the intermediate classes of CA as a counterexample to PCE. The argument is that though intermediate degrees exist, their construction is critically linked to universal computation. While the universal computation is invisible when only the output of the system is observed, the associated computational process includes the whole computation and is thus universal. As a case in point, consider the standard Friedberg–Muchnik construction for an intermediate semi-decidable set A. The construction actually builds two semi-decidable sets A and B that are mutually incomparable with respect to Turing reducibility. Only A is output and B remains hidden. However, even ignoring all the intricate technical details of the whole construction, if we consider both A and B as output then the computation is indeed universal: the disjoint union $A \oplus B$ is Σ_1-complete, see [51]. It remains to be seen if similar arguments can be put forth in connection with priority-free constructions of intermediate degrees or if natural examples of intermediate sets can be found. At any rate, by considering only the reachability relation instead of a whole segment of the orbit we also achieve information-hiding, much as in the classical Friedberg–Muchnik construction.

Conclusion

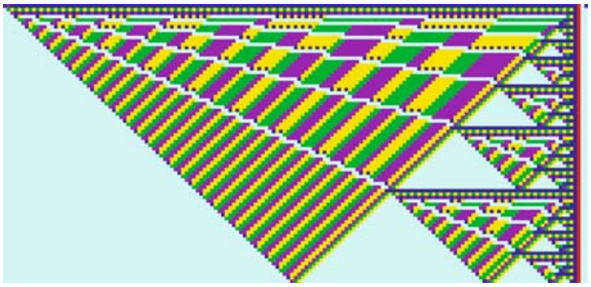
Classification schemes of cellular automata based on the long-term evolution of pattern are typically undecidable, even if the property in question can be expressed in a fairly week system. While it is easy to construct examples of CA in particular classes it is usually very difficult to establish the position of a given CA in a particular classification. An excellent example for the difficulty of analyzing a given CA is Cook's proof of the universality of elementary CA number 110 whose local rule is given by $\rho(x, y, z) = (\overline{x} \wedge y \wedge z) \oplus y \oplus z$ where $\oplus$ denotes exclusive or, see [9], ► Cellular Automata, Universality of. The argument shows that cyclic tag systems, which are known to be complete, can be simulated by elementary CA 110 provided one allows an almost periodic background. Recent work by Turlough and Woods has shown that the whole simulation can be effected with only a polynomial slow-down, see [42,43]. This result suggests that the appropriate setting for classifications is the space of almost periodic configurations rather than finite ones.

In light of the successful analysis of elementary CA 110 it is tempting to ask about the classification of elementary CA 30. Figure 3 shows a segment of the orbit of a one-point seed configuration under rule 30. It is striking how chaotic and apparently random the image is. As a matter of fact, rule 30 has been used for many years as the default random number generator in the commercial computer algebra system Mathematica, see [70]. The underlying local map is simply $\rho(x, y, z) = x \oplus (y \vee z)$. Alas, there appear to be no structures in the evolution of configurations under rule 30 such as "moving particles" that might be exploited in a universality argument along the lines of rule 110. On the other hand, it is unclear how a decision procedure for reachability could be developed. This makes it tempting to conjecture that rule 30 in C_{ap} might be a member of one of the intermediate classes $\mathbb{C}_{\mathbf{d}}$, though at present there seems to be no way to either establish or refute this conjecture.

While undecidability results rule out the possibility of automatic classification mechanisms there is still ample room for the development of sufficient criteria for membership in certain classes, see [1,64,72]. For example, a proof of computational universality in a CA that has not been artificially constructed to simulate some other device often rests on the presence of "particles" or "gliders" that can be used to send "signals" between spatially separated locations. Moreover, one has to be able to process these signals much in the way of Boolean logic gates, to store state and so forth. A good example for complicate interactions between signals are the various solutions to the firing squad problem, albeit not in the context of simulating arbitrary computations; see Fig. 4, [36]. A more recent example is Cook's ingenious method of using natural gliders in elementary CA 110 to implement a cyclic tag system in C_{ap}, thereby establishing computational universality of rule 110, see [9]. Notable here is the fact that the automaton was fixed from the start and the the appropriate coding mechanisms had to be developed in a very constrained environment. This is in stark contrast to other hardness arguments where the CA is carefully constructed to display the desired behavior. Careful visual inspection of rule 110 orbits was a crucial component in Cook's proof, it is difficult to imagine that the result could have been established in a purely combinatorial or algebraic fashion. One



Cellular Automata, Classification of, Figure 3
A pseudo-random pattern generated by elementary CA 30



Cellular Automata, Classification of, Figure 4
Interacting signals in Mazoyer's optimal solution to the firing squad problem

can envision an interactive software system that helps to tackle some algorithmically unsolvable classification problems in special cases, much as Baumslag's Magnus project in group theory, see [5].

Bibliography

1. Adamatzky A (1994) Identification of Cellular Automata. Taylor & Francis, London
2. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tesselation structures. J Comput Syst Sci 6:448–464
3. Baldwin JT (2002) Computation versus simulation. http://www.math.uic.edu/~jbaldwin/pub/cafom.ps. Accessed May 2007
4. Baldwin JT, Shelah S (2000) On the classifiability of cellular automata. Theor Comput Sci 230(1–2):117–129
5. Baumslag G () Magnus. http://caissny.org/. Accessed May 2007
6. Beal M-P, Perrin D (1997) Symbolic dynamics and finite automata. In: Rozenberg G, Salomaa A (eds) Handbook of Formal Languages. Springer, Berlin
7. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17:525–532
8. Börger E, Grädel E, Gurevich Y (2001) The Classical Decision Problem. Springer, Berlin
9. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15(1):1–40
10. Culik K (1987) On invertible cellular automata. Complex Syst 1(6):1035–1044
11. Culik K, Sheng Y (1988) Undecidability of CA classification schemes. Complex Syst 2(2):177–190
12. Davis M (1956) A note on universal Turing machines. Shannon CE, McCarthy J (eds) Automata Studies. Annals of Mathematics Studies, vol 34. Princeton University Press, Princeton, pp 167–175
13. Davis M (1957) The definition of universal Turing machines. Proc Am Math Soc 8:1125–1126
14. Delorme M, Mazoyer J (1999) Cellular Automata: A Parallel

Model. Mathematics and Its Applications, vol 460. Kluwer, Dordrecht

15. Delvenne J-C, Kůrka P, Blondel V (2006) Decidability and universality in symbolic dynamical systems. Fundamenta Informaticae 74(4):463–490
16. Durand-Lose J (2001) Representing reversible cellular automata with reversible block automata. Disc Math Theor Comp Sci Proc AA:145–154
17. Elgaard J, Klarlund N, Møller A (1998) MONA 1.x: new techniques for WS1S and WS2S. In: Proc. 10th International Conference on Computer-Aided Verification, CAV '98. LNCS, vol 1427. Springer, Berlin, pp 516–520
18. Friedberg RM (1957) Two recursively enumerable sets of incomparable degrees of unsolvability. Proc Natl Acad Sci USA 43:236–238
19. Grädel E, Thomas W, Wilke T (eds) (2002) Automata, Logics, and Infinite Games. LNCS, vol 2500. Springer, Berlin
20. Gutowitz H (1996) Cellular automata and the sciences of complexity, part I. Complexity 1(5):16–22
21. Gutowitz H (1996) Cellular automata and the sciences of complexity, part II. Complexity 1(6)
22. Harrington L, Shelah S (1982) The undecidability of the recursively enumerable degrees. Bull Amer Math Soc 6:79–80
23. Head T (1989) Linear CA: injectivity from ambiguity. Complex Syst 3(4):343–348
24. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
25. Hooper PK (1966) The undecidability of the Turing machine immortality problem. J Symb Log 31(2):219–234
26. Kari J (1990) Reversibility of 2D cellular automata is undecidable. Physica D 45:397–385
27. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergod Th Dyn Syst 17:417–433
28. Kůrka P (2003) Topological and Symbolic Dynamics. Cours Spécialisés, vol 11. Societe Mathematique de France, Paris
29. Langton CG (1990) Computation at the edge of chaos. Physica D 42:12–37
30. Lecerf Y (1963) Machine de Turing réversible. Insolubilité récursive en $n \in N$ de l'équation $u = \theta^n u$, où θ est un "isomorphisme de codes". C R Acad Sci Paris 257:2597–2600
31. Lerman M (1983) Degrees of Unsolvability. Perspectives in Mathematical Logic. Springer, Berlin
32. Li W, Packard N (1990) The structure of the elementary cellular automata rule space. Complex Syst 4(3):281–297
33. Li W, Packard N, Langton CG (1990) Transition phenomena in CA rule space. Physica D 45(1–3):77–94
34. Libkin L (2004) Elements of Finite Model Theory. Springer, Berlin
35. Manzini G, Margara L (1999) A complete and efficiently computable topological classification of *D*-dimensional linear cellular automata over Z_m. Theor Comput Sci 221(1–2):157–177
36. Mazoyer J (1987) A six state minimal time solution to the firing squad synchronization problem. Theor Comput Sci 50:183–238
37. Mitchell M, Crutchfield JP, Hraber PT (1994) Evolving cellular automata to perform computations: Mechanisms and impediments. Physica D (75):361–369
38. Morita K (1994) Reversible cellular automata. J Inf Process Soc Japan 35:315–321
39. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comput Sci 148:157–163
40. Morita K, Harao M (1989) Computation universality of 1 dimensional reversible (injective) cellular automata. Trans Inst Electron, Inf Commun Eng E 72:758–762
41. Muchnik AA (1956) On the unsolvability of the problem of reducibility in the theory of algorithms. Dokl Acad Nauk SSSR 108:194–197
42. Neary R, Woods D (2006) On the time complexity of 2-tag systems and small universal turing machines. In: FOCS, pp 439–448. IEEE Comput Soc, Berkeley
43. Neary R, Woods D (2006) Small fast universal turing machines. Theor Comput Sci 362(1–3):171–195
44. Oliveira G, Oliveira P, Omar N (2001) Definition and application of a five-parameter characterization of one-dimensional cellular automata rule space. Artif Life 7(3):277–301
45. Ollinger N (2003) The intrinisic universality problem of one-dimensional cellular automata. In: Alt H, Habib M (eds) Proc. STACS. LNCS, vol 2607. Springer, Berlin, pp 632–641,
46. Packard NH (1988) Adaptation towards the Edge of Chaos. In: Dynamic Patterns in Complex Systems. World Scientific, Singapore, pp 29–301
47. Rogers H (1967) Theory of Recursive Functions and Effective Computability. McGraw Hill, New York
48. Rozenberg G, Salomaa A (1997) Handbook of Formal Languages. Springer, Berlin
49. Shepherdson JC (1965) Machine configuration and word problems of given degree of unsolvability. Z Math Logik Grundl Math 11:149–175
50. Shoenfield JR (1967) Mathematical Logic. Addison Wesley
51. Soare RI (1972) The Friedberg-Muchnik theorem re-examined. Can J Math 24:1070–1078
52. Soare RI (1987) Recursively Enumerable Sets and Degrees. Perspectives in Mathematical Logic. Springer, Berlin
53. Sutner K (1989) A note on Culik-Yu classes. Complex Syst 3(1):107–115
54. Sutner K (1990) Classifying circular cellular automata. Physica D 45(1–3):386–395
55. Sutner K (1991) De Bruijn graphs and linear cellular automata. Complex Syst 5(1):19–30
56. Sutner K (1995) The complexity of finite cellular automata. J Comput Syst Sci 50(1):87–97,
57. Sutner K (2002) Cellular automata and intermediate reachability problems. Fundamentae Informaticae 52(1–3):249–256
58. Sutner K (2003) Almost periodic configurations on linear cellular automata. Fundamentae Informaticae 58(3,4):223–240
59. Sutner K (2003) Cellular automata and intermediate degrees. Theor Comput Sci 296:365–375
60. Sutner K (2004) The complexity of reversible cellular automata. Theor Comput Sci 325(2):317–328
61. Taati S (2007) Cellular automata reversible over limit set. J Cell Autom 2(2):167–177
62. Toffoli T, Margolus N (1990) Invertible cellular automata: A review. Physica D 45:229–253
63. Turing AM (1936) On computable numbers, with an application to the Entscheidungsproblem. P Lond Math Soc 42: 230–65
64. Vorhees B (1996) Computational Analysis of One-Dimensional Cellular Automata. World Scientific, Singapore
65. Wang H (1993) Popular Lectures on Mathematical Logic. Dover Publications, Dover, New York
66. Weihrauch K (2000) Computable Analysis. EATCS Monographs. Springer, Berlin

67. Wolfram S (1984) Computation theory of cellular automata. Comm Math Phys 96(1):15–57
68. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10:1–35
69. Wolfram S (1985) Twenty problems in the theory of cellular automata. Phys Scr T9:170–183
70. Wolfram S (2002) The Mathematica Book. Cambridge University Press, Cambridge
71. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign
72. Wuensche A (1999) Classifying cellular automata automatically. Complexity 4(3):47–66

Cellular Automata, Emergent Phenomena in

JAMES E. HANSON
IBM T.J. Watson Research Center,
Yorktown Heights, USA

Article Outline

Glossary

Cellular automaton A spatially-extended dynamical system in which spatially-discrete cells take on discrete values, and evolve according to a spatially-localized discrete-time update rule.

Emergent phenomenon A phenomenon that arises as a result of a dynamical system's intrinsic dynamical behavior.

Domain A spatio-temporal region of a cellular automation that conforms to a specific pattern.

Particle A spatially-localized region of a cellular automaton that exists as a boundary or defect in a domain, and persists for a significant amount of time.

Definition of the Subject

In a dynamical system, an "emergent" phenomenon is one that arises out of the system's own dynamical behavior, as opposed to being introduced from outside. Emergent phenomena are ubiquitous in the natural world; as just one example, consider a shallow body of water with a sandy bottom. It often happens that small ridges form in the sand. These ridges emerge spontaneously, have a characteristic size and shape, and move across the bottom in a characteristic way – all due to the interaction of the sand and the water.

In cellular automata (CA), the system's state consists of an N-dimensional array of discrete cells that take on discrete values and the dynamics is given by a discrete time update rule (see below). The "phenomena" that emerge in CA therefore necessarily consist of spatio-temporal patterns and/or statistical regularities in the cell values. Therefore, the study of emergent phenomena is CA is the study of the spatio-temporal patterns and statistical regularities that arise spontaneously in cellular automata.

Introduction

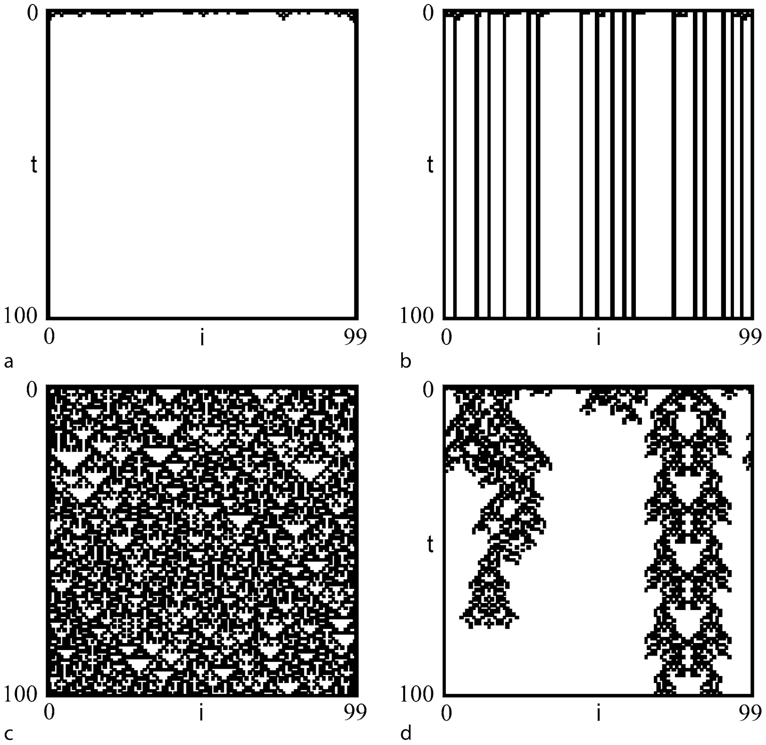
The study of emergent phenomena in cellular automata dates back at least to the beginnings of the modern era of CA investigation inaugurated by Stephen Wolfram and collaborators. Indeed, it was a central theme of the landmark paper that introduced the four "Wolfram classes" [15] shown in Fig. 1. Ever since, emergent phenomena have been the driving force behind a great deal of CA research.

To be genuinely emergent, a phenomenon must arise out of configurations in which it is not present; and furthermore, to be of any significance, it must do so with non-vanishing likehood, and persist for a measurable amount of time. Thus the proper study of emergent phenomena in CA excludes from consideration a broad subcategory of systems in which the initial condition and update rule are chosen a priori to exhibit some particular structural feature (lattice gases are a representative example). The fact that such systems are CA is an implementation detail; the CA is merely a substrate or means for the simulation of higher-order structures. Note also that the essential issue is not whether the phenomena were intentionally designed into the CA rule; it is whether they arise naturally with any degree of frequency from configurations in which they are not present.

Notation and Terminology

A **cellular automaton** (CA) consists of a discrete N-dimensional array of sites or **cells** and a discrete-time **local update rule** ϕ applied to all cells in parallel.

The location of a cell is given by the N integer-valued coordinates $\{i, j, k, \ldots\}$. Cells take on values in a discrete set or **alphabet**, conventionally written $0, 1, \ldots, k-1$,



Cellular Automata, Emergent Phenomena in, Figure 1
Examples of Wolframs four qualitative classes. a Class 1: Spatio-temporally uniform configuration of ECA 32. b Class 2: Separated simple or periodic structures of ECA 44. c Class 3: Chaotic space-time pattern of ECA 90. d Class 4: Complex localized structures of Binary radius-2 CA 1771476584. In all cases the initial condition is random. In this and subsequent figures, cells with value 0 are shown as *white squares*, cells with value 1 are *black*

with k the **alphabet size**. An assignment of values to cells is called the **configuration** of those cells. The value 0 is sometimes treated as a special "quiescent" value, particularly in rules that obey the **quiescence condition** $\phi(\ldots 0 \ldots) = 0$.

The local update rule determines the value of a cell at time $t+1$ as a function of the values at time t of the cells around it. Typical neighborhoods are symmetrical, centered on the cell to be updated, and are parametrized by the **radius** r, which is the greatest distance from the center cell to any cell in the neighborhood. An assignment of values to the cells in a neighborhood is called a **parent neighborhood**, denoted by η, and the value $\phi(\eta)$ to which that parent neighborhood is mapped under the local update rule is its **child value**. The set of ordered pairs $\{\eta, \phi(\eta)\}$ is the **rule table**. The **speed of light** of a CA is the maximal rate at which information about a cell's value may travel; in general it is given by the radius r.

In two dimensions there are two common alternatives for the neighborhood's shape: the **von Neumann** neighborhood includes the center cell and its four neighbors up, down, left, and right; and the **Moore** neighborhood, which also includes the four cells diagonally adjacent to the center cell.

The so-called **elementary** cellular automata (ECA) are one-dimensional CA with $k = 2$, $r = 1$; a cell is denoted σ^i, takes on values in $\{0, 1\}$ and evolves over time according to the rule $\sigma^i_{t+1} = \phi(\sigma^{i-1}_t, \sigma^i_t, \sigma^{i+1}_t)$. A neighborhood consists of three consecutive cells, so there are 8 distinct parent neighborhoods and 256 different rule tables. It is convenient to refer to an elementary CA by its **rule number**, which is determined as follows. The different parent neighborhoods η are regarded as numbers in base k and are arranged in decreasing numerical order, from left to right. Immediately beneath each parent neighborhood its child value $\phi(\eta)$ is written. The rule number is obtained by regarding the sequence of child symbols as another number, again in base k. This numbering scheme may be used for one-dimensional CA with any k and r, and may be extended to higher-dimensional CA by the adoption of a convention for assigning numerical values to parent neighborhoods.

Different formulations of the local update rule are possible for CA in which symmetry or other constraints are present. For example, one important subclass of CA rules are the **totalistic** rules, in which the child value depends only on the sum of the values in the parent neighborhood, not on their positions. Totalistic rules may also be assigned a rule number, by writing down the different possible sums of cell values in the parent neighborhood in order, writing the child cell beneath each such sum, and interpreting the sequence of child cells as a number.

In describing patterns in one-dimensional configurations, it is convenient to adopt a simplified form of regular expression notation, as follows:

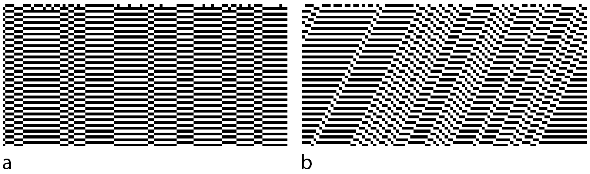
- symbols $0, 1, \ldots, k-1$ denote literal cell values
- the symbol Σ denotes a "wild card" that may take on any value in the alphabet
- the expression x^* denotes any number of repetitions of the pattern x
- $[\ldots]$ denotes grouping
- concatenation denotes spatial adjacency.

For example, 0^* represents any number of consecutive 0s, while $[10]^*1$ is any configuration consisting of some number of repetitions of the pattern 10 followed by a 1: e. g., 101, 10101, 1010101, and so forth.

Synchronization

Possibly the simplest type of emergent phenomenon in CA is **synchronization**, which is the growth of spatial regions in which all cells have the same value. A synchronized region remains synchronized over time (except possibly at its borders) and it may either temporally invariant (i. e., the cell values to not change in time) or periodic (the cells all cycle together through the same temporal sequence of values). The temporal periodicity in the latter case is not greater than the alphabet size k.

About the trivial case in which the CA rule maps all neighborhoods to the same value (e. g., ECA 0 or ECA 255), there is little to be said. However, other cases exist in which the synchronized regions emerge only gradually. Characteristic examples in one dimension are shown in Fig. 1a and c. It is evident from these examples that any initial condition can be roughly, but usefully, described in terms of four patterns: (a) pattern 0^*, which represents the synchronized regions; (b,c) boundary regions $0^*\Sigma^*$ and Σ^*0^*; and (d) Σ^* for the interior of the non-synchronized regions. The behavior of the boundary regions determines whether the synchronized regions grow or shrink. For example, in ECA 32 (Fig. 1a), the parent neighborhoods in the boundary region are $\eta = \{0\Sigma\Sigma, \Sigma\Sigma 0\}$, all of which have child value 0; this means that the synchronized region grows as fast as is possible. Also not that since the only parent neighborhood that is *not* mapped to 0 is $\eta = 101$. the time taken for a given configuration in ECA 32 to reach a globally synchronized state is governed by the length of the longest region of pattern $[10]^*1$.



Cellular Automata, Emergent Phenomena in, Figure 2
Synchronization and phase defects: a ECA 55. b ECA 17

In general, the growth (or shrinkage) of synchronized regions is determined by the aggregate behavior of the neighborhoods that occur its boundaries; if they recede from each other, the region will grow. The boundaries need not move at the speed of light; the left and right boundaries need not move at the same speed; and their motion need not be perfectly uniform over time.

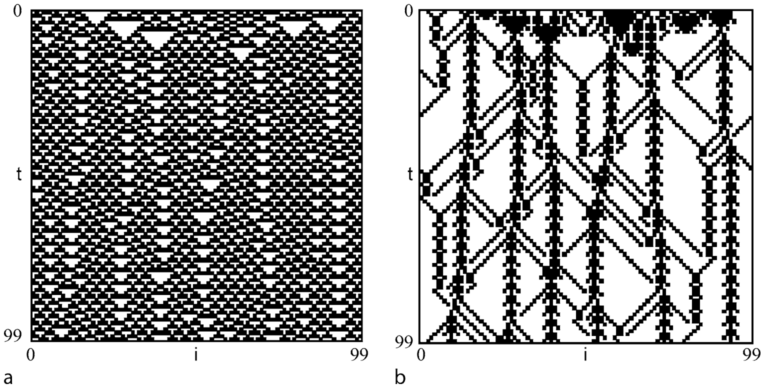
Figure 2a shows ECA 55, in which synchronized regions with temporal period $p = 2$ emerge from random initial conditions. Note, however, that multiple distinct synchronized regions persist indefinitely. This is an example of a **temporal phase defect**, which is a boundary between spatio-temporal regions that have the same overall pattern, but one of which is *ahead* of the other in time.

In general, phase defects need not be stationary: an example is shown in Fig. 2b. Also note that for CA with $k > 2$ it is possible for several different synchronized patterns to emerge and coexist. For example, consider a CA with $k = 3$ in which the pattern 0^* is temporally invariant, while 1^* and 2^* are mapped into each other to form a period-2 cycle.

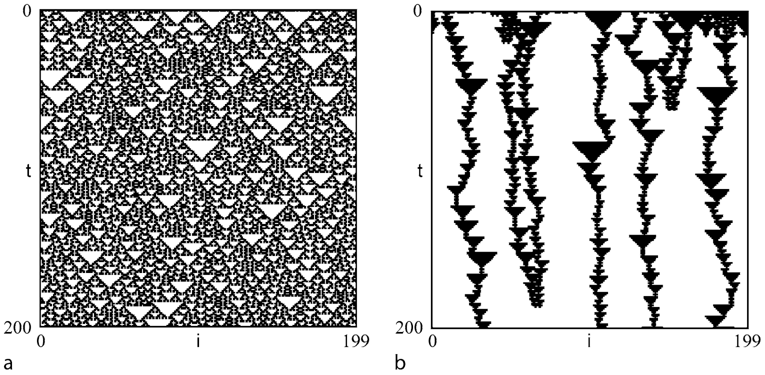
Domains in One Dimension

Synchronization is a special case of a more general emergent phenomenon, the **domain**. A domain is spatial region that conforms to some specific pattern which persists over time. As has been seen in the case of synchronization, the emergence of a domain is governed by the behavior of its boundaries.

An important subclass of domain is the **regular domain**, in which the spatial pattern may be expressed in terms of a regular language (or equivalently, a finite state machine) [11]. As defined in [9], a regular domain has two properties: all spatial sequences of cells in the domain are in a given regular language; and (2) the set of all sequences in that regular language is itself temporally invariant or periodic. Regular domains are a powerful tool for identify-



Cellular Automata, Emergent Phenomena in, Figure 3
Raw and domain-filtered space-time diagrams of ECA 54



Cellular Automata, Emergent Phenomena in, Figure 4
Raw and domain-filtered ECA 18

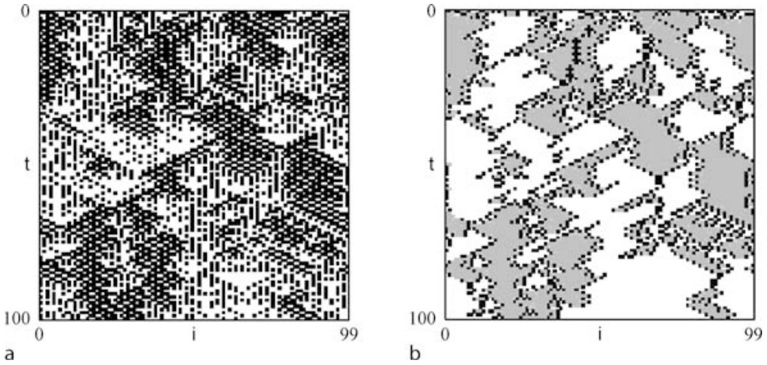
ing and analyzing emergent phenomena in CA of one dimension. Generalization to two or more dimensions has proven challenging, though [12] made a significant step in that direction.

In studying domains in CA, it is useful to pass the space-time data through a **domain filter** to help visualize them. A domain filter, which may be constructed for any regular domain, maps every cell that is in the domain to a chosen value (0, say) and maps all cells not in the domain to other values in a prescribed way. Multi-domain filters may be constructed in a similar fashion, to map cells in any of a set of distinct domains $\Lambda_1, \Lambda_2, \ldots$ onto distinct values $\sigma_1, \sigma_2, \ldots$. See [3] for details.

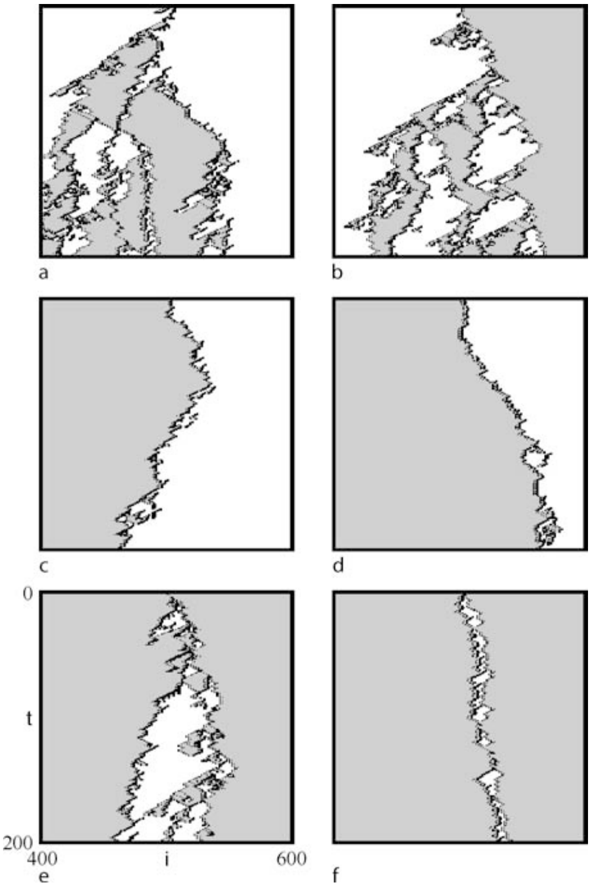
An illustrative example is ECA 54, shown in Fig. 3. On the left is the unfiltered data; and on the right, the same dat after passing through the domain filter for ECA 54's primary domain. The domain has temporal period $p = 2$ and alternates between patterns [0001] and [110] The two patterns line up to form the interlocking white and black "T" shapes visible in the unfiltered data.

As the filtered plot clearly shows, the cells *not* in the domain have patterns of their own; this will be discussed in the next section. For now, it is sufficient to note that, in addition to the *temporal* phase defects seen in the emergence of temporally periodic synchronized regions, domains with nontrivial spatial structure may also show *spatial* phase defects, in which the pattern, in effect, skips or slips by a few cells.

The spatial regions that make up a domain may themselves contain disorder; such domains are called **chaotic**. ECA 90 is the archetypical example of this see Fig. 1c. From a random initial condition, ECA 90 quickly evolves so that entire configuration is in the domain $[0\Sigma]^*$. ECA 18 see Fig. 4a, attempts to do the same, except that the global synchronization is frustrated by long-lived spatial phase defects. This is clearly visible in the filtered space-time diagram shown in Fig. 4b. In this case the boundaries of the domain are inherently ambiguous: the pattern $[0\Sigma]^*[00]^*[\Sigma 0]^*$ contains exactly one spatial phase defect, but it may be regarded as lying anywhere in the cen-



Cellular Automata, Emergent Phenomena in, Figure 5
Multiple coexisting chaotic domains



Cellular Automata, Emergent Phenomena in, Figure 6
Domain interfaces in the CA of Fig. 5

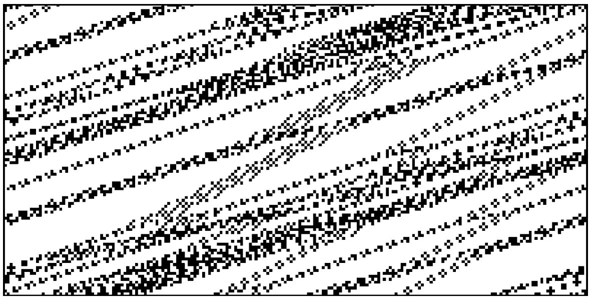
tral [00]* region. The filter used maps all cells in regions that contain a spatial phase defect to 1s.

A single CA may support the emergence of multiple different domain patterns. In many cases one domain dominates and will eventually take over. But this is not always true. An interesting case in which two domains, both chaotic, compete on roughly equal status, is binary radius-2 rule 2614700074, shown in Fig. 5. The two domains have patterns $\Lambda_0 = [0\Sigma]^*$ and $\Lambda_1 = [110\Sigma]^*$, respectively. In the filtered plot, cells in Λ_0 are shown in white, cells in Λ_1 are gray, and all other cells are black. It appears that by about $t = 200$ Λ_0 appears to be winning, but in fact, by about $t = 700$, the entire configuration was in Λ_1, where it remained indefinitely. Depending on the initial condition, one or the other domain was always found to eventually take over with Λ_0 winning about 80% of the time.

The coexistence of multiple domains, each with its own spatial structure, gives rise to a large number of possible interfaces. in general, the number of distinct interface types is governed by the complexity of the pattern in each domain; for 2614700074 it turns out that there are 8 distinct possibilities. Six of these show qualitatively distinct behavior, and are plotted (in filtered form only) in Fig. 6. Note that of the six interfaces, two show a quickly growing region in which defects continually multiply, three of them appear to remain spatially localized, and one (at bottom left) is ambiguous.

Particles in One Dimension

An immediate consequence of the emergence of domains is the simultaneous emergence of boundaries between them. These boundaries may be phase defects, as mentioned in Sect. "Synchronization", but they may also take the form of **particles**. A particle is a small region of cells that separates two domains, persists for a relatively long period of time and remains spatially localized. Particles may be stationary or may move; they may themselves exhibit a pattern that is temporally invariant, periodic, or even disordered.



Cellular Automata, Emergent Phenomena in, Figure 7
Examples of solitons in the one-dimensional Filtering Rule

Solitons

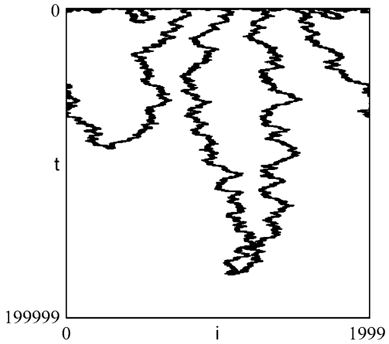
An interesting type of particle emerges in the so-called **soliton** CA, shown in Fig. 7. These CA rules received their name in analogy with the solitons of fluid dynamics, which are solitary traveling waves with the interesting property that two solitons may collide, interact, and pass safely through each other, ultimately recovering their original form as if no collision had taken place. In soliton CA, something similar occurs.

In the simplest case, $k = 2$, the quiescence condition holds with the usual quiescent symbol 0. The solitons or particles embedded in a large lattice of 0s are finite sequences of 1s and 0s that are both temporally periodic (up to a spatial shift) and can collide and pass through each other without being destroyed. A particle consists of a finite sequence of basic strings of length $r + 1$ (where r is the CA radius). The leftmost cell of a particle is always a non-quiescent cell. A particle is bounded on the right by a sequence of $r + 1$ quiescent cells. Under the action of the CA rule, a particle may move to the left or right, may grow or shrink, but ultimately will come back to its original configuration after a finite time p – though possibly shifted by some number of cells. The ratio of the shift and temporal period p determines the particle's velocity V defined in the obvious way: $V =$ (spatial shift)/(temporal period).

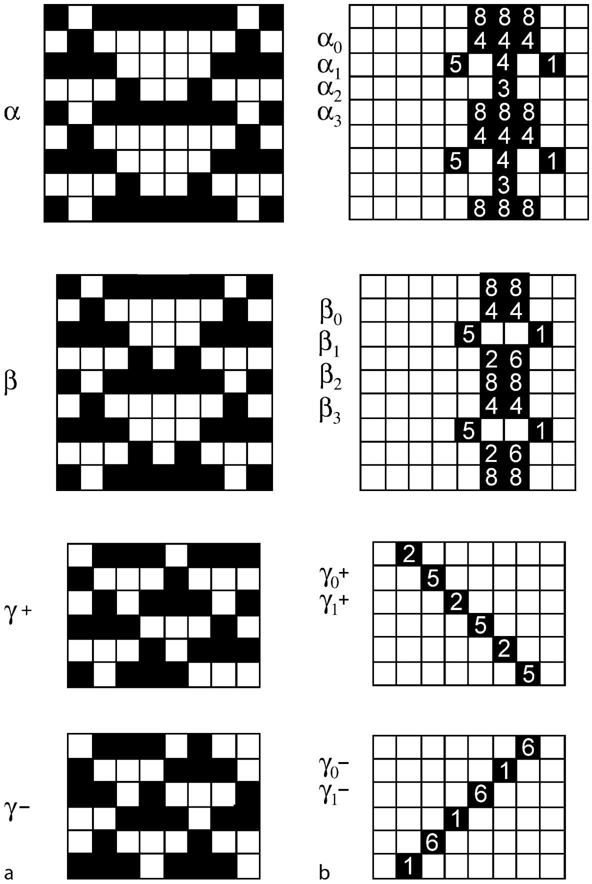
A particle may even temporarily split into two or more smaller particles, so long as eventually they rejoin to form the original configuration. And, as the name implies, two particles with different velocities may collide and pass through each other without being destroyed.

Particles and Defects Defined by Domains

Given the wide variety of domains that arise in CA, the resultant variety of particles that they support is apparently limitless. However, two simple examples may suffice to illustrate these phenomena: ECA 18 and ECA 54, both of which were discussed in the previous section.



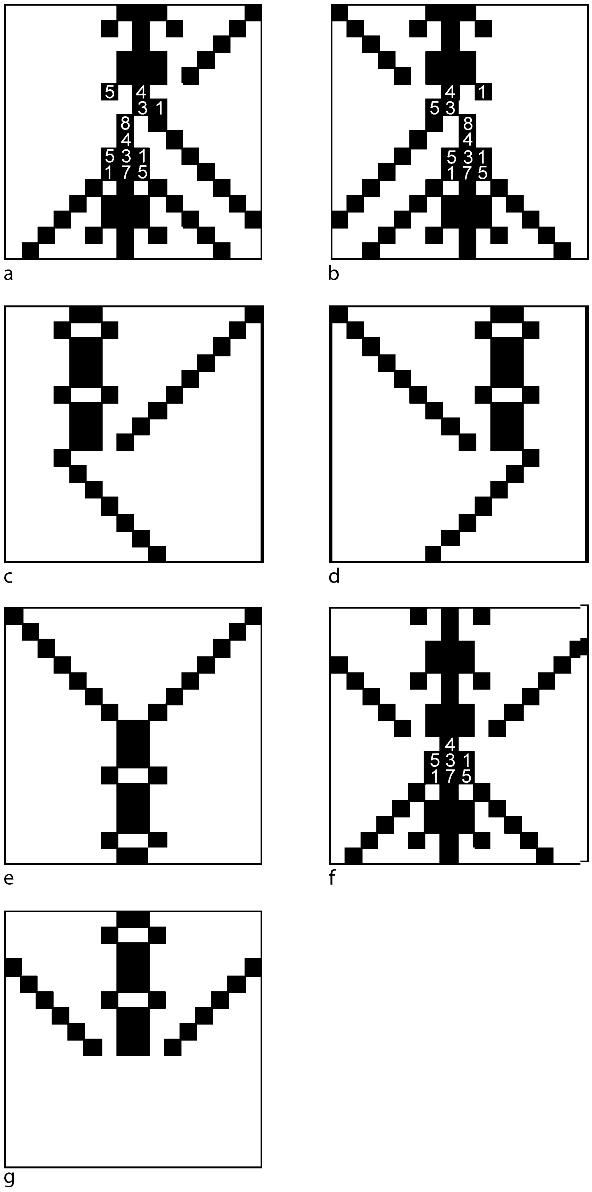
Cellular Automata, Emergent Phenomena in, Figure 8
Long-term behavior of ECA 18



Cellular Automata, Emergent Phenomena in, Figure 9
Fundamental particles in ECA54

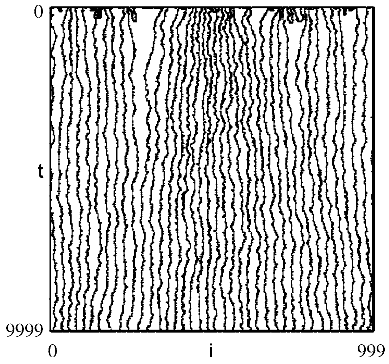
Particles in ECA 18

The spatial phase defects that occur in the domain of ECA 18 (see Fig. 4b) appear, on casual inspection, to be moving more or less at random. It turns out that to a very good approximation, an isolated defect performs a random walk



Cellular Automata, Emergent Phenomena in, Figure 10
Pairwise interactions between fundamental particles in ECA 54

on the lattice [4,7]. When two of them meet, they mutually annihilate. This behavior is purely deterministic, of course; it is caused entirely by the iterated action of the update rule on the initial condition. In effect, the disorder in the domains is causing disorder in the motion of their boundaries. For small systems, and *eventually* on all systems, finite-size effects cause departures from statistical randomness; but otherwise, except for a few highly atypical system-sizes, the defects' behavior is statistically indistinguishable from random motion. Figure 8 shows the long-term behavior of a random initial condition on a relatively large lattice.



Cellular Automata, Emergent Phenomena in, Figure 11
Long-term behavior of ECA 54

Particles in ECA 54

ECA 54 represents an interesting case which can serve to illustrate many the emergent phenomena in one dimensional CA [1,10]. The primary domain gives rise to the so-called "fundamental particles" α, β and γ, shown in Fig. 9. The unfiltered space-time diagrams are shown on the left, and their filtered counterparts on the right. The interactions between the fundamental particles are shown in Fig. 10. In the filtered figures, the numbers inscribed in the black squares are the different outputs of the domain filter; each different sequence of numbers represents a different way in which the domain pattern has been violated.

The long-term behavior of the particles can be seen in Fig. 11. The βs decay relatively quickly, leaving only αs and γs – except for rare cases where a β is created by the interaction in Fig. 10e and persists for a short while, and rarer cases where some other pattern is momentarily present. (Note that the scale of the figure is so compressed that only the αs are visible.) It appears, and is borne out by numerical experiments, that the number of αs decays extremely slowly, and that the system settles into a state in which the αs are roughly equidistant, but move back and forth slightly in a disordered way. Unlike the case of ECA 18, the domains are not disordered, so the particle motion cannot be caused by disorder in the domain. Instead, it comes from the α–γ interactions.

Emergent Phenomena in Two and Higher Dimensions

As might be expected, the emergent phenomena in CA of more than one spatial dimension are at once richer and less systematically studied. All of the phenomena that are observed in one dimension have their analogues in higher dimensions: domains and particle abound. In 2 or more dimensions, "particle" is no longer synonymous with "boundary"; one sees particles that are entirely surrounded by a domain, and spatially-extended boundaries that separate domains. Fundamentally new types of emergent phenomena appear as well.

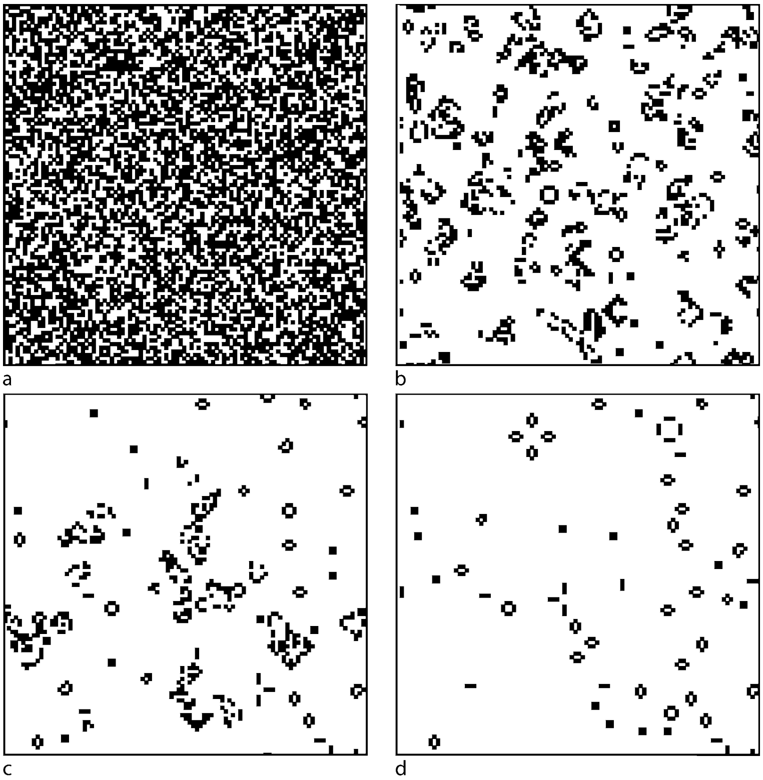
Domains, Particles, and Interfaces

Many of the coherent structures found to exist in Conway's famous Game of Life can be observed to arise spontaneously from random initial conditions, so they properly fall into the category of emergent phenomena. In Fig. 12 a configuration of 100x100 cells is shown at four successive times $t = 0, 50, 900, 1350$. From the random initial condition, a background pattern of 0s quickly emerges, against which there exist a rich variety of particles and disordered structures. By $t = 1350$ the configuration has settled to its final state, in which only a few particles remain, all of which are stationary and have temporal period $p = 1$ or $p = 2$. At intermediate times, various moving structures may be identified: see, for example, the "glider" at $t = 900$, about halfway between the center and the top. In moving about, these inevitably collide with each other or with the stationary particles, eventually leading to the final state.

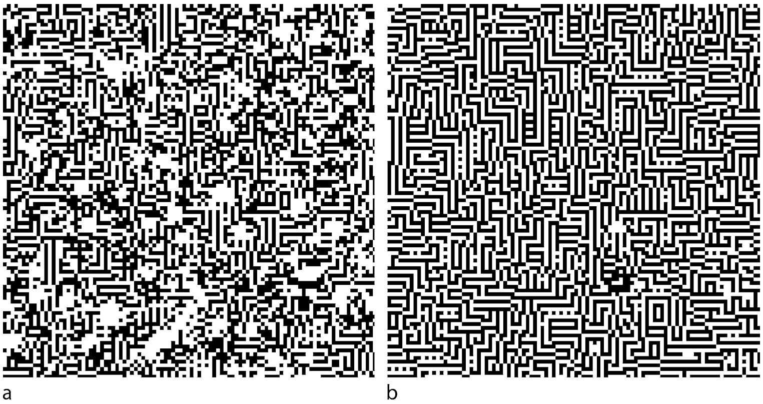
Interestingly enough, a minor variation on the rule gives rise to the patterns shown in Fig. 13. Small regions of horizontal or vertical stripes emerge quickly. Boundaries between them settle down. By $t = 100$, a few non-striped areas persist, along with a few "dotted lines" that take the place of a stripe, and in which the "dots" oscillate. The non-striped areas eventually all disappear. The dotted lines persist indefinitely.

As these examples suggest, 2-dimensional CA support the emergence of synchronized regions, "domains", and particles in close analogy to 1-D CA. The striped regions in Fig. 13 are an example of a two-dimensional, temporally-invariant domain.

Fundamentally new features also appear in two and higher dimensions as well. The most obvious of these is the spatially-extended **interface** or boundary between two adjacent domains. Unlike the one-dimensional case, in



Cellular Automata, Emergent Phenomena in, Figure 12
Conway's Game of Life, starting from a random initial condition. a $t = 0$. b $t = 50$. c $t = 900$. d $t = 1350$

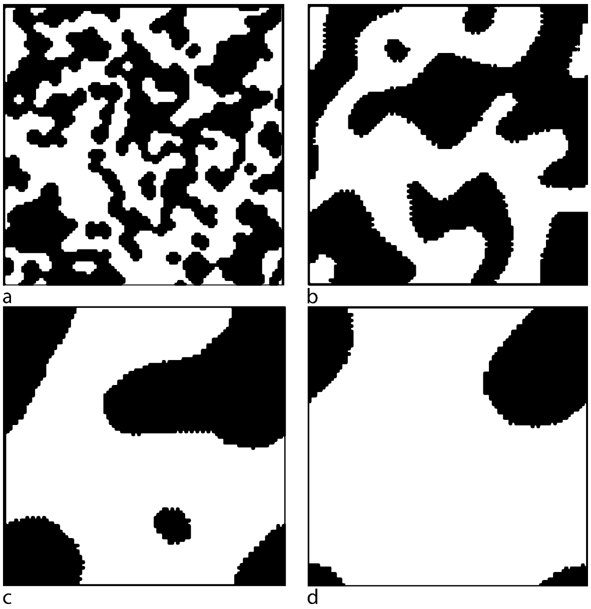


Cellular Automata, Emergent Phenomena in, Figure 13
Variant on Conway's Game of Life, starting from the same random initial condition as in Fig. 12. a $t = 10$. b $t = 100$

which particles and interfaces are more or less the same thing, interfaces in two dimensions are themselves one-dimensional. A characteristic example is seen in the *voting rule*, a 2-D binary CA with von Neumann neighborhood, in which the child cell is determined by the majority of the the local update rule maps a the child cell is equal to the value held by the majority of cells in the parent neighborhood, or if the vote is a tie, by a 0. Figure 14a shows a snapshot at $t = 50$ of the voting rule starting from a random initial condition. The system has organized itself into regions of two domain patterns. The pattern has stabilized by this time and does not change thereafter.

A stochastic variation on the voting rule uses a random variable to break tie votes, resulting it patterns such as Fig. 14b, c and d. Over time, the long boundaries gradually straighten, and small regions of one domain embedded in the other gradually shrink.

A number of extensive tours of patterns observed in selected 2-d CA may be found online; see, for example, [8,14].



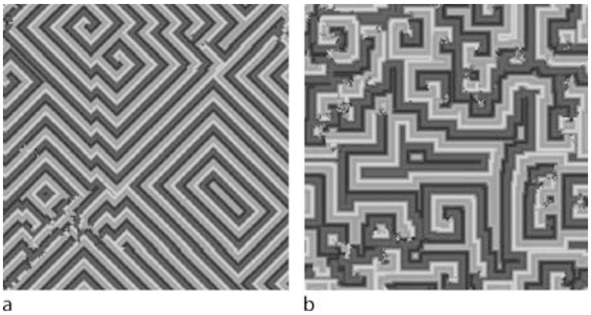
Cellular Automata, Emergent Phenomena in, Figure 14
Two variants of voter rule. a Voter rule at $t = 50$. This configuration is time-invariant. b Voter rule with random tie-breaking at $t = 50$. c Voter rule with random tie- breaking at $t = 250$. d Voter rule with random tie-breaking at $t = 750$

Spiral Waves

Another important class of patterns in 2-D CA are expanding wavelike patterns, as shown in Fig. 15. These are typical of the class of rules called **cyclic CA** [5], and generally evolve to configurations of spirals (as shown). These patterns are not domains in the usual sense, because they have a geometric center. The shape of the spiral is closely related to the shape of the parent neighborhood. Starting from a random initial condition, eventually some number of centers form out from which the spiral waves emanate.

Quasiperiodicity

The final phenomenon to be mentioned here is an intriguing form of emergent phenomenon fundamentally different from what has been discussed above: the emergence of quasiperiodic oscillations in coarse statistical properties of the configuration (such as, percentage of 1s). [2,6] The evidence consists of **return maps**, in which the fraction m_t of 1s at time t, is plotted against the fraction m_{t+1} at time $t + 1$. A synchronized system would show a return map consisting of a single point: $m_t = m_{t+1}$. A periodic sys-



Cellular Automata, Emergent Phenomena in, Figure 15
Spiral waves. a Cyclic CA with $k = 16$, von Neumann neighborhood. b Cyclic CA with $k = 16$, Moore neighborhood

tem would show a sequence of points for the different values of m at the different temporal phases of the sequence, and would have $m_t = m_{t+p}$, where p is the period. The observed return plots, however, showed the characteristic shape of quasiperiodic behavior in nonlinear dynamical systems, which is a sequence of points that eventually map out a roughly continuous, closed curve in the plane. This quasiperiodic behavior was found to occur only in CA of dimension $N > 3$.

Future Directions

This short survey has only been able to hint at the vast wealth of emergent phenomena that arise in CA. Much work yet remains to be done, in classifying the different structures, identifying general laws governing their behavior, and determining the the causal mechanisms that lead them to arise.

For example, there are as yet no general techniques for determining whether a given domain is stable in a given CA; for characterizing the set of initial conditions that will eventually give rise to it; or for working out the particles that it supports. In CA or two or more dimensions, a large body of *descriptive* results are available, but these are more frequently anecdotal than systematic. A significant barrier to progress has been the lack of good mathematical techniques for identifying, describing, and classifying domains. One promising development in this area is an information-theoretic filtering technique that can operate on configurations of any dimension [13].

Bibliography

Primary Literature

1. Boccara N, Nasser J, Roger M (1991) Particlelike structures and their interactions in spatio-temporal patterns generated by one-dimensional deterministic cellular automaton rules. Phys Rev A 44:866
2. Chate H, Manneville P (1992) Collective behaviors in spatially extended systems with local interactions and synchronous updating. Profress Theor Phys 87:1
3. Crutchfield JP, Hanson JE (1993) Turbulent pattern bases for cellular automata. Physica D 69:279
4. Eloranta K, Nummelin E (1992) The kink of cellular automaton rule 18 performs a random walk. J Stat Phys 69:1131
5. Fisch R, Gravner J, Griffeath D (1991) Threshold-range scaling of excitable cellular automata. Stat Comput 1:23–39
6. Gallas J, Grassberger P, Hermann H, Ueberholz P (1992) Noisy collective behavior in deterministic cellular automata. Physica A 180:19
7. Grassberger P (1984) Chaos and diffusion in deterministic cellular automata. Phys D 10:52
8. Griffeath D (2008) The primordial soup kitchen. http://psoup.math.wisc.edu/kitchen.html
9. Hanson JE, Crutchfield JP (1992) The attractor-basin portrait of a cellular automaton. J Stat Phys 66:1415
10. Hanson JE, Crutchfield JP (1997) Computational mechanics of cellular automata: An example. Physica D 103:169
11. Hopcroft JE, Ullman JD (1979) Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, Reading
12. Lindgren K, Moore C, Nordahl M (1998) Complexity of two-dimensional patterns. J Stat Phys 91:909
13. Shalizi C, Haslinger R, Rouquier J, Klinker K, Moore C (2006) Automatic filters for the detection of coherent structure in spatiotemporal systems. Phys Rev E 73:036104
14. Wojtowicz M (2008) Mirek's cellebration. http://www.mirekw.com/ca/
15. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10:1

Books and Reviews

Das R, Crutchfield JP, Mitchell M, Hanson JE (1995) Evolving Globally Synchronized Cellular Automata. In: Eshelman LJ (ed) Proceedings of the Sixth International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo

Gerhardt M, Schuster H, Tyson J (1990) A cellular automaton model of excitable medai including curvature and dispersion. Science 247:1563

Gutowitz HA (1991) Transients, Cycles, and Complexity in Cellular Automata. Phys Rev A 44:R7881

Henze C, Tyson J (1996) Cellular automaton model of three-dimensional excitable media. J Chem Soc Faraday Trans 92:2883

Hordijk W, Shalizi C, Crutchfield J (2001) Upper bound on the products of particle interactions in cellular automata. Physica D 154:240

Iooss G, Helleman RH, Stora R (ed) (1983) Chaotic Behavior of Deterministic Systems. North-Holland, Amsterdam

Ito H (1988) Intriguing Properties of Global Structure in Some Classes of Finite Cellular Automata. Physica 31D:318

Jen E (1986) Global Properties of Cellular Automata. J Stat Phys 43:219

Kaneko K (1986) Attractors, Basin Structures and Information Processing in Cellular Automata. In: Wolfram S (ed) Theory and Applications of Cellular Automata. World Scientific, Singapore, pp 367

Langton C (1990) Computation at the Edge of Chaos: Phase transitions and emergent computation. Physica D 42:12

Lindgren K (1987) Correlations and Random Information in Cellular Automata. Complex Syst 1:529

Lindgren K, Nordahl M (1988) Complexity Measures and Cellular Automata. Complex Syst 2:409

Lindgren K, Nordahl M (1990) Universal Computation in Simple One-Dimensional Cellular Automata. Complex Syst 4:299

Mitchell M (1998) Computation in Cellular Automata: A Selected Review. In: Schuster H, Gramms T (eds) Nonstandard Computation. Wiley, New York

Packard NH (1984) Complexity in Growing Patterns in Cellular Automata. In: Demongeot J, Goles E, Tchuente M (eds) Dynamical Behavior of Automata: Theory and Applications. Academic Press, New York

Packard NH (1985) Lattice Models for Solidification and Aggregation. Proceedings of the First International Symposium on Form, Tsukuba

Pivato M (2007) Defect Particle Kinematics in One-Dimensional Cellular Automata. Theor Comput Sci 377:205–228

Weimar J (1997) Cellular automata for reaction-diffusion systems. Parallel Comput 23:1699

Wolfram S (1984) Computation Theory of Cellular Automata. Comm Math Phys 96:15

Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific Publishers, Singapore

Wuensche A, Lesser MJ (1992) The Global Dynamics of Cellular Automata. Santa Fe Institute Studies in the Science of Complexity, Reference vol 1. Addison-Wesley, Redwood City

Cellular Automata and Groups

TULLIO CECCHERINI-SILBERSTEIN[1],
MICHEL COORNAERT[2]

[1] Dipartimento di Ingegneria, Università del Sannio, Benevento, Italy

[2] Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, Strasbourg, France

Article Outline

Glossary

Groups A *group* is a set G endowed with a binary operation $G \times G \ni (g, h) \mapsto gh \in G$, called the *multiplication*, that satisfies the following properties: (i) for all g, h and k in G, $(gh)k = g(hk)$ (associativity); (ii) there exists an element $1_G \in G$ (necessarily unique) such that, for all g in G, $1_G g = g 1_G = g$ (existence of the identity element); (iii) for each g in G, there exists an element $g^{-1} \in G$ (necessarily unique) such that $gg^{-1} = g^{-1}g = 1_G$ (existence of the inverses).

A group G is said to be *Abelian* (or *commutative*) if the operation is commutative, that is, for all $g, h \in G$ one has $gh = hg$.

A group F is called *free* if there is a subset $S \subset F$ such that any element g of F can be uniquely written as a *reduced word* on S, i. e. in the form $g = s_1^{\alpha_1} s_2^{\alpha_2} \cdots s_n^{\alpha_n}$, where $n \geq 0$, $s_i \in S$ and $\alpha_i \in \mathbb{Z} \setminus \{0\}$ for $1 \leq i \leq n$, and such that $s_i \neq s_{i+1}$ for $1 \leq i \leq n-1$. Such a set S is called a *free basis* for F. The cardinality of S is an invariant of the group F and it is called the *rank* of F.

A group G is *finitely generated* if there exists a finite subset $S \subset G$ such that every element $g \in G$ can be expressed as a product of elements of S and their inverses, that is, $g = s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_n^{\epsilon_n}$, where $n \geq 0$ and $s_i \in S$, $\epsilon_i = \pm 1$ for $1 \leq i \leq n$. The minimal n for which such an expression exists is called the *word length* of g with respect to S and it is denoted by $\ell(g)$. The group G is a (discrete) metric space with the distance function $d\colon G \times G \to \mathbb{R}_+$ defined by setting $d(g, g') = \ell(g^{-1}g')$ for all $g, g' \in G$. The set S is called a *finite generating subset* for G and one says that S is *symmetric* provided that $s \in S$ implies $s^{-1} \in S$.

The *Cayley graph* of a finitely generated group G w.r.t. to a symmetric finite generating subset $S \subset G$ is the (undirected) graph $Cay(G, S)$ with vertex set G and where two elements $g, g' \in G$ are joined by an edge if and only if $g^{-1}g' \in S$.

A group G is *residually finite* if the intersection of all subgroups of G of finite index is trivial.

A group G is *amenable* if it admits a *right-invariant mean*, that is, a map $\mu\colon \mathcal{P}(G) \to [0, 1]$, where $\mathcal{P}(G)$ denotes the set of all subsets of G, satisfying the following conditions: (i) $\mu(G) = 1$ (*normalization*); (ii) $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathcal{P}(G)$ such that $A \cap B = \varnothing$ (*finite additivity*); (iii) $\mu(Ag) = \mu(A)$ for all $g \in G$ and $A \in \mathcal{P}(G)$ (*right-invariance*).

Rings A *ring* is a set R equipped with two binary operations $R \times R \ni (a, b) \mapsto a + b \in R$ and $R \times R \ni (a, b) \mapsto ab \in R$, called the *addition* and the *multiplication*, respectively, such that the following properties are satisfied: (i) R, with the addition operation, is an Abelian group with identity element 0, called the *zero element*, (the inverse of an element $a \in R$ is denoted by $-a$); (ii) the multiplication is associative and admits an identity element 1, called the *unit element*; (iii) multiplication is distributive with respect to addition, that is, $a(b+c) = ab + ac$ and $(b+c)a = ba + ca$ for all a, b and $c \in R$.

A ring R is *commutative* if $ab = ba$ for all $a, b \in R$.

A *field* is a commutative ring $\mathbb{K} \neq \{0\}$ where every non-zero element $a \in \mathbb{K}$ is invertible, that is there exists $a^{-1} \in \mathbb{K}$ such that $aa^{-1} = 1$.

In a ring R a non-trivial element a is called a *zero-divisor* if there exists a non-zero element $b \in R$ such that either $ab = 0$ or $ba = 0$.

A ring R is *directly finite* if whenever $ab = 1$ then necessarily $ba = 1$, for all $a, b \in R$. If the ring $M_d(R)$ of $d \times d$ matrices with coefficients in R is directly finite for all $d \geq 1$ one says that R is *stably finite*.

Let R be a ring and let G be a group. Denote by $R[G]$ the set of all formal sums $\sum_{g \in G} \alpha_g g$ where $\alpha_g \in R$ and $\alpha_g = 0$ except for finitely many elements $g \in G$. We define two binary operations on $R[G]$, namely the addition, by setting

$$\left(\sum_{g \in G} \alpha_g g\right) + \left(\sum_{h \in G} \beta_h h\right) = \sum_{g \in G} (\alpha_g + \beta_g) g \,,$$

and the multiplication, by setting

$$\left(\sum_{g \in G} \alpha_g g\right)\left(\sum_{h \in G} \beta_h h\right) = \sum_{g,h \in G} \alpha_g \beta_h gh$$
$$\equiv_{k=gh} \sum_{g,k \in G} \alpha_g \beta_{g^{-1}k} k.$$

Then, with these two operations, $R[G]$ becomes a ring; it is called the *group ring* of G with coefficients in R.

Cellular automata Let G be a group, called the *universe*, and let A be a set, called the *alphabet*. A *configuration* is a map $x \colon G \to A$. The set A^G of all configurations is equipped with the right action of G defined by $A^G \times G \ni (x, g) \mapsto x^g \in A^G$, where $x^g(g') = x(gg')$ for all $g' \in G$.

A *cellular automaton* over G with coefficients in A is a map $\tau \colon A^G \to A^G$ satisfying the following condition: there exists a finite subset $M \subset G$ and a map $\mu \colon A^M \to A$ such that $\tau(x)(g) = \mu(x^g|_M)$ for all $x \in A^G$, $g \in G$, where $x^g|_M$ denotes the restriction of x^g to M. Such a set M is called a *memory set* and μ is called a *local defining map* for τ.

If $A = V$ is a vector space over a field $\mathbb{K}$, then a cellular automaton $\tau \colon V^G \to V^G$, with memory set $M \subset G$ and local defining map $\mu \colon V^M \to V$, is said to be *linear* provided that μ is linear.

Two configurations $x, x' \in A^G$ are said to be *almost equal* if the set $\{g \in G; x(g) \neq x'(g)\}$ at which they differ is finite. A cellular automaton is called *pre-injective* if whenever $\tau(x) = \tau(x')$ for two almost equal configurations $x, x' \in A^G$ one necessarily has $x = x'$.

A *Garden of Eden configuration* is a configuration $x \in A^G \setminus \tau(A^G)$. Clearly, GOE configurations exist if and only if τ is not surjective.

Definition of the Subject

A cellular automaton is a self-mapping of the set of configurations of a group defined from local and invariant rules. Cellular automata were first only considered on the n-dimensional lattice group $\mathbb{Z}^n$ and for configurations taking values in a finite alphabet set but they may be formally defined on any group and for any alphabet. However, it is usually assumed that the alphabet set is endowed with some mathematical structure and that the local defining rules are related to this structure in some way. It turns out that general properties of cellular automata often reflect properties of the underlying group. As an example, the Garden of Eden theorem asserts that if the group is amenable and the alphabet is finite, then the surjectivity of a cellular automaton is equivalent to its pre-injectivity (a weak form of injectivity). There is also a linear version of the Garden of Eden theorem for linear cellular automata and finite-dimensional vector spaces as alphabets. It is an amazing fact that famous conjectures of Kaplansky about the structure of group rings can be reformulated in terms of linear cellular automata.

Introduction

The goal of this paper is to survey results related to the Garden of Eden theorem and the surjunctivity problem for cellular automata.

The notion of a cellular automaton goes back to John von Neumann [37] and Stan Ulam [34]. Although cellular automata were firstly considered only in theoretical computer science, nowadays they play a prominent role also in physics and biology, where they serve as models for several phenomena (► Cellular Automata Modeling of Physical Systems, ► Chaotic Behavior of Cellular Automata), and in mathematics. In particular, cellular automata are studied in ergodic theory (► Entropy in Ergodic Theory, ► Ergodic Theory: Basic Examples and Constructions, ► Ergodic Theory of Cellular Automata, ► Ergodicity and Mixing Properties) and in the theory of dynamical systems (► Topological Dynamics of Cellular Automata, ► Symbolic Dynamics), in functional and harmonic analysis (▶ Spectral Theory of Dynamical Systems), and in group theory.

In the classical framework, the universe U is the lattice $\mathbb{Z}^2$ of integer points in Euclidean plane and the alphabet A is a finite set, typically $A = \{0, 1\}$. The set $A^U = \{x : U \to A\}$ is the configuration space, a map $x : U \to A$ is a configuration and a point $(n, m) \in U$ is called a cell. One is given a neighborhood M of the origin $(0, 0) \in U$, typically, for some $r > 0$, $M = \{(n, m) \in \mathbb{Z}^2 : |n| + |m| \leq r\}$ (von Neumann r-ball) or $M = \{(n, m) \in \mathbb{Z}^2 : |n|, |m| \leq r\}$ (Moore's r-ball) and a local map $\mu : A^M \to A$. One then "extends" μ to the whole universe obtaining a map $\tau : A^U \to A^U$, called a cellular automaton, by setting $\tau(x)(n, m) = \mu(x(n + s, m + t)_{(s,t)\in M})$. This way, the value $\tau(x)(n, m) \in A$ of the configuration x at the cell $(n, m) \in U$ only depends on the values $x(n+s, m+s)$ of x at its neighboring cells $(x + s, y + t) \equiv (x, y) + (s, t) \in (x, y) + M$, in other words, τ is $\mathbb{Z}^2$-equivariant. M is called a *memory set* for τ and μ a *local defining map*.

In 1963 E.F. Moore proved that if a cellular automaton $\tau : A^{\mathbb{Z}^2} \to A^{\mathbb{Z}^2}$ is surjective then it is also pre-injective, a weak form of injectivity. Shortly later, John Myhill proved the converse to Moore's theorem. The equivalence of surjectivity and pre-injectivity of cellular automata is referred to as the Garden of Eden theorem (briefly GOE theorem), this biblical terminology being motivated by the fact that it gives necessary and sufficient conditions for the existence of configurations x that are not in the image of τ, i. e. $x \in A^{\mathbb{Z}^2} \setminus \tau(A^{\mathbb{Z}^2})$, so that, thinking of $(\tau, A^{\mathbb{Z}^2})$ as a discrete dynamical system, with τ being the time, they can appear only as "initial configurations".

It was immediately realized that the GOE theorem was holding also in higher dimension, namely for cellular automata with universe $U = \mathbb{Z}^d$, the lattice of integer points in the d-dimensional space. Then, Machì and Mignosi [27] gave the definition of a cellular automaton over a finitely generated group and extended the GOE theorem to the class of groups G having sub-exponential growth, that is for which the growth function $\gamma_G(n)$, which counts the elements $g \in G$ at "distance" at most n from the unit element 1_G of G, has a growth weaker than the exponential, in formulæ, $\lim_{n\to\infty} \sqrt[n]{\gamma_G(n)} = 1$. Finally, in 1999 Ceccherini-Silberstein, Machì and Scarabotti [9] extended the GOE theorem to the class of amenable groups. It is interesting to note that the notion of an amenable group was also introduced by von Neumann [36]. This class of groups contains all finite groups, all Abelian groups, and in fact all solvable groups, all groups of sub-exponential growth and it is closed under the operation of taking subgroups, quotients, directed limits and extensions. In [27] two examples of cellular automata with universe the free group F_2 of rank two, the prototype of a non-amenable group, which are surjective but not pre-injective and, conversely, pre-injective but not surjective, thus providing an instance of the failure of the theorems of Moore and Myhill and so of the GOE theorem. In [9] it is shown that this examples can be extended to the class of groups, thus necessarily non-amenable, containing the free group F_2. We do not know whether the GOE theorem only holds for amenable groups or there are examples of groups which are non-amenable and have no free subgroups: by results of Olshanskii [30] and Adyan [1] it is know that such class is non-empty. In 1999 Misha Gromov [20], using a quite different terminology, reproved the GOE for cellular automata whose universes are infinite amenable graphs Γ with a dense pseudogroup of holonomies (in other words such Γs are rich in symmetries). In addition, he considered not only cellular automata from the full configuration space A^Γ into itself but also between subshifts $X, Y \subset A^\Gamma$. He used the notion of entropy of a subshift (a concept hidden in the papers [27] and [9].

In the mid of the fifties W. Gottschalk introduced the notion of surjunctivity of maps. A map $f : X \to Y$ is surjunctive if it is surjective or not injective. We say that a group G is surjunctive if all cellular automata $\tau : A^G \to A^G$ with finite alphabet are surjunctive. Lawton [18] proved that residually finite groups are surjunctive. From the GOE theorem for amenable groups [9] one immediately deduce that amenable groups are surjunctive as well. Finally Gromov [20] and, independently, Benjamin Weiss [38] proved that all sofic groups (the class of sofic groups contains all residually finite groups and all amenable groups) are surjunctive. It is not known whether or not all groups are surjunctive.

In the literature there is a notion of a linear cellular automaton. This means that the alphabet is not only a finite set but also bears the structure of an Abelian group and that the local defining map μ is a group homomorphism, that is, it preserves the group operation. These are also called *additive cellular automata* (▶ Additive Cellular Automata).

In [5], motivated by [20], we introduced another notion of linearity for cellular automata. Given a group G and a vector space V over a (not necessarily finite) field $\mathbb{K}$, the configuration space is V^G and a cellular automaton $\tau : V^G \to V^G$ is linear if the local defining map $\mu : V^B \to V$ is $\mathbb{K}$-linear. The set $\mathrm{LCA}(V, G)$ of all linear cellular automata with alphabet V and universe G naturally bears a structure of a ring.

The finiteness condition for a set A in the classical framework is now replaced by the finite dimensionality of V. Similarly, the notion of entropy for subshifts $X \subset A^G$

is now replaced by that of mean-dimension (a notion due to Gromov [20]). In [5] we proved the GOE theorem for linear cellular automata $\tau\colon V^G \to V^G$ with alphabet a finite dimensional vector space and with G an amenable group. Moreover, we proved a linear version of Gottschalk surjunctivity theorem for residually finite groups.

In the same paper we also establish a connection with the theory of group rings. Given a group G and a field $\mathbb{K}$, there is a one-to-one correspondence between the elements in the group ring $\mathbb{K}[G]$ and the cellular automata $\tau\colon \mathbb{K}^G \to \mathbb{K}^G$. This correspondence preserves the ring structures of $\mathbb{K}[G]$ and $\mathrm{LCA}(\mathbb{K}, G)$. This led to a reformulation of a long standing problem, raised by Irving Kaplansky [23], about the absence of zero-divisors in $\mathbb{K}[G]$ for G a torsion-free group, in terms of the pre-injectivity of all $\tau \in \mathrm{LCA}(\mathbb{K}, G)$.

In [6] we proved the linear version of the Gromov–Weiss surjunctivity theorem for sofic groups and established another application to the theory of group rings. We extended the correspondence above to a ring isomorphism between the ring $\mathrm{Mat}_d(\mathbb{K}[G])$ of $d \times d$ matrices with coefficients in the group ring $\mathbb{K}[G]$ and $\mathrm{LCA}(\mathbb{K}^d, G)$. This led to a reformulation of another famous problem, raised by Irving Kaplansky [24] about the structure of group rings. A group ring $\mathbb{K}[G]$ is stably finite if and only if, for all $d \geq 1$, all linear cellular automata $\tau\colon (\mathbb{K}^d)^G \to (\mathbb{K}^d)^G$ are surjunctive. As a byproduct we obtained another proof of the fact that group rings over sofic groups are stably finite, a result previously established by G. Elek and A. Szabó [11] using different methods.

The paper is organized as follows. In Sect. “Cellular Automata” we present the general definition of a cellular automaton for any alphabet and any group. This includes a few basic examples, namely Conway's Game of Life, the majority action and the discrete Laplacian. In the subsequent section we specify our attention to cellular automata with a finite alphabet. We present the notions of Cayley graphs (for finitely generated groups), of amenable groups, and of entropy for G-invariant subsets in the configuration space. This leads to a description of the setting and the statement of the Garden of Eden theorem for amenable groups. We also give detailed expositions of a few examples showing that the hypotheses of amenability cannot, in general, be removed from the assumption of this theorem. We also present the notion of surjunctivity, of sofic groups and state the surjunctivity theorem of Gromov and Weiss for sofic groups. In Sect. “Linear Cellular Automata” we introduce the notions of linear cellular automata and of mean dimension for G-invariant subspaces in V^G. We then discuss the linear analogue of the Garden of Eden theorem and, again, we provide explicit examples showing that the assumptions of the theorem (amenability of the group and finite dimensionality of the underlying vector space) cannot, in general, be removed. Finally we present the linear analogue of the surjunctivity theorem of Gromov and Weiss for linear cellular automata over sofic groups. In Sect. “Group Rings and Kaplansky Conjectures” we give the definition of a group ring and present a representation of linear cellular automata as matrices with coefficients in the group ring. This leads to the reformulation of the two long standing problems raised by Kaplansky about the structure of group rings.

Finally, in Sect. “Future Directions” we present a list of open problems with a description of more recent results related to the Garden of Eden theorem and to the surjunctivity problem.

Cellular Automata

The Configuration Space

Let G be a group, called the *universe*, and let A be a set, called the *alphabet* or the *set of states*. A *configuration* is a map $x\colon G \to A$. The set A^G of all configurations is equipped with the right action of G defined by $A^G \times G \ni (x, g) \mapsto x^g \in A^G$, where $x^g(g') = x(gg')$ for all $g' \in G$.

Cellular Automata

A *cellular automaton* over G with coefficients in A is a map $\tau\colon A^G \to A^G$ satisfying the following condition: there exists a finite subset $M \subset G$ and a map $\mu\colon A^M \to A$ such that

$$\tau(x)(g) = \mu(x^g|_M) \tag{1}$$

for all $x \in A^G$, $g \in G$, where $x^g|_M$ denotes the restriction of x^g to M. Such a set M is called a *memory set* and μ is called a *local defining map* for τ.

It follows directly from the definition that every cellular automaton $\tau\colon A^G \to A^G$ is *G-equivariant*, i. e., it satisfies

$$\tau(x^g) = \tau(x)^g \tag{2}$$

for all $g \in G$ and $x \in A^G$.

Note that if M is a memory set for τ, then any finite set $M' \subset G$ containing M is also a memory set for τ. The local defining map associated with such an M' is the map $\mu'\colon A^{M'} \to A$ given by $\mu' = \mu \circ \pi$, where $\pi\colon A^{M'} \to A^M$ is the restriction map. However, there exists a unique memory set M_0 of minimal cardinality. This memory set M_0 is called the *minimal* memory set for τ.

We denote by $\mathrm{CA}(G, A)$ the set of all cellular automata over G with alphabet A.

Examples

Example 1 (Conway's Game of Life [3]*)* The most famous example of a cellular automaton is the *Game of Life* of John Horton Conway. The set of states is $A = \{0, 1\}$. State 0 corresponds to *absence of life* while state 1 indicates *life*. Therefore passing from 0 to 1 can be interpreted as *birth*, while passing from 1 to 0 corresponds to *death*.

The universe for Life is the group $G = \mathbb{Z}^2$, that is, the free Abelian group of rank 2. The minimal memory set is $M = \{-1, 0, 1\}^2 \subset \mathbb{Z}^2$. The set M is the *Moore neighborhood* of the origin in $\mathbb{Z}^2$. It consists of the origin $(0, 0)$ and its eight neighbors $\pm(1, 0), \pm(0, 1), \pm(1, 1), \pm(-1, 1)$. The corresponding local defining map $\mu\colon A^M \to A$ given by

$$\mu(y) = \begin{cases} 1 & \text{if } \begin{cases} \sum_{s\in S} y(s) = 3 \\ \text{or} \\ \sum_{s\in S} y(s) = 4 \text{ and } y((0,0)) = 1\,, \end{cases} \\ 0 & \text{otherwise}\,. \end{cases}$$

Example 2 (The majority action [21]*)* Let G be a group, M a finite subset of G, and $A = \{0, 1\}$. The automaton $\tau\colon A^G \to A^G$ with memory set M and local defining map $\mu\colon A^M \to A$ given by

$$\mu(y) = \begin{cases} 1 & \text{if } \sum_{m\in M} y(m) > \frac{|M|}{2}\,, \\ 0 & \text{if } \sum_{m\in M} y(m) < \frac{|M|}{2}\,, \\ y(1_G) & \text{if } \sum_{m\in M} y(m) = \frac{|M|}{2}\,, \end{cases}$$

for all $y \in A^M$, is the *majority action automaton* associated with G and M.

Example 3 Let G be a group and let A be any alphabet. Let $f\colon A \to A$ be a map and consider the map $\tau_f\colon A^G \to A^G$ defined by setting $\tau_f(x)(g) = f(x(g))$ for all $x \in A^G$ and $g \in G$. Then τ_f is a cellular automaton with memory set $M = \{1_G\}$ and the local defining map $\mu\colon A^M \to A$ given by $y \mapsto f(y(1_G))$ for all $y \in A^M$. When $f = \iota_A$ is the identity map on A then $\tau_{\iota_A} =: I$ is the identity map on A^G. On the other hand, given $c \in A$, if $f = f_c$ is the constant map given by $f_c(a) = c$ for all $a \in A$, then τ_{f_c} is the *constant* cellular automaton defined by $\tau_{f_c}(x) = x_c$ for all $x \in A^G$, where $x_c(g) = c$ for all $g \in G$.

Example 4 (The discrete Laplacian) Let G be a group, S a finite subset of G not containing 1_G, and $A = \mathbb{R}$, the field of *real numbers*. The (linear) map $\Delta_S\colon \mathbb{R}^G \to \mathbb{R}^G$ defined by

$$\Delta_S(x)(g) = x(g) - \frac{1}{|S|}\sum_{s\in S} x(gs)$$

is a cellular automaton over G with memory set $M = S \cup \{1_G\}$ and local defining map $\mu\colon \mathbb{R}^M \to \mathbb{R}$ given by $\mu(y) = y(1_G) - \frac{1}{|S|}\sum_{s\in S} y(s)$, for all $y \in \mathbb{R}^M$. It is called the *Laplacian* or *Laplace operator* on G relative to S.

Let $\tau_1, \tau_2 \in \mathrm{CA}(G, A)$ be two cellular automata (with memory sets M_1 and M_2, respectively). It is easy to see that their composition $\tau_1 \circ \tau_2$, defined by $[\tau_1 \circ \tau_2](x) = \tau_1(\tau_2(x))$ for all $x \in A^G$, is a cellular automaton (admitting $M = M_1M_2$ as a memory set). Since the identity map $I\colon A^G \to A^G$ is a cellular automaton, it follows that $\mathrm{CA}(G, A)$ is a monoid for the composition operation.

Cellular Automata with a Finite Alphabet

The Configuration Space as a Metric Space

Let G be a countable group, e. g., a finitely generated group (see Subsect. "Cayley Graphs") and let A be a finite alphabet with at least two elements.

The set A^G of all configurations can be equipped with a metric space structure as follows. Let $\emptyset = E_1 \subset E_2 \subset \cdots \subset E_n \subset E_{n+1} \subset \cdots$ be an increasing sequence of finite subsets of G such that $\cup_{n\geq 1} E_n = G$. Then, given any two configurations $x, x' \in A^G$, we set:

$$d(x, x') = 1/\sup\left\{n \in \mathbb{N} : x|_{E_n} = x'|_{E_n}\right\} \qquad (3)$$

(we use the convention that $1/\infty = 0$). In this way, A^G becomes a compact totally disconnected space homeomorphic to the middle third Cantor set.

We then have Hedlund's topological characterization of cellular automata.

Theorem 1 (Hedlund) *Suppose that A is a finite set. A map $\tau\colon A^G \to A^G$ is a cellular automaton if and only if it is continuous and G-equivariant.*

Corollary 1 *Suppose that A is a finite set. Let $\tau\colon A^G \to A^G$ be a bijective cellular automaton. Then the inverse map $\tau^{-1}\colon A^G \to A^G$ is also a cellular automaton.*

The Garden of Eden Theorem of Moore and Myhill

Let G be any group and A be any alphabet. Let $F \subset G$ be a finite subset. A *pattern* with *support* F is a map $p\colon A^F \to A$.

Let now $\tau\colon A^G \to A^G$ be a cellular automaton. One says that τ is *surjective* if $\tau(A^G) = A^G$. One often thinks of τ as describing time evolution: if $x \in A^G$ is the configuration of the universe at time t, then $\tau(x)$ is the configuration of the universe at time $t + 1$. An *initial configuration* is a configuration at time $t = 0$. A configuration x which is not in the image of τ, namely such that

$x \in A^G \setminus \tau(A^G)$, is called a *Garden of Eden* (briefly *GOE*) *configuration*. This biblical terminology is motivated by the fact that GOE configurations may only appear as initial configurations. Analogously, a pattern p with support $F \subset G$ is called a *GOE pattern* if $p \neq \tau(x)|_F$ for all $x \in A^G$. Using topological methods it is easy to see that, when the alphabet is finite, the existence of GOE patterns for τ is equivalent to the existence of GOE configurations for τ, i. e., to the non-surjectivity of τ.

One says that τ is *injective* if, for $x, x' \in A^G$, one has $x = x'$ whenever $\tau(x) = \tau(x')$.

Two configurations $x, x' \in A^G$ are *almost equal*, and we write $x \sim_{a.e.} x'$, if they coincide outside a finite subset of G, namely $|\{g \in G : x(g) \neq x'(g)\}| < \infty$. Finally, using terminology introduced by Gromov, one says that τ is *pre-injective* if, for $x, x' \in A^G$ s.t. $x \sim_{a.e.} x'$, one has $x = x'$ whenever $\tau(x) = \tau(x')$.

Two patterns p, p' with the same support F are *mutually erasable* if they are distinct and whenever $x, x' \in A^G$ are two configurations which extend in the same way p and p' outside of F (i. e. $x|_F = p$, $x'|_F = p'$ and $x|_{G\setminus F} = x'|_{G\setminus F}$), then $\tau(x) = \tau(x')$. The non-existence of mutually erasable patterns is equivalent to the pre-injectivity of the cellular automaton. Finally note that injectivity implies pre-injectivity (but the converse is false, in general).

The following is the celebrated Garden of Eden theorem of Moore and Myhill.

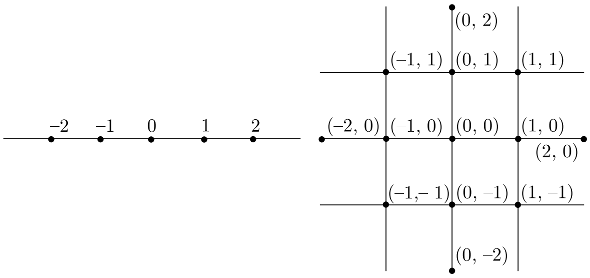
Theorem 2 (Moore and Myhill) *Let $\tau \in CA(\mathbb{Z}^2, A)$ be a cellular automaton with coefficients in a finite set A. Then τ is surjective if and only if it is pre-injective.*

The necessary condition is due to Moore, the converse implication to Myhill.

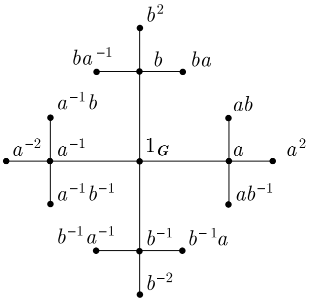
As Conway's Game of Life is concerned, we have that this cellular automaton is clearly not pre-injective (the constant dead configuration and the configuration with only one live cell have the same image) and by the previous theorem it is not surjective either. We mention that the non-surjectivity of the Game of Life is not trivial: the smallest GOE pattern known up to now has as a support a rectangle 13 × 12 with 81 live cells.

Cayley Graphs

A group G is said to be *finitely generated* if there exists a finite subset $S \subset G$ such that every element $g \in G$ can be expressed as a product of elements of S and their inverses, that is, $g = s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_n^{\epsilon_n}$, where $n \geq 0$ and $s_i \in S, \epsilon_i = \pm 1$ for $1 \leq i \leq n$. The minimal n for which such an expression exists is called the *word length* of g with respect to S and it is denoted by $\ell(g)$. The group G is a (discrete) metric space with the distance function $d \colon G \times G \to \mathbb{R}_+$ defined by setting $d(g, g') = \ell(g^{-1}g')$ for all $g, g' \in G$. The set S is called a *finite generating subset* for G and one says that S is *symmetric* provided that $s \in S$ implies $s^{-1} \in S$.



Cellular Automata and Groups, Figure 1
The ball $B(2, 1_G)$ in $\mathbb{Z}$ and in $\mathbb{Z}^2$, respectively



Cellular Automata and Groups, Figure 2
The ball $B(2, 1_G)$ in F_2

Suppose that G is finitely generated and let S be a symmetric finite generating subset of G. The *Cayley graph* of G w.r.t. S is the (undirected) graph $Cay(G, S)$ with vertex set G and two elements $g, g' \in G$ are joined by an edge if and only if $g^{-1}g' \in S$. The group G becomes a (discrete) metric space by introducing the distance $d \colon G \times G \to \mathbb{R}_+$ defined by $d(g, g') = \ell(g^{-1}g')$ for all $g, g' \in G$. Note that the distance d coincides with the graph distance on the Cayley graph $Cay(G, S)$. For $g \in G$ and $n \in \mathbb{N}$ we denote by $B(n, g) = \{g' \in G \colon d(g, g') \leq n\}$ the *ball of radius n with center g*.

Amenable Groups

The notion of an amenable group is also due to J. von Neumann [36]. Let G be a group and denote by $\mathcal{P}(G)$ the set of all subsets of G. The group G is said to be *amenable* if there exists a *right-invariant mean*, that is, a map $\mu \colon \mathcal{P}(G) \to [0, 1]$ satisfying the following conditions:

1. $\mu(G) = 1$ (*normalization*);

2. $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathcal{P}(G)$ such that $A \cap B = \varnothing$ (*finite additivity*);
3. $\mu(Ag) = \mu(A)$ for all $g \in G$ and $A \in \mathcal{P}(G)$ (*right-invariance*).

We mention that if G is amenable, namely such a right-invariant mean exists, then also left-invariant and in fact even bi-invariant means do exist. The class of amenable groups includes, in particular, all finite groups, all Abelian groups (and, more generally, all solvable groups), and all finitely generated groups of subexponential growth. It is closed under the operations of taking subgroups, taking factors, taking extensions and taking direct limits.

It was observed by von Neumann himself [36] that the free group F_2 based on two generators is non-amenable. Therefore, all groups containing a subgroup isomorphic to the free group F_2 are non-amenable as well.

However, there are examples of non-amenable groups which do not contain subgroups isomorphic to F_2. The first examples are due to Olshanski [30]; later Adyan [1] showed that also the free Burnside groups $B(m, n) = \langle s_1, s_2, \dots, s_m : w^n \rangle$ of rank $m \geq 2$ and odd *exponent* $n \geq 665$ are non-amenable.

It follows from a result of Følner [16] that a countable group G is amenable if and only if it admits a *Følner sequence*, i. e., a sequence $(F_n)_{n\in\mathbb{N}}$ of non-empty finite subsets of G such that

$$\lim_{n\to\infty} \frac{|F_n \triangle gF_n|}{|F_n|} = 0 \quad \text{for all } g \in G, \tag{4}$$

where $F_n \triangle gF_n = (F_n \cup gF_n) \setminus (F_n \cap gF_n)$ is the symmetric difference of F_n and gF_n.

For instance, for $G = \mathbb{Z}$ one can take as Følner sets the intervals $[-n, n]$ where $[-n, n] = \{-n, -n+1, \dots, -1, 0, 1, \dots, n\}$, $n \in \mathbb{N}$. Analogously, for $G = \mathbb{Z}^2$ one can take as Følner sets the squares $F_n = [-n, n] \times [-n, n]$.

Suppose that G is countable and amenable, and fix a Følner sequence $(F_n)_{n\in\mathbb{N}}$. Let A be a finite alphabet. A subset $X \subset A^G$ is said to be *G-invariant* if $x \in X$ implies that $x^g \in X$ for all $g \in G$. The *entropy* $\operatorname{ent}(X)$ of a G-invariant subset $X \subset A^G$ is defined by

$$\operatorname{ent}(X) = \lim_{n\to\infty} \frac{\log |X_{F_n}|}{|F_n|} \tag{5}$$

where, for any subset $F \subset G$

$$X_F = \{x|_F : x \in X\} \tag{6}$$

denotes the set of restrictions to F of all configurations in X.

By using a result of Ornstein and Weiss [31], it can be shown that the above limit in (5) exists and does not depend on the particular Følner sequence $(F_n)_{n\in\mathbb{N}}$.

One clearly has $\operatorname{ent}(A^G) = \log |A|$ and $\operatorname{ent}(X) \leq \operatorname{ent}(Y)$ if $X \subset Y$ are G-invariant subsets of A^G.

Theorem 3 (Ceccherini-Silberstein, Machì and Scarabotti [9]) *Let G be a countable amenable group and let A be a finite set. Let $\tau\colon A^G \to A^G$ be a cellular automaton. The following are equivalent:*

(a) τ is surjective (i. e. there are no GOE configurations);
(b) τ is pre-injective;
(c) $\operatorname{ent}(\tau(A^G)) = \log |A|$.

Example 5 Let G be a group. Let M be a finite subset of G with at least three elements. Let $A = \{0, 1\}$ and consider the majority action cellular automaton $\tau\colon A^G \to A^G$ associated with G and M (see Subsect. "Cellular Automata"). Clearly τ is not pre-injective. Indeed the configurations $x_1, x_2 \in A^G$ defined by $x_1(g) = 0$ for all $g \in G$ and

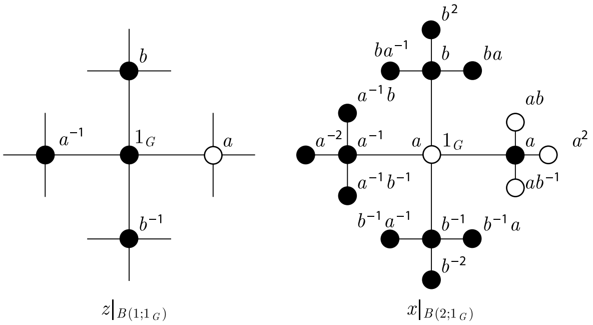
$$x_2(g) = \begin{cases} 1 & \text{if } g = 1_G \\ 0 & \text{otherwise} \end{cases}$$

are almost equal and $\tau(x_2) = x_1 = \tau(x_1)$. By applying Theorem 3 we deduce that τ is not surjective when G is a countable amenable group.

In the example below we show that for the non-amenable group F_2, the free group of rank two, the implication (a) $\Rightarrow$ (b) fails to hold. In Example 9 in Sect. "Linear Cellular Automata" we show that also the converse implication fails to hold, in general, for cellular automata over F_2.

Example 6 Let $G = F_2$ be the free group on two generators a and b. Take $A = \{0, 1\}$ and $M = \{a, a^{-1}, b, b^{-1}\} = S$. Consider the majority action cellular automaton $\tau\colon A^G \to A^G$ associated with G and M. As observed above, τ is not pre-injective. However, τ is surjective. To see this, let $z \in A^G$. Let us show that there exists $x \in A^G$ such that $\tau(x) = z$. We first set $x(1_G) = 0$. For $g \in G$ such that $g \neq 1_G$, denote by $g' \in G$ the unique element such that $\ell(g') = \ell(g) - 1$ and $g = g's'$ for some $s' \in S$. Then set $x(g) = z(g')$. We clearly have $\tau(x) = z$. This shows that τ is surjective (see Fig. 3).

Recently, Laurent Bartholdi [2] proved that if G is a non-amenable group, then there exists a cellular automaton $\tau\colon A^G \to A^G$ with finite alphabet which is surjective but not pre-injective. In other words, the implication "(a) τ is surjective $\Rightarrow$ (b) τ is pre-injective" in Theorem 3 (which corresponds to the generalization of Moore's theorem)



Cellular Automata and Groups, Figure 3
The construction of $x \in A^G$ such that $\tau(x) = z$

holds true only if the group G is amenable. In particular, the Garden of Eden Theorem (Theorem 3) holds true if and only if the universe G is amenable. This clearly gives a new characterization of amenability for groups in terms of cellular automata.

However, up to now, it is not known whether the validity of the implication (b) ⇒ (a) in Theorem 3 (which corresponds to the generalization of Myhill's theorem) holds true only if the group G is amenable.

Surjunctivity

A group G is said to be *surjunctive* (a terminology due to Gottschalk [17]) if given any finite alphabet A, every injective cellular automaton $\tau: A^G \to A^G$ is surjective. In other words, uniqueness implies existence for solutions of the equation $y = \tau(x)$. This property is reminiscent of several other classes of mathematical objects for which injective endomorphisms are always surjective (finite sets, finite-dimensional vector spaces, Artinian modules, complex algebraic varieties, co-Hopfian groups, etc.).

Recall that a group G is said to be *residually finite* if for every element $g \neq 1_G$ in G there exist a finite group F and a homomorphism $h: G \to F$ such that $h(g) \neq 1_F$. This amounts to saying that the intersection of all subgroups of finite index of G is reduced to the identity element. From a dynamical viewpoint we have the following characterization of residual finiteness. Given a finite set A a configuration $x \in A^G$ is said to be *periodic* if its *G-orbit* $\{x^g: g \in G\} \subset A^G$ is finite. Then G is residually finite if and only if the set of periodic configurations is dense in A^G.

The class of residually finite groups is quite large. For instance, every finitely generated subgroup of $\mathrm{GL}_n(\mathbb{C})$, the group of n by n invertible matrices over the complex numbers, is residually finite. However, there are finitely generated amenable groups which are not residually finite.

Lawton [17,18] proved that residually finite groups are surjunctive.

From Theorem 3 one immediately deduces the following

Corollary 2 *Amenable groups are surjunctive.*

Note that the implication "surjectivity ⇒ injectivity" fails to hold, in general, for cellular automata with finite alphabet over amenable groups, even for $G = \mathbb{Z}$. Take, for instance $A = \{0, 1\}$ and $\tau: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ defined by $\tau(x)(n) = x(n+1) - x(n)$ for all $x \in A^{\mathbb{Z}}$ and $n \in \mathbb{Z}$. This cellular automaton is surjective but not injective. See also Example 8 below.

Sofic Groups

Let S be a set. An *S-labeled graph* is a triple (Q, E, λ), where Q is a set, E is a symmetric subset of $Q \times Q$ and λ is a map from E to S. The set Q is the set of *vertices*, E is the set of *edges* and $\lambda: E \to S$ is the *labeling map* of the S-labeled graph (Q, E, λ). We shall view every subgraph of a labeled graph as a labeled graph in the obvious way. Also, for $r \in \mathbb{R}$ and $q \in Q$, we denote by $B(q, r) = \{q' \in Q: d(q, q') \leq r\}$ the *ball* of *radius* r centered at q (here d denotes the *graph distance* in Q).

Let (Q, E, λ) and (Q', E', λ') be S-labeled graphs. Two vertices $q \in Q$ and $q' \in Q'$ are said to be *r-equivalent*, and we write $q \sim_r q'$, if the balls $B(q, r)$ and $B(q', r)$ are isomorphic as labeled graphs (i. e. there is a bijection $\varphi: B(q, r) \to B(q', r)$ sending q to q' such that $(q_1, q_2) \in E \cap (B(q, r) \times B(q, r))$ if and only if $(\varphi(q_1), \varphi(q_2)) \in E' \cap (B(q', r) \times B(q', r))$ and $\lambda(q_1, q_2) = \lambda'(\varphi(q_1), \varphi(q_2))$.

Let G be a finitely generated group and let S be a finite symmetric ($S = S^{-1}$) generating subset of G. We denote by *Cay*(G, S) the *Cayley graph* of G with respect to S. Its vertex set is G and $(g, g') \in G \times G$ is an edge if $s := g^{-1}g' \in S$, if this is the case, its label is $\lambda(g, g') = s$.

The group G is said to be *sofic* if for all $\varepsilon > 0$ and $r \in \mathbb{N}$ there exists a finite S-labeled graph (Q, E, λ) such that the set $Q(r) \subset Q$ defined by $Q(r) = \{q \in Q: q \sim_r 1_G\}$ (here 1_G is considered as a vertex in *Cay*(G, S)) satisfies

$$|Q(r)| \geq (1 - \varepsilon)|Q|. \qquad (7)$$

It can be shown (see [38]) that this definition does not depend on the choice of S and that it can be extended as follows. A (not necessarily finitely generated) group G is said to be *sofic* if all of its finitely generated subgroups are sofic.

Sofic groups were introduced by M. Gromov in [20]. The sofic terminology is due to B. Weiss [38]. The class of sofic groups contains, in particular, all residually finite groups, all amenable groups, and it is closed under direct products, free products, taking subgroups and extensions by amenable groups, and taking direct limits [12].

The following generalizes Lawton's result mentioned in Subsect. "Surjunctivity" as well as Corollary 2.

Theorem 4 (Gromov and Weiss [20,38]) *Sofic groups are surjunctive.*

We end this section by mentioning that there is no known example of a non-surjunctive group nor even of a non-sofic group up to now.

Linear Cellular Automata

Let G be a group and V be a vector space over a field $\mathbb{K}$. The configuration space $V^G = \{x\colon G \to V\}$ is a vector space over $\mathbb{K}$. Simply set $(x+y)(g) = x(g) + y(g)$ and $(\lambda x)(g) = \lambda x(g)$ for all $x, y \in V^G$, $g \in G$ and $\lambda \in \mathbb{K}$. The zero vector is the zero configuration $\mathbf{0}(g) = 0$ for all $g \in G$. The *support* of a configuration $x \in A^G$ is the subset $\mathrm{supp}(x) = \{g \in G\colon x(g) \neq 0\} \subset G$. We denote by $V[G] = \{x \in V^G\colon x \sim_{a.e.} \mathbf{0}\}$ the subspace of all configurations with *finite* support.

A linear cellular automaton is a cellular automaton $\tau\colon V^G \to V^G$ which is a linear map, that is, $\tau(x+y) = \tau(x) + \tau(y)$ and $\tau(\lambda x) = \lambda\tau(x)$ for all $x, y \in V^G$ and $\lambda \in \mathbb{K}$. This is equivalent to the linearity of the local defining map $\mu\colon V^M \to V$. We denote by $\mathrm{LCA}(G, V)$ the space of all linear cellular automata over G with coefficients in V.

Example 7 The Laplacian (cf. Example 4) is a linear cellular automaton.

Remark If the field is finite, so that necessarily $|\mathbb{K}| = p^n$ with p a prime number, and V has finite dimension over $\mathbb{K}$, then the vector space V is also finite.

It is easy to see that if $\tau \in \mathrm{LCA}(G, V)$ and $x \in V^G$ has finite support, then $\tau(x)$ also has finite support (in fact $\mathrm{supp}(\tau(x)) \subset \mathrm{supp}(x)M$). In other words $\tau(V[G]) \subset V[G]$. We denote by $\tau_0 = \tau|_{V[G]}\colon V[G] \to V[G]$ the restriction of τ to $V[G]$. We then have the following characterization of pre-injectivity for linear cellular automata.

Proposition 1 *The linear cellular automaton* $\tau \in \mathrm{LCA}(G, V)$ *is pre-injective if and only if* $\tau_0\colon V[G] \to V[G]$ *is injective.*

Note that if G is countable, V^G admits also a structure of a *metric space* (the distance function (3) is defined in the same way). Then $V[G]$ is dense in V^G with respect to the topology induced by the distance (3). However, V^G is no longer a compact space so that many topological arguments based on compactness need to be obtained with an alternative method. As an example, the following linear analogue of Corollary 2 needs an appropriate proof.

Theorem 5 ([6]) *Let G be a countable group and let V be a finite-dimensional vector space over a field $\mathbb{K}$. Suppose that $\tau\colon V^G \to V^G$ is a linear cellular automaton.*

(i) $\tau(V^G)$ is a closed subspace of V^G.
(ii) If τ is bijective then the inverse map $\tau^{-1}\colon V^G \to V^G$ is also a linear cellular automaton.

Mean Dimension and the GOE Theorem

Let G be a countable amenable group and V a finite-dimensional vector space over a field $\mathbb{K}$. Fix a Følner sequence $(F_n)_{n\in\mathbb{N}}$ for G. The *mean dimension* of a G-invariant vector subspace $X \subset V^G$, which plays the role of entropy used in the finite alphabet case, is the non-negative number

$$\mathrm{mdim}(X) = \lim_{n\to\infty} \frac{\dim(X_{F_n})}{|F_n|}\,, \tag{8}$$

where X_{F_n} is defined as in (6).

The result of Ornstein and Weiss [31] already mentioned above implies that the limit (8) exists and does not depend on the particular choice of the Følner sequence $(F_n)_{n\in\mathbb{N}}$ for G.

Note that it immediately follows from this definition that $\mathrm{mdim}(V^G) = \dim(V)$ and that $\mathrm{mdim}(X) \leq \mathrm{mdim}(Y) \leq \dim(V)$ for all G-invariant vector subspaces $X \subset Y$ of V^G.

The linear analogue of the Garden of Eden theorem for linear cellular automata states as follows.

Theorem 6 (Ceccherini-Silberstein and Coornaert, [5]) *Let V be a finite-dimensional vector space over a field $\mathbb{K}$ and let G be a countable amenable group. Let $\tau\colon V^G \to V^G$ be a linear cellular automaton. Then the following are equivalent.*

(a) τ is surjective (i. e. there are no GOE configurations);
(b) τ is pre-injective;
(c) $\mathrm{mdim}(\tau(V^G)) = \dim(V)$.

As an application we present the following example.

Example 8 Let G be a finitely generated group. Let $S \subset G$ be a finite generating subset (not necessarily symmetric) such that $1_G \notin S$, and denote by $\Delta_S\colon \mathbb{R}^G \to \mathbb{R}^G$ the corresponding Laplacian (cf. Example 4). It follows from the Maximum Principle, see also Proposition 6.4 in [5], that if the group G is infinite, then the linear cellular automaton

Δ_S is pre-injective (though not injective since the constant configurations are in the kernel of Δ_S).

Thus, as a consequence of Theorem 6, we deduce that Δ_S is also surjective if G is an infinite amenable group.

Actually, one has that Δ_S is always surjective whenever G is infinite. Indeed, denoting by $P_S = I - \Delta_S$ the *Markov operator* associated with S, one has that if G is non-amenable then G is *transient* i. e. the series $\sum_{n=0}^{\infty}(P_S)^n$ converges [39] (in fact, by a profound result of N. Varopoulos (1986, see, e. g. [35]), G is transient if and only if it has no finite index subgroup isomorphic to either $\mathbb{Z}$ or $\mathbb{Z}^2$). We denote by $\mathcal{G}_S$ the sum of this series. It is called the *Green operator* of G.

But then, for $f \in \mathbb{R}[G]$ the function $g = \mathcal{G}_S f \in \mathbb{R}^G$ clearly satisfies $\Delta_S g = (I - P_S)g = f$. This shows that $\Delta_S(\mathbb{R}^G) \supset \mathbb{R}[G]$ and, by virtue of Theorem 5 (i) and the density of $\mathbb{R}[G]$ in $\mathbb{R}^G$, one has indeed $\Delta_S(\mathbb{R}^G) = \mathbb{R}^G$, that is, Δ_S is surjective. We thank Vadim Kaimanovich and Nic Varopoulos for clarifying this point to us.

In the example below we show that the implication (b) $\Rightarrow$ (a) in Theorem 6 fails to hold, in general, for linear cellular automata over the free group of rank two. Note that if the field is finite, then this example also provides an instance of the failure of the implication (b) $\Rightarrow$ (a) in Theorem 3 when $G = F_2$.

Example 9 Let $G = F_2$ be the free group on two generators a and b. Let $\mathbb{K}$ be a field and set $V = \mathbb{K}^2$. Consider the endomorphisms p and q of V defined by $p(\alpha,\beta) = (\alpha, 0)$ and $q(\alpha,\beta) = (\beta, 0)$ for all $(\alpha,\beta) \in V$. Let $\tau\colon V^G \to V^G$ be the linear cellular automaton, with memory set $M = \{a, b, a^{-1}, b^{-1}\}$, given by

$$\tau(x)(g) = p(x(ga)) + q(x(gb)) + p(x(ga^{-1})) + q(x(gb^{-1}))$$

for all $x \in V^G$, $g \in G$. The image of τ is contained in $(\mathbb{K} \times \{0\})^G$. Therefore τ is not surjective. Let us show that τ is pre-injective. Assume that there is an element $x_0 \in V^G$ with non-empty finite support $\Omega \subset G$ such that $\tau(x_0) = 0$. Consider a vertex $g_0 \in \Omega$ at maximal distance n_0 from the identity in the Cayley graph of G. The vertex g_0 has at least three adjacent vertices at distance $n_0 + 1$ from the identity. It follows from the definition of τ that $\tau(x_0)$ does not vanish at (at least) one of these three vertices. This gives a contradiction. Thus τ is pre-injective.

The following, which is a linear version of Example 6, provides an instance of the failure of the implication (a) $\Rightarrow$ (b) in Theorem 6 when $G = F_2$.

Example 10 Let $G = F_2$ be the free group on two generators a and b. Let $\mathbb{K}$ be a field and set $V = \mathbb{K}^2$. Consider the endomorphisms p' and q' of V defined by $p'(\alpha,\beta) = (\alpha, 0)$ and $q'(\alpha,\beta) = (0,\alpha)$ for all $(\alpha,\beta) \in V$.

Let $\tau\colon V^G \to V^G$ be the $\mathbb{K}$-linear cellular automaton, with memory set $S = \{a, b, a^{-1}, b^{-1}\}$, given by

$$\tau(x)(g) = p'(x(ga)) + p'(x(ga^{-1})) + q'(x(gb)) + q'(x(gb^{-1}))$$

for all $x \in V^G$ and $g \in G$.

Consider the configuration $x_0 \in V^G$ defined by

$$x_0(g) = \begin{cases} (0,1) & \text{if } g = 1_G \\ (0,0) & \text{otherwise}\,. \end{cases}$$

Then, x_0 is almost equal to 0 and $\tau(x_0) = 0$. This shows that τ is not pre-injective (cf. Proposition 1).

However, τ is surjective. To see this, let $z = (z_1, z_2) \in \mathbb{K}^G \times \mathbb{K}^G = V^G$. Let us show that there exists $x \in V^G$ such that $\tau(x) = z$. We define $x(g)$ by induction on the graph distance, which we denote by $|g|$, of $g \in G$ from 1_G in the Cayley graph of G. We first set $x(1_G) = (0, 0)$.

Then, for $s \in S$ we set

$$x(s) = \begin{cases} (z_1(1_G), 0) & \text{if } s = a \\ (z_2(1_G), 0) & \text{if } s = b \\ (0,0) & \text{otherwise}\,. \end{cases}$$

Suppose that $x(g)$ has been defined for all $g \in G$ with $|g| \le n$, for some $n \ge 1$. For $g \in G$ with $|g| = n$, let $g' \in G$ and $s' \in S$ be the unique elements such that $|g'| = n - 1$ and $g = g's'$. Then, for $s \in S$ with $s's \neq 1_G$, we set

$$x(gs) = \begin{cases} (z_1(g) - x_1(g'), 0) & \text{if } s' \in \{a, a^{-1}\} \text{ and } s = s' \\ (z_2(g), 0) & \text{if } s' \in \{a, a^{-1}\} \text{ and } s = b \\ (z_1(g), 0) & \text{if } s' \in \{b, b^{-1}\} \text{ and } s = a \\ (z_2(g) - x_2(g'), 0) & \text{if } s' \in \{b, b^{-1}\} \text{ and } s = s' \\ (0,0) & \text{otherwise}\,. \end{cases}$$

Then one easily checks that $\tau(x) = z$. This shows that τ is surjective.

We now show that, for any group, both implications of the equivalence (a) $\Leftrightarrow$ (b) in Theorem 6 fail to hold, in general, when the vector space V is infinite-dimensional.

Example 11 Let V be an infinite-dimensional vector space over a field $\mathbb{K}$ and let G be any group. Let us choose a basis B for V. Every map $\alpha\colon B \to B$ uniquely extends to a linear map $\widetilde{\alpha}\colon V \to V$. The product map $\tau = \widetilde{\alpha}^G\colon V^G \to V^G$ is a linear cellular automaton with memory set $M = \{1_G\}$ and local defining map $\widetilde{\alpha}$. Since B

is infinite, we can find a map $\alpha\colon B \to B$ which is surjective but not injective (resp. injective but not surjective). Clearly, the associated linear cellular automaton τ is surjective but not pre-injective (resp. injective but not surjective).

We say that a group G is *L-surjunctive* if for any field $\mathbb{K}$ and for any finite dimensional vector space V over $\mathbb{K}$, every injective linear cellular automaton $\tau\colon V^G \to V^G$ is surjective.

The following is the linear analogue of the Gromov–Weiss theorem (Theorem 4).

Theorem 7 ([6]) *Sofic groups are L-surjunctive.*

Group Rings and Kaplansky Conjectures

Irving Kaplansky [23,24] posed some famous problems in the theory of group rings. Here we establish some connections between these problems and the theory of linear cellular automata.

Group Rings

Let G be a group and $\mathbb{K}$ a field. A natural basis for $\mathbb{K}[G]$, the subspace of finitely supported configurations in $\mathbb{K}^G$, is given by $\{\delta_g\colon g \in G\}$, where $\delta_g\colon G \to \mathbb{K}$ is defined by $\delta_g(g) = 1$ and $\delta_g(g') = 0$ if $g' \neq g$. Also, $\mathbb{K}[G]$ can be endowed with a ring structure by defining the *convolution* product xy of two elements $x, y \in \mathbb{K}[G]$ by setting, for all $g \in G$,

$$[xy](g) = \sum_{h\in G} x(h)y(h^{-1}g). \tag{9}$$

One has $\delta_g\delta_h = \delta_{gh}$ and $\delta_h x = x^{h^{-1}}$ for all $g, h \in G$ and $x \in \mathbb{K}[G]$, in particular, δ_{1_G} is the unit element in $\mathbb{K}[G]$. The ring $\mathbb{K}[G]$ is called the *group ring* of G with coefficients in $\mathbb{K}$. Note that the product map $(x, y) \to xy$ is $\mathbb{K}$-linear so that $\mathbb{K}[G]$ is in fact a $\mathbb{K}$-algebra.

Note also that (9) makes sense for $x, y \in K^G$ and at least one is finitely supported. Moreover, the group G, via the map $G \ni g \mapsto \delta_g \in \mathbb{K}[G]$, can be identified with a subgroup of the group of invertible elements in $\mathbb{K}[G]$. This way, every element x of $\mathbb{K}[G]$ can be uniquely expressed as $x = \sum_{g\in G} x(g)g$.

The Matrix Representation of Linear Cellular Automata

Let $d \geq 1$ be an integer. Denote by $\mathrm{Mat}_d(\mathbb{K}[G])$ the $\mathbb{K}$-algebra of $d \times d$ matrices with coefficients in $\mathbb{K}[G]$. For $x = (x_1, x_2, \dots, x_d) \in (\mathbb{K}^G)^d$ and $\alpha = (\alpha_{ij})_{i,j=1}^d \in \mathrm{Mat}_d(\mathbb{K}[G])$, we define $x\alpha = (y_1, y_2, \dots, y_d) \in (\mathbb{K}^G)^d$ by setting $y_j = \sum_{i=1}^d x_i\alpha_{ij}$ for all $j = 1, 2, \dots, d$, where $x_i\alpha_{ij}$ is the convolution product of $x_i \in \mathbb{K}^G$ and $\alpha_{ij} \in \mathbb{K}[G]$ defined using (1).

The map $\mathrm{Mat}_d(\mathbb{K}[G]) \ni \alpha \mapsto \overline{\alpha} \in \mathrm{Mat}_d(\mathbb{K}[G])$, where $\overline{\alpha}_{ij}(g) = \alpha_{ji}(g^{-1})$ for all $g \in G$ and $i, j = 1, 2, \dots, d$ is an *anti-involution* of the algebra $\mathrm{Mat}_d(\mathbb{K}[G])$, since $\overline{\overline{\alpha}} = \alpha$ and $\overline{\alpha\beta} = \overline{\beta}\overline{\alpha}$, for all $\alpha, \beta \in \mathrm{Mat}_d(\mathbb{K}[G])$.

Let $\alpha \in \mathrm{Mat}_d(\mathbb{K}[G])$ and define the map $\tau_\alpha : (\mathbb{K}^d)^G \to (\mathbb{K}^d)^G$ by setting

$$\tau_\alpha(x) = x\overline{\alpha}$$

for all $x = (x_1, x_2, \dots, x_d) \in (\mathbb{K}^G)^d = (\mathbb{K}^d)^G$.

Theorem 8 ([6]) *For $\alpha \in \mathrm{Mat}_d(\mathbb{K}[G])$ the map $\tau_\alpha\colon (K^d)^G \to (\mathbb{K}^d)^G$ is a linear cellular automaton. Moreover, the map $\mathrm{Mat}_d(\mathbb{K}[G]) \ni \alpha \to \tau_\alpha \in \mathrm{LCA}(G, \mathbb{K}^d)$ is an isomorphism of $\mathbb{K}$-algebras.*

Remark When $G = \mathbb{Z}$ and $\mathbb{K}$ is a finite field, linear cellular automata $\tau_\alpha \in \mathrm{LCA}(\mathbb{Z}, \mathbb{K}^d)$ with $\alpha \in \mathrm{Mat}_d(\mathbb{K}[\mathbb{Z}])$, are called *convolutional encoders* in Sect. 1.6 of [26].

Zero-Divisors in Group Rings

Let R be a ring. A non zero element $a \in R$ is said to be a *left* (resp. *right*) *zero-divisor* provided there exists a non zero element $b \in R$ such that $ab = 0$ (resp. $ba = 0$).

The following result relates the notion of a zero-divisor in a group ring $\mathbb{K}[G]$ with the pre-injectivity of one-dimensional linear cellular automata over the group G. We use the same notation as in Theorem 8 (here $d = 1$).

Lemma 1 ([5]) *Let G be a group and let $\mathbb{K}$ be a field. An element $\alpha \in \mathbb{K}[G]$ is a left zero-divisor if and only if the linear cellular automaton $\tau_\alpha\colon \mathbb{K}^G \to \mathbb{K}^G$ is not pre-injective.*

Let G be a group and suppose that it contains an element g_0 of finite order $n \geq 2$. Then we have

$$(1_G - g_0)\left(1_G + g_0 + g_0^2 + \dots + g_0^{n-1}\right) = 0\,,$$

showing that $\mathbb{K}[G]$ has zero-divisors.

A group is *torsion-free* if it has no non-trivial element of finite order. *Kaplansky zero-divisor problem* [23] asks whether $\mathbb{K}[G]$ has no zero-divisors whenever G is torsion-free. In virtue of Lemma 1 and Theorem 8 we can state it as follows (see [5]).

Problem 1 (Kaplansky zero-divisor problem reformulated in terms of cellular automata) *Let G be a torsion-free group and let $\mathbb{K}$ be a field. Is it true that every non-identically-zero linear cellular automaton $\tau\colon \mathbb{K}^G \to \mathbb{K}^G$ is pre-injective?*

The zero-divisor problem is known to have an affirmative answer for a wide class of groups including the free groups, the free Abelian groups, the fundamental groups of surfaces and the braid groups B_n.

Combining Lemma 1 with Theorem 6 we deduce the following.

Corollary 3 *Let G be a countable amenable group and let $\mathbb{K}$ be a field. Suppose that $\mathbb{K}[G]$ has no zero-divisors. Then every non-identically-zero linear cellular automaton $\tau\colon \mathbb{K}^G \to \mathbb{K}^G$ is surjective.*

The class of *elementary amenable* groups is the smallest class of groups containing all finite and all Abelian groups that is closed under taking extensions and directed unions. It is known, see Theorem 1.4 in [25], that if G is a torsion-free elementary amenable group, then $\mathbb{K}[G]$ has no zero-divisor for any field $\mathbb{K}$. As a consequence, the conclusion of Corollary 3 holds for all torsion-free elementary amenable groups.

Stable Finiteness of Group Rings

Recall that a ring R with identity element 1_R is said to be *directly finite* if one-sided inverses in R are also two-sided inverses, i. e., $ab = 1_R$ implies $ba = 1_R$ for $a, b \in R$. The ring R is said to be *stably finite* if the ring $M_d(R)$ of $d \times d$ matrices with coefficients in R is directly finite for all integers $d \geq 1$.

Commutative rings and finite rings are obviously directly finite. Also observe that if elements a and b of a ring R satisfy $ab = 1_R$ then $(ba)^2 = ba$, that is, ba is an idempotent. Therefore if the only idempotent of R are 0_R and 1_R (e. g. if R has no zero-divisors) then R is directly finite. The ring of endomorphisms of an infinite-dimensional vector space yields an example of a ring which is not directly finite.

Kaplansky [24] observed that, for any group G and any field $\mathbb{K}$ of characteristic 0, the group ring $\mathbb{K}[G]$ is stably finite and asked whether this property remains true for fields of characteristic $p > 0$. We have that this holds for L-surjunctive groups.

Using the matrix representation of linear cellular automata (Theorem 8) one has the following characterization of L-surjunctivity.

Theorem 9 *For a group G, a field $\mathbb{K}$, and an integer $d \geq 1$ the following conditions are equivalent:*

(a) Every injective linear cellular automaton $\tau\colon (\mathbb{K}^d)^G \to (\mathbb{K}^d)^G$ is surjective;
(b) The ring $\mathrm{Mat}_d(\mathbb{K}[G])$ is directly finite.

As a consequence, a group G is L-surjunctive if and only if the group ring $\mathbb{K}[G]$ is stably finite for any field $\mathbb{K}$.

From Theorem 7 and Theorem 9 we deduce the following result previously established by G. Elek and A. Szabó using different methods.

Corollary 4 ([6,11]) *Let G be a sofic group and $\mathbb{K}$ any field. Then the group ring $\mathbb{K}[G]$ is stably finite.*

Future Directions

We indicate some open problems related to the topics that we have treated in this article.

Garden of Eden Theorem

As we mentioned in the Subsect. "Amenable Groups", it would be interesting to determine whether the Myhill theorem (i. e. the implication (b) ⇒ (a) in Theorem 3), which holds for amenable groups, but fails to hold for groups containing the free group F_2 (cf. Example 9), holds or not for the non-amenable groups with no free subgroups (such as the free Burnside groups $B(m, n)$, $m \geq 665$ odd, see [1]). Note that a negative answer would give another new characterization of amenability for groups.

Problem 2 *Determine whether the Myhill theorem (pre-injectivity implies surjectivity) for cellular automata with finite alphabet holds only for amenable groups.*

It turns out that Bartholdi's cellular automata ([2], see Subsect. "Amenable Groups") are not linear, so that the question whether the linear GOE theorem (Theorem 6) holds also for non-amenable groups remains an open problem.

Problem 3 *Determine whether the GOE theorem for linear cellular automata over finite-dimensional vector spaces holds only for amenable groups or not. More precisely, determine whether Moore's theorem and/or Myhill's theorem for linear cellular automata over finite dimensional vector spaces hold only for amenable groups or not.*

In [7] we generalized the GOE theorem to linear cellular automata over semisimple modules (over a, not necessarily commutative, ring R) of finite length with universe an amenable group. A vector space is a (semisimple) left module over a field. The finite length condition for modules (which corresponds to the ascending chain condition (Noetherianity) and to the descending chain condition (Artinianity)) is the natural analogue of the notion of finite dimension for vector spaces.

The Garden of Eden theorem can be also generalized by looking at subshifts. Given an alphabet A and a countable group G a *subshift* $X \subset A^G$ is a subset which is closed (in the topology induced by the metric (3)) and G-invariant (if $x \in X$ then $x^g \in X$ for all $g \in G$). If G is amenable and $X \subset A^G$ is a subshift the quantity (5) is the entropy of X. Given two subshifts $X, Y \subset A^G$ we define a cellular automaton $\tau \colon X \to Y$ as the restriction to X of a cellular automaton $\overline{\tau} \colon A^G \to A^G$ such that $\overline{\tau}(X) \subset Y$.

Problem 4 *Let G be a countable amenable group, A a finite alphabet and $X, Y \subset A^G$ two subshifts with* $\mathrm{ent}(X) = \mathrm{ent}(Y)$. *Prove, under suitable conditions, the GOE theorem for cellular automata $\tau \colon X \to Y$.*

We mention that the GOE theorem for subshifts over amenable groups fails to hold in general with no additional hypotheses on the subshifts.

Let $G = \mathbb{Z}$ be the infinite cyclic group and let A be a finite alphabet. For $n, m \in \mathbb{Z}$ we set $[n, m] = \{x \in \mathbb{Z} \colon n \leq x \leq m\}$. Also, we denote by $A^* = \{a_1 a_2 \cdots a_n \colon a_i \in A, 1 \leq i \leq n, n \in \mathbb{N}\}$ the set of all *words* over A. The *length* of a word $w = a_1 a_2 \cdots a_n \in A^*$ is $\ell(w) = n$. Given a subset $X \subset A^G$ we denote by $W(X) = \{w \in A^* \colon \exists x \in X, \text{ s.t. } w = x|_{[0,\ell(w)]}\}$ the *language* associated with X. Then one says that a subshift $X \subset A^{\mathbb{Z}}$ is irreducible if, for any two subwords $w_1, w_2 \in W(X)$ there exists $w_3 \in A^*$ (necessarily in $W(X)$) such that $w_1 w_3 w_2 \in W(X)$.

Also it can be shown that, for a subshift $X \subset A^{\mathbb{Z}}$, there exists a set $\mathcal{F} \subset A^*$ of so-called *forbidden words*, such that setting

$$X_{\mathcal{F}} := \{x \in A^{\mathbb{Z}} \colon x^g|_{[0,n]} \notin \mathcal{F}, \forall n \in \mathbb{N}, \forall g \in \mathbb{Z}\} ,$$

one has $X = X_{\mathcal{F}}$. Then a subshift $X \subset A^{\mathbb{Z}}$ is of *finite type* if $X = X_{\mathcal{F}}$ for some *finite* set $\mathcal{F} \subset A^*$. Finally, X is *sofic* (same etymology but a different meaning from that used for groups in Subsect. "Sofic Groups") if $X = \tau(Y)$ for some cellular automaton $\tau \colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ and some subshift $Y \subset A^{\mathbb{Z}}$ of finite type.

In [13], F. Fiorenzi considered cellular automata on irreducible subshifts of finite type inside $A^{\mathbb{Z}}$(A finite). She proved the GOE theorem for such cellular automata and provided examples of cellular automata on subshifts of finite type which are not irreducible and for which both implications of the GOE theorem fail to hold. She also provided an example of a cellular automaton on an irreducible subshift which is sofic but not of finite type for which the Moore theorem (namely the implication (a) ⇒ (b) in Theorem 3) fails to hold. Note that it is well known (cf. 2.1 in [13]) that, under the same hypotheses, the Myhill theorem (namely the implication (b) ⇒ (a) in Theorem 3) always holds true ($G = \mathbb{Z}$).

More generally, for groups G other than the integers $\mathbb{Z}$, one needs appropriate notions of *irreducibility* for the subsets $X \subset A^G$ as investigated by Fiorenzi [14,15].

Surjunctivity

Problem 5 *Prove or disprove that all groups are surjunctive.*

A positive answer to the previous problem could be derived by positively answering to the following

Problem 6 *Prove or disprove that all groups are sofic.*

By considering the linear analogue of Problem 5 we have

Problem 7 (Kaplansky's conjecture on stable finiteness of group rings) *Prove or disprove that all groups are L-surjunctive. Equivalently (cf. Theorem 9), prove or disprove that the group ring $\mathbb{K}[G]$ is stably finite for any group G and any field $\mathbb{K}$.*

Also, one could look for surjunctivity results for cellular automata with alphabets other than the finite ones and the finite dimensional vector spaces. A ring is called *left Artinian* (see [40]) if it satisfies the *descending condition on left ideals*, namely every decreasing sequence $R \supset I_1 \supset I_2 \cdots \supset I_n \supset I_{n+1} \supset \cdots$ of left ideals eventually stabilizes (there exists n_0 such that $I_n = I_{n_0}$ for all $n \geq n_0$).

In [4] we showed that if G is a residually finite group and M is an Artinian left module over a ring R (e. g. if M is a finitely generated left module over a left Artinian ring R), then every injective R-linear cellular automaton $\tau \colon M^G \to M^G$ is surjective.

In [8] we showed that if G is a sofic group (thus a weaker condition than being residually finite) and M is a left module of finite length over a ring R (thus a stronger condition than being just Artinian), then every injective R-linear cellular automaton $\tau \colon M^G \to M^G$ is surjective. As a consequence (cf. Theorem 9) we have that the group ring $R[G]$ is stably finite for any left (or right) Artinian ring R and any sofic group G.

It is therefore natural to consider the following generalization of Problem 7.

Problem 8 *Prove or disprove that the group ring $R[G]$ is stably finite for any group G and any left (or right) Artinian ring R.*

Problem 9 (Kaplansky's zero divisor conjecture for group rings) *Prove or disprove that if G is torsion-free then any non-identically zero linear cellular automaton $\tau \colon \mathbb{K}^G \to \mathbb{K}^G$, where $\mathbb{K}$ is a field, is preinjective. Equivalently (cf. Lemma 1 and Theorem 8), prove or disprove that if G is torsion-free, the group ring $\mathbb{K}[G]$ has no zero divisors.*

Bibliography

1. Adyan SI (1983) Random walks on free periodic groups. Math USSR Izvestiya 21:425–434

2. Bartholdi L (2008) A converse to Moore's and Hedlund's theorems on cellular automata. J Eur Math Soc (to appear). Preprint arXiv:0709.4280
3. Berlekamp ER, Conway JH, Guy RK (1982) Winning ways for your mathematical plays, vol 2, Chap. 25. Academic Press, London
4. Ceccherini-Silberstein T, Coornaert M (2007) On the surjunctivity of Artinian linear cellular automata over residually finite groups. In: Geometric Group Theory. Trends in Mathematics. Birkhäuser, Basel, pp 37–44
5. Ceccherini-Silberstein T, Coornaert M (2006) The garden of eden theorem for linear cellular automata. Ergod Th Dynam Syst 26:53–68
6. Ceccherini-Silberstein T, Coornaert M (2007) Injective linear cellular automata and sofic groups. Israel J Math 161:1–15
7. Ceccherini-Silberstein T, Coornaert M (2008) Amenability and linear cellular automata over semisimple modules of finite length. Comm Algebra 36:1320–1335
8. Ceccherini-Silberstein T, Coornaert M (2007) Linear cellular automata over modules of finite length and stable finiteness of group rings. J Algebra 317:743–758
9. Ceccherini-Silberstein TG, Machì A, Scarabotti F (1999) Amenable groups and cellular automata. Ann Inst Fourier 49:673–685
10. Ceccherini-Silberstein TG, Fiorenzi F, Scarabotti F (2004) The garden of eden theorem for cellular automata and for symbolic dynamical systems. In: Random walks and geometry (Vienna 2001), pp 73–108. de Gruyter, Berlin
11. Elek G, Szabó A (2004) Sofic groups and direct finiteness. J Algebra 280:426–434
12. Elek G, Szabó A (2006) On sofic groups. J Group Theor 9: 161–171
13. Fiorenzi F (2000) The Garden of eden theorem for sofic shifts. Pure Math Appl 11(3):471–484
14. Fiorenzi F (2003) Cellular automata and strongly irreducible shifts of finite type. Theor Comput Sci 299(1–3):477–493
15. Fiorenzi F (2004) Semistrongly irreducible shifts. Adv Appl Math 32(3):421–438
16. Følner E (1955) On groups with full Banach mean value. Math Scand 3:245–254
17. Gottschalk W (1973) Some general dynamical systems. In: Recent advances in topological dynamics. Lecture Notes in Math, vol 318. Springer, Berlin, pp 120–125
18. Gottschalk WH, Hedlund GA (1955) Topological dynamics. In: American Mathematical Society Colloquium Publications, vol 36. American Mathematical Society, Providence
19. Greenleaf FP (1969) Invariant means on topological groups and their applications. Van Nostrand, New York
20. Gromov M (1999) Endomorphisms of symbolic algebraic varieties. J Eur Math Soc 1:109–197
21. Ginosar Y, Holzman R (2000) The majority action on infinite graphs: Strings and puppets. Discret Math 215:59–71
22. Hungerford TW (1987) Algebra, graduate texts in mathematics. Springer, New York
23. Kaplansky I (1957) Problems in the theory of rings. Report of a conference on linear algebras, June, 1956, pp 1–3. National Academy of Sciences-National Research Council, Washington, Publ. 502
24. Kaplansky I (1969) Fields and rings. Chicago Lectures in Math. Univ. of Chicago Press, Chicago
25. Kropoholler PH, Linnell PA, Moody JA (1988) Applications of a new K-theoretic theorem to soluble group rings. Proc Amer Math Soc 104:675–684
26. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
27. Machì A, Mignosi F (1993) Garden of eden configurations for cellular automata on Cayley graphs of groups. SIAM J Discret Math 6:44–56
28. Moore EF (1963) Machine models of self-reproduction. Proc Symp Appl Math AMS 14:17–34
29. Myhill J (1963) The converse of Moore's garden of eden theorem. Proc Amer Math Soc 14:685–686
30. Ol'shanskii AY (1980) On the question of the existence of an invariant mean on a group. Uspekhi Mat Nauk 35(214)4: 199–200
31. Ornstein DS, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. J Anal Math 48:1–141
32. Passman DS (1985) The algebraic structure of group rings. Reprint of the 1977 original. Robert E. Krieger Publishing, Melbourne, FL
33. Paterson A (1988) Amenability, AMS mathematical surveys and monographs, vol 29. American Mathematical Society, Providence
34. Ulam S (1952) Processes and Transformations. Proc Int Cong Math 2:264–275
35. Varopoulos NT, Saloff-Coste L, Coulhon T (1992) Analysis and geometry on groups. Cambridge University Press, Cambridge
36. von Neumann J (1930) Zur allgemeinen Theorie des Maßes. Fun Math 13:73–116
37. von Neumann J (1966) In: Burks A (ed) The theory of self-reproducing automata. University of Illinois Press, Urbana, London
38. Weiss B (2000) Sofic groups and dynamical systems, (Ergodic theory and harmonic analysis, Mumbai, 1999). Sankhya Ser A 62:350–359
39. Woess W (2000) Random walks on infinite graphs and groups, Cambridge Tracts in Mathematics 138. Cambridge University Press, Cambridge
40. Zariski O, Samuel P (1975) Commutative algebra. vol 1 With the cooperation of IS Cohen. Corrected reprinting of the 1958 edition. Graduate Texts in Mathematics, No. 28. Springer, New York

Cellular Automata in Hyperbolic Spaces

MAURICE MARGENSTERN
Université Paul Verlaine, Metz, France

Article Outline

Glossary

Fibonacci sequence A sequence of natural integers, denoted by f_n and defined by the recurrent equation $f_{n+2} = f_{n+1} + f_n$, for all $n \in \mathbf{N}$, and by the initial values $f_0 = f_1 = 1$.

Hyperbolic geometry This geometry was discovered independently by both Nikolaj Lobachevsky and Jànos Bolyai around 1830. This geometry satisfies the axioms of Euclidean geometry, the axiom of parallels being excepted and replaced by the following one: through a point out of a line, there are exactly two parallels to the line. In this geometry, there are also lines which never meet: they are called **non-secant**. They are characterized by the existence, for any couple of such lines, of a unique common perpendicular. Also, in this geometry, the sum of the interior angles of a triangle is always less than π. The difference to π defines the area of the triangle. In hyperbolic geometry, distances are absolute: there is no notion of similarity. See also **Poincaré's disc**.

Pentagrid The tiling $\{5, 4\}$, with five sides and four tiles around a vertex. The angles are right angles.

Poincaré's disc A model of the hyperbolic plane inside the Euclidean plane. The points are the points which are interior to a fixed disc D. The lines are the trace in D of diameters or circles which are orthogonal to the border of D. The model was first found by Beltrami and then by Poincaré who also devised the half-plane model also called after his name. The half-plane model is a conformal image of the disc model.

Ternary heptagrid The tiling $\{7, 3\}$ with seven sides and three tiles around a vertex.

Tessellation A particular case of a finitely generated tiling. It is defined by a polygon and by its reflections in its sides and, recursively, of the images in their sides.

Tiling A partition of a geometric space; the closure of the elements of the partition are called the **tiles**. An important case is constituted by **finitely generated** tilings: there is a finite set of tiles G such that any tile is a copy of an element of G.

Tiling $\{p, q\}$ This is tessellation based on the regular polygon with p sides and with vertex angle $2\pi/q$.

Invariant group of a tiling A group of transformations which define a **bijection** on the set of tiles. Usually, in a geometrical space, they are required to belong to the group of isometries of the space.

Definition of the Subject

Cellular Automata in Hyperbolic Spaces, in short **hyperbolic cellular automata** (abbreviated HCA), consists in the implementation of Cellular Automata in the environment of a regular tiling of a hyperbolic space.

The first implementation of such an object appeared in a paper by the present author and Kenichi Morita in 1999, see [18]. In this first paper, a first solution to the location problem in this context was given. The paper also focused on one advantage of this implementation: it allows to solve **NP**-problems in polynomial time. In 2000, a second paper appeared, by the present author, where a decisive solution of the location problem was given.

The study of HCA's is a new domain in computer science, at the border of mathematics and physics. They involve hyperbolic geometry as well as elementary arithmetics and algebra with the connections of polynomials with matrices, and also some theory of fields. They also involve the theory of formal languages in connection with their properties of elementary arithmetics.

To be a melting pot of such different techniques is already something which is very interesting.

But the new field has very striking properties. Their complexity classes offer a very different landscape than that of the classical theory of complexity based on the Turing machine. They also provide a bridge between the classical theory of computation and super-Turing computations.

HCA's are a novel object with rich properties: they inherit the richness of the infinitely many regular tilings which live in the hyperbolic plane. We are at the beginning of the study and still, there are a lot of surprising results. HCA's might appear as successful as their Euclidean relatives in various domains as astrophysics, nuclear physics and computer science.

For many results indicated in this paper, we quote the book [15], where the results and its proof can be found there.

Introduction

Before the appearance of HCA's, there were a few papers on possible implementations of cellular automata in abstract contexts, especially in the case of Caley graphs, see [25]. However, as infinitely many tilings of the hyperbolic plane are not Caley graphs of a their invariant group, this method cannot solve the problem in full generality. The difficulty was the location of the tiles, the **locating**

problem. The problem is already difficult in the simple case of tessellations. Note that there are infinitely many of them in the hyperbolic plane.

The study appeared to be possible thanks to a partial solution to the locating problem, see [18,19]. A decisive step was done in [7], where the already mentioned mixing of various techniques appear. This first solution in the case of a particular tiling, the pentagrid, is dealt with in Sect. "The Locating Problem in Hyperbolic Tilings". A significant advance was performed at the occasion of the meetings organized in 2002 for the bicentury of the birth of Jànos Bolyai, the co-inventor with Nikolaj Lobachevsky of hyperbolic geometry, and also at the occasion of SCI'2002. At this conference, seven papers were presented on the topic of this article, and they had a strong impact on the later development.

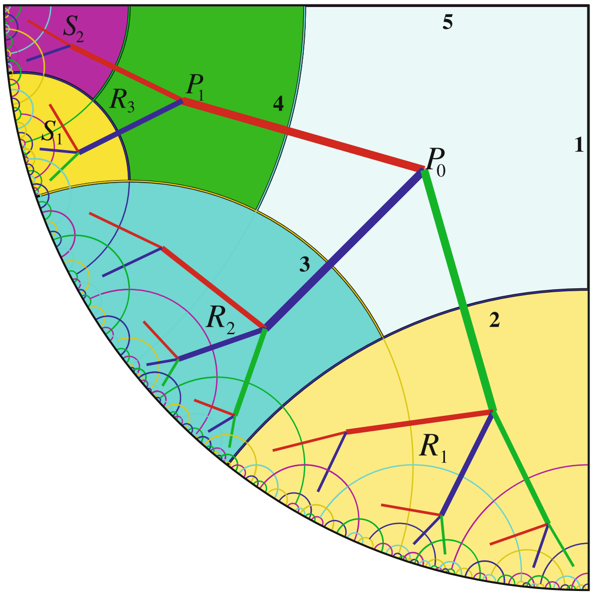
This introduction should contain a paragraph on hyperbolic geometry. If the reader is not familiar with this geometry and has some time, we recommend to him/her the first chapter of [15], or any other book introducing hyperbolic geometry. For a reader which is not familiar and who has no time, we recommend the following solution. First, forget everything of Euclidean geometry and try to remember the few elements given in the glossary. Don't worry, the Euclidean objects will always be the first thing to come to your mind and, most often, they will be misleading. Second, never forget that in traveling over hyperbolic spaces, you are in the situation of a pilot of a plane flying with instruments only. You can see nothing in the usual sense of these words and, sorry to repeat it again, the usual intuition is misleading. The best introduction is to imagine that when you venture into the hyperbolic plane, always keep with you the Ariadne thread of the way backwards. Otherwise you will definitely be lost. With this precaution, you will never regret the trip. The landscape changes very quickly and you are always fascinated by its unbelievable beauty.

The Locating Problem in Hyperbolic Tilings

The Classical Case of the Pentagrid

The method introduced in [18] consists of constructing a bijection between the tiling and a tree, the **spanning tree** of the tiling. The tree is constructed in a recursive way, defined as follows, also see Fig. 1.

> *Initial step*: P_0 is the root of the tree; it is called the **leading pentagon** of the quarter $\mathcal{Q}_0$ it is defined by its sides 1 and 5;
>
> *Induction step*: Let P be the current pentagon;
> if P is the leading pentagon of a quarter $\mathcal{Q}$, see P_0 in Fig. 1, the complement of P in $\mathcal{Q}$ splits into two quarters, $\mathcal{R}_1$ and $\mathcal{R}_3$ and a remaining region, $\mathcal{R}_3$ which we call a **strip**;
> if P is the leading pentagon of a strip S, see P_1 in Fig. 1, the complement of P in S splits into a quarter S_1 and again a strip, S_2.

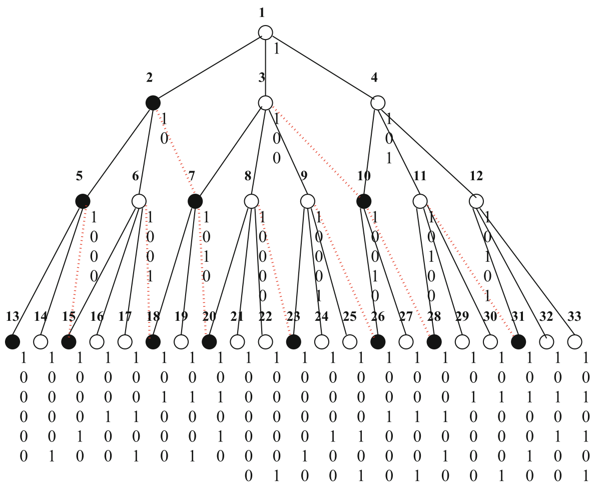


Cellular Automata in Hyperbolic Spaces, Figure 1
The pentagrid: regular pentagons with vertex angle $\pi/2$

As proved in [15], the set of tiles attached to the tree generated in this way, the **leading pentagons** of the above algorithm, is exactly the set of pentagons contained in the quarter $\mathcal{Q}_0$.

With [7,15], a new ingredient is brought in: number the nodes of the tree from the root, to which we attach 1, and then level by level, from left to right on each level, see Fig. 2. As already noticed in [18], the number of nodes of the tree which spans the tiling of a quarter which are on the same level n is f_{2n+1}, where $\{f_n\}_{n\in N}$ with $f(0) = f(1) = 1$. For this reason, the spanning tree of the pentagrid is called the **standard Fibonacci tree**, illustrated by Fig. 2.

The above splitting induces a particular structure on the standard Fibonacci tree. Define **white nodes** as nodes which have three sons and **black nodes** as nodes which have two sons. Black and white are the two possible values of the **status** of a node. Then, there is a **rule** to define the status of the sons of a node. We can write them as follows, in self-explained notations:



Cellular Automata in Hyperbolic Spaces, Figure 2
The Fibonacci tree

$$
\begin{aligned}
W &\longrightarrow BWW\\
B &\longrightarrow BW
\end{aligned}
$$

Now, let us represent the numbers attached to the nodes of the Fibonacci tree in the numeration basis defined by the Fibonacci sequence itself, starting from f_1. The representation is not unique. Choose the longest representation with respect to the lexicographic order and call it the **coordinate** of the node to which the corresponding number is attached.

First, we have that the set of coordinates is a **regular language**, which is a corollary of a well known theorem, see [2]. Now we have a more interesting property, which was noticed in [7] and which we call the **preferred son property**. Let $\alpha_k..\alpha_0$ be the coordinate of a node ν of the standard Fibonacci tree, with α_0 as the lightest digit of the representation. The property says that for each node ν of the standard Fibonacci tree, with coordinate $\alpha_k..\alpha_0$, there is exactly one son of ν whose coordinate is $\alpha_k..\alpha_0 00$. This son is called **preferred**. Moreover, there is a rule to find out the preferred son from the status of a node: in a black node, the preferred son is the black son; in a white node, the preferred son is the middle one.

Generalization: The Splitting Method

The generalization was first announced in [8]. It was then presented in [9], with a new visit to Poincaré's theorem, at the occasion of the second century of the birth of Jànos Bolyai.

The method defines a **basis of splitting** and then, the notion of a **combinatoric tiling**. Two important consequences can be derived from these very definitions to which we turn now.

Let $S_0, \dots, S_k$ be finitely many parts of some geometric metric space X which are supposed to be closed, with nonempty interior, unbounded and simply connected. Consider also finitely many closed simply connected bounded sets $P_1, \dots, P_h$ with $h \leq k$. Say that the S_i's and P_ℓ's constitute a **basis of splitting** *if and only if: (i) X splits into finitely many copies of S_0, (ii) any S_i splits into one copy of some P_ℓ, the* **leading tile** *of S_i, and finitely many copies of S_j's, where* **copy** *means an* **isometric image**, *and where, in the condition (ii), the copies may be of different S_j's, S_i possibly included. As usual, it is assumed that the interiors of the copies of P_ℓ and of the copies of the S_j's are pairwise disjoint. The set S_0 is called the* **head** *of the basis and the P_ℓ's are called the* **generating tiles** *and the S_i's are called the* **regions** *of the splitting.*

On the example of the pentagrid, a basis of splitting is given by a quarter $\mathcal{Q}$ and a strip S. When there is a basis of a splitting, we then define:

Say that a tiling of X is **combinatoric** *if X has a basis of splitting and if the spanning tree of the splitting yields exactly the restriction of the tiling to the head S_0 of the basis.*

In [15,18], the pentagrid is proven to be combinatoric. A lot of other tilings of the hyperbolic plane are combinatoric. In particular, all the tilings $\{p, q\}$, with $q \geq 4$, possess this property. In higher dimensions, the following tilings were proved combinatoric: the 3D tiling $\{5, 3, 4\}$, based on Poincaré's dodecahedron, see [15,21] and the 4D tiling $\{5, 3, 3, 4\}$, based on the 120-cell, see [11,15].

Once a tiling is combinatoric, from the definition of its basis of splitting, we can derive a square matrix M called the **matrix of the splitting**, see [8,15]. Its lines indicate, for each column, the number of copies of S_j's entering in the splitting of S_i. The **polynomial of the splitting** is the characteristic polynomial of M, divided by the greatest power of X it contains as a factor. In our cases, this polynomial has a greatest real root. The polynomial of the splitting induces a recurrent equation which defines the **sequence of the splitting** with appropriate initial values. The maximal representations of numbers in the basis defined by the sequence of the splitting constitute the **language of the splitting**. As proved in [15,20], the language of the splitting of the tilings $\{p, q\}$ is regular when $q \geq 4$ and $p \geq 4$.

Implementation of Cellular Automata in Hyperbolic Spaces

The implementation of HCA's is induced by the results mentioned in the previous section.

But first, let us go back to the general definition of CA's. Three conditions must be fulfilled by a set of cells to be called a cellular automaton. The cells of the automaton must uniformly be distributed in the considered space. The neighborhood of each cell is defined in a uniform way. At each top of the discrete clock, all the cells update their own state according to the same function applied to the state of the cell and the sequence of states of its neighbors.

To implement cellular automata, we have to satisfy these three requirements.

The first two conditions are easily satisfied in a tessellation. Note that this is the standard frame for CA's in the Euclidean plane and in the 3D Euclidean space.

The third condition already requires that we have a system of coordinates for the tiles at our disposal. More than three centuries after Descartes' discovery of the system of coordinates which everybody uses for the Euclidean plane, this condition is trivially fulfilled. This is not only the three-century usage. This is also the case because the mathematical structure of the group of displacements which leaves the considered tessellations of the Euclidean plane globally invariant is a very simple structure.

The situation is very different in the case of hyperbolic spaces. Before [7], there was no convenient, or at least fast, procedure to define the coordinates of the tiles in a way which is in connection with the geometrical properties of the tiling.

Now, the splitting method gives such a solution. First, it effectively exhibits a tree which generates the tiling. As Gormov pointed out, [3], hyperbolic spaces are characterized by a tree structure. Second, it provides **fast** algorithms to handle these coordinates. By fast, we mean that the basic algorithms we need are **linear** in time with respect to the coordinate of the initial point. Note that nobody really matters with the fact that addition of vectors in Euclidean coordinates is linear while multiplication of coordinates by a scalar is not. Here, we have no addition, no multiplication, no nice formula. We have algorithms only, but they turn out to work in the best time.

The result of these considerations is that the directions, **north**, **south**, **east** and **west** which play a nice role in the Euclidean case no longer exist. In fact, we have infinitely many directions, each of which defines an essential direction in the space: if you follow other directions, you will never go to the area covered by this one. Of course, an infinite amount of information is ruled out in computer science. And so, we replace this basic indetermination of the direction by the direction of the **father**. Of course, we are led to a root and a central cell, but nobody complains about using an origin in the Euclidean case. Moreover, as shown in [12,15], it is also possible in the case of tessellations of the hyperbolic plane to get rid of the origin. We just mention this point, here, and refer the interested reader to the quoted papers for a closer study.

Once again, we illustrate how we proceed by the case of the pentagrid. It is repeated in the case of the ternary heptagrid, see [12,14] and in the case of the 3D tiling of Poincaré's dodecahedron, see [13].

For the implementation, we first fix a basis of splitting and the representation of the tiling. As indicated in [7,15], there are a lot of choices with the same basis of splitting. Moreover, in the case of the pentagrid and of the ternary heptagrid in which the standard Fibonacci tree is also a spanning tree, we have the choice between using the Fibonacci sequence, as we did in Sect. "The Locating Problem in Hyperbolic Tilings", or using the basis derived from the polynomial of the splitting. The difference is that the Fibonacci sequence is defined by the golden mean $(1+\sqrt{5})/2$, while the sequence of the splitting is defined by the square of the golden mean, $(3+\sqrt{5})/2$.

Let us go on with the Fibonacci sequence, as is it used in the majority of papers.

The preferred son property allows us to compute very easily the coordinates of the neighbors of a cell ν from the coordinate c of ν: the computation is linear in time with respect to the length of c, see [10,15]. Similarly, the path from a cell c to the root of its tree can be computed in a linear time with respect to the length of c, again see [10,15]. As a simple example, if $c = \alpha_k \ldots \alpha_1\alpha_0$, the father of ν has for number $\mathcal{A}(\alpha_k..\alpha_2) + \alpha_1$, where $\mathcal{A}(\beta_h \ldots \beta_0)$ computes the number m whose coordinate is $\beta_{\mathrm{h}} \ldots \beta_0$.

What we indicated up to now fixes the coordinates for a cell whose supporting tile is in a given quarter. Now, it is enough to number the five initial quarters which lie around the central pentagon in order to completely define the coordinates of a cell. The central pentagon has 0 as an unique coordinate. All other cells are defined by two numbers: (α, ν). The first number, α is in $\{1..5\}$ and defines the quarter. The second number, ν, defines the tile in the indicated quarter. Together with its coordinate, a cell is associated with other data: the status of its supporting tile, and the indication of which side is shared with its father. On one hand, note that the coordinate is a hardware feature: it is never known by the cell and it cannot: it does not have a bounded size. Note that this is the same for CA's in Euclidean spaces. On the other hand, the status of the supporting node can be known by the cell. As shown in [10], one can define rules for a cellular automaton to dispatch this information. As it is a finite information which can be provided by the hardware, we may assume that the cell knows it.

Complexity of Cellular Automata in Hyperbolic Spaces

Now, we are ready to give the results about the complexity classes of HCA's.

SAT and NP-complete Problems

In [18], HCA's are proved to be able to solve **SAT** in polynomial time. Historically, the possibility to solve **NP**-complete problems in the hyperbolic plane was first announced in [24]. Although the authors of [18] were not aware of paper [24], the latter paper does not involve cellular automata and does not provide a precise description of how **SAT** can be solved in the new frame. On the contrary, [18] describes an HCA which is able to solve the problem. In [18], the computation is estimated as quadratic. In fact it can be proved to be linear in the size of the input.

The solution for **SAT** is easy: it makes use of a Fibonacci tree, in which only two nodes are selected among the sons of a node. Each level represents the possible assignment of true and false values to the variable indexed by this level. The computation of all possible assignments until the level n, where n is the number of variables is triggered at initial time. Once it is reached, the information comes back to the root from the leaves of the tree, i. e. the nodes which are on the level n of the tree: each node computes the OR on the values of its left-hand side and right-hand side sons. Accordingly, the root gives **true** if and only if there is a branch from it to a leaf along which the value is always **true**.

From this, applying classical tools of the theory of complexity, we obtain that any **NP**-complete problem can be solved in polynomial time by an appropriate cellular automaton of the hyperbolic plane.

P = NP in the Hyperbolic Plane

From what we have seen previously, we have that the classical class **NP** is contained in the class of HCA's which work in polynomial time, denoted by $\mathbf{P}_h$. Now, it is also possible to define $\mathbf{NP}_h$ for HCA's, taking the classical definition of non-deterministic computations in polynomial time.

As shown in [6], it turns out that $\mathbf{P}_h = \mathbf{NP}_h$. The key point is that the computation of a non-deterministic Turing machine in time $O(t(n))$, with $t(n) \geq n$, can be computed by a deterministic HCA in time $O(t^2(n))$.

From this theorem, the following surprising result can easily be derived, see [6]:

$$\mathbf{P}_h = \mathbf{NP}_h = \mathbf{PSPACE}$$

where **PSPACE** is the classical class of functions computed in polynomial space by a Turing machine.

Of course, in these results, a basic ingredient is the possibility, given by the hyperbolic plane, to occupy a working space of exponential area within a polynomial time. The above process for solving **SAT** is a basic example of such a possibility.

Other Parts of the Complexity Hierarchy of HCA's

In fact, if we look at the hierarchy of complexity classes for HCA's, we get a landscape which is very different from the classical situation.

We have the following situation, described in [5]:

$$\begin{aligned}\mathbf{DLOG}_h &= \mathbf{NLOG}_h = \mathbf{P}_h = \mathbf{NP}_h = \mathbf{PSPACE}\\ &\subsetneq \mathbf{PSPACE}_h = \mathbf{EXPTIME}_h = \mathbf{NEXPTIME}_h\\ &= \mathbf{EXPSPACE}\end{aligned}$$

We notice that compared to the Euclidean analogs, the hyperbolic hierarchy seems to be very flat. As, by construction, $\mathbf{P}_h \subsetneq \mathbf{EXPTIME}_h$, there are indeed two classes on which the hierarchy concentrates.

We also have $\mathbf{NP}_h \subsetneq \mathbf{AP}_h$, unless $\mathbf{PSPACE} = \mathbf{NEXPTIME}$, where $\mathbf{AP}_h$ denotes the class of alternate HCA's. As with classical machines, an alternate HCA is defined on the set of configurations of a non-deterministic HCA. In the tree of these configurations, certain nodes are called existential, others are called universal. At an existential node, the node is accepting if and only if it has at least one accepting child. At a universal node, the node is accepting if and only if all its children are accepting. The result about $\mathbf{AP}_h$ indicates a similar situation with the Euclidean classes where $\mathbf{P} \subsetneq \mathbf{AP}$, unless $\mathbf{P} = \mathbf{PSPACE}$. Accordingly, we may expect that alternating HCA's should be more powerful than HCA's, either deterministic or non-deterministic.

On Specific Problems of Cellular Automata

Synchronization of an HCA

Although no paper is especially devoted to this problem, we mention it because it has an analog to the standard problem of the firing squad in one-dimensional CA's, and we shall use it in the next section.

In fact, as mentioned more or less explicitly in papers devoted to HCA, see [5,16], for instance, it is very easy to synchronize a disc or a sector inside a disc, defined by a tree rooted at the center of the disc. The idea is simply to simulate any classical algorithm of synchronization of a one-dimensional CA on each branch of the tree or each radius of the disc.

The synchronization is linear in the radius of the disc or the height of the tree.

Communications between HCA's

Another problem, more specific to HCA's is the communication between HCA's, possibly distant ones. Two papers study the problem in different settings, see [12] and [14].

In [12] the question is: how to establish a contact between two cells of a HCA, possibly distant ones? The paper provides a solution based on a new system of coordinates in which there is not an origin. The new system is based on the possibility to represent the hyperbolic plane as a union of growing quarters. We fix such a sequence in an appropriate way. Each term of the sequence is a Fibonacci tree, indexed by an integer n, and it contains all the trees indexed by m when $m > n$. Inside a given Fibonacci tree, we use the standard system of coordinates, indicated in Sect. "The Locating Problem in Hyperbolic Tilings". In the construction, the roots of the mentioned trees belong to a line δ. It is not difficult to see that sending signals on δ makes it possible for the cell to establish a contact in a linear time with respect to their mutual distance. Once the signals of each cell reach δ, they send in both directions a new signal at speed 1/2. A pair of such signals meet, which triggers both the communication signal and a killing signal, at speed 1, to erase the now useless signals sent in the wrong directions.

In [14], another problem is considered. This time all cells may dispatch messages, and each cell forwards the messages it receives and to which it does not want to reply. Accordingly, the same cell may be an emitter of messages, a receiver of messages, and a relay in the message system. The idea is to use the tree property to be in bijection with the tiling as follows: each emitting cell considers that it is the center of the hyperbolic plane, and the message is accompanied by an address which is updated by the relays and which is the address in the tree whose root is the sender of the message. This allows any receiver willing to answer the message to send it to the right emitter. Again, the complexity of the computation is linear in the mutual distance of a sender and a receiver.

Universality in Cellular Automata in Hyperbolic Spaces

Of course, from the existence of universal cellular automata on the line, we conclude that there are universal HCA's. This means that there are HCA's which are able to simulate any universal device, a Turing machine for instance.

Paradoxically, the study of universal HCA's has very few results. There is a universal HCA in the hyperbolic plane with 9 states, see [22], a recent result improving [4] where the HCA had 22 states. There is a universal HCA in the hyperbolic 3D space with 5 states, see [13]. At last, there is an intrinsically universal HCA in the hyperbolic plane, see [16].

Universal HCA's with a Small Number of States

First, we have to notice that the just mentioned universal HCA's with a small number of states are in fact **weakly** universal HCA's. The term **weak** refers to two conditions:

- the HCA needs an infinite initial configuration;
- the initial configuration is ultimately periodic.

Note that these conditions are standardly used with ordinary CA's where universality with a small number of states is reached with such conditions in many cases.

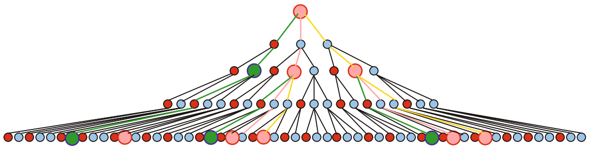
The second condition requires some explanation. In the context of a hyperbolic space, the notion of periodicity is not as clear as it is in the Euclidean case. Accordingly, we mean, by ultimate periodicity that at a large, *i. e.* outside a big enough domain, the configuration is globally invariant under a shift.

The above mentioned universal HCA's with a small number of states are obtained by a similar construction. They both simulate a railway circuit with the kind of switches, described by [26]. While in [26] a Turing machine is simulated, in [4,22] and in [13], we simulate a register machine. It can be remarked that the bigger number of states in the hyperbolic plane is due to the management of **crossings** of tracks which can be simply replaced by **bridges** in the hyperbolic 3D space. Moreover, as a cell in the hyperbolic 3D space has 12 neighbors, there are much more combinations of states which can be used to differentiate the relevant steps of the computation.

An Intrinsically Universal HCA

The intrinsically universal HCA is required to simulate any HCA in the same space. Of course, both the simulating HCA and the simulated one are required to work starting from finite configurations only.

In [16], two ingredients are used to achieve the simulation. One ingredient is the synchronization algorithm mentioned in Sect. "On Specific Problems of Cellular Automata". The second is the construction of **scaled** trees. The construction consists of building a new Fibonacci tree inside the tiling, but with a constant distance k between two consecutive nodes on a same branch. It is not difficult to construct such a tree, which is illustrated by Fig. 3.



Cellular Automata in Hyperbolic Spaces, Figure 3
A scaled tree by a factor 2

The constant k is computed in such a way that a disc of radius k contains both an encoding of the initial configuration of the HCA to be simulated, say A, and an encoding of the transition table of A. Figure 4 illustrates the mechanism of propagation of the scaled tree.

Each step of the simulated HCA A is simulated by a **cycle** of steps of the simulating HCA U. The number of steps of U in a cycle is not constant. It may be increasing, especially if the simulated configuration is growing during its own computation. The synchronization algorithm of Sect. “On Specific Problems of Cellular Automata” is used to delimit the stages into which a cycle is split. These stages are the reception of the current states of the neighbors of the simulated cell of A, for each simulating cell of U. When this is achieved, possibly at different times for each simulating cell, the new state is determined and it is installed in the appropriate region, controlled by the simulating cell. When this is performed, the cell waits until it is informed by its simulating sons that their step of computation is completed. When this is the case, the cell informs its father in the scaled tree that it finished its computation. Accordingly, when the central cell receives the message of completion from all its neighbors of the scaled tree, it knows that the computation of this step of A is finished. Then the comparison with the previous configuration is performed, thanks to a synchronization. Depending on the result of the comparison, the computation is stopped if there was no difference, or it goes on, when a difference was noticed.

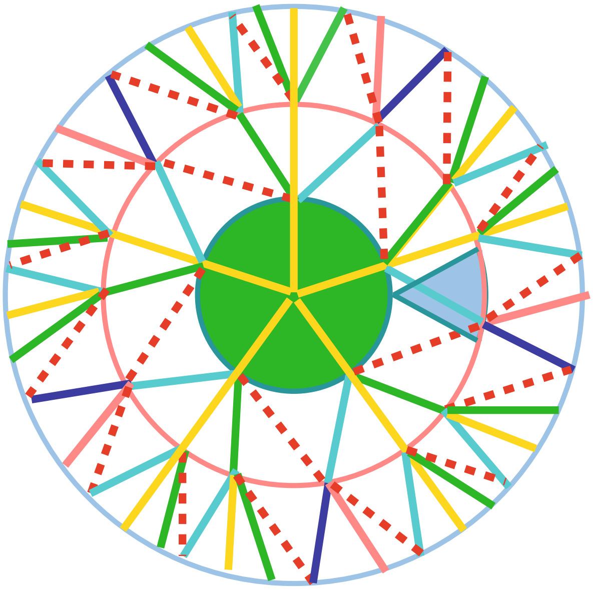
The Connection with Tiling Problems

As usual, cellular automata have deep connections with tilings.

This is probably the case with HCA’s, although, up to now, the single connection is the possibility to implement them in the tilings, thanks to the coordinate system.

However, this system itself appeared to be useful in order to investigate the properties of tilings in the hyperbolic plane and in the hyperbolic spaces of higher dimensions, namely the dimensions 3 and 4.

Indeed, the splitting method could be applied to the tiling $\{5, 3, 4\}$ of the hyperbolic 3D space, see [15,21]. It turned out to be possible to use an old tool of the late 19th century, Schlegel diagrams, to both represent the tiles and the construction of the tiling as a process which is infinite in time. The application of the splitting method revealed an interesting property. The language of the splitting of this tiling provides us with a **natural** example of a language which is neither rational nor context free. As a corollary, the algorithm to compute the path from a tile to the root of its tree is cubic in time with respect to the size of the coordinate of the cell supported by the tile.



Cellular Automata in Hyperbolic Spaces, Figure 4
Propagation of a scaled tree

The splitting method could also be applied to the tiling $\{5, 3, 3, 4\}$ of the hyperbolic 4D space. It provides us with a simple system of coordinates to explore this tiling which is the natural extension of the tiling $\{5, 3, 4\}$ of the hyperbolic 3D space. Note that the same process which allows to go from the pentagon with right angles to Poicaré’s dodecahedron also allows to go from that dodecahedron to the 120-cell. This process is called **orthogonal completion** in [11,15]. Together with an appropriate notion of interior and exterior, it allows to get a correct orientation in the hyperbolic 4D space and to correctly use the dimensional analogy with the spaces of lower dimension.

There are also important applications to tilings of the hyperbolic plane. As already mentioned about the communication between cells of a HCA, these algorithms benefit from the study of the tilings which are made possible by using the splitting method. The new results which are obtained in an algorithmic way are indicated in [15], in particular in its Chapter 4. An important application to the study of tilings in the hyperbolic plane is given by the recent result given in [17]: there, it is proved that the tiling problem of the hyperbolic plane is undecidable.

Future Directions

There are still a lot of problems to investigate about cellular automata in hyperbolic spaces. Let us look at a few of them.

In the classical study of cellular automata in an Euclidean space, there are well known theorems about the characterization of cellular automata in terms of mappings over the space of configurations. There are also well known studies about properties of the global transition function, which transforms a given configuration into the next one which is obtained, for each cell, by application of the transition table of the automaton. What can be said in the hyperbolic spaces? Are these theorems, or their adaptions, still valid in this new setting?

Another direction is a closer study of HCA in the hyperbolic 4D space. The foundation is ready: it is enough, now, to pave the way.

In the very beginning of this article, we mentioned possible applications. We think that this is also a promising direction for future investigations. A few applications were recently performed, see [1,23], for instance, but nothing significant enough to enter this article. We think that what was said in Sect. "Communications Between HCA's" is a serious foundation for future developments.

At last, the connection with tilings gives us the hope that the method initiated by the implementation of cellular automata in hyperbolic spaces will still help to improve the study of tilings in hyperbolic spaces.

Acknowledgments

The author is particularly in debt to Andrew Adamatzky for giving him the task to write this article.

Bibliography

Primary Literature

1. Chelghoum K, Margenstern M, Martin B, Pecci I (2004) Celluladr automata in the hyperbolic plane: proposal for a new environment. In: Proceedings of ACRI'2004, Amsterdam, 25–27 October 2004. Lecture Notes in Computer Sciences, vol 3305. Springer, Berlin, pp 678–687
2. Fraenkel AS (1985) Systems of numerations. Amer Math Mon 92:105–114
3. Gromov M (1981) Groups of polynomial growth and expanding maps. Publ Math IHES 53:53–73
4. Herrmann F, Margenstern M (2003) A universal cellular automaton in the hyperbolic plane. Theor Comput Sci 296:327–364
5. Iwamoto C, Margenstern M (2003) A survey on the Complexity Classes in Hyperbolic Cellular Automata. In: Proceedings of SCI'2003, V, pp 31–35
6. Iwamoto C, Margenstern M, Morita K, Worsch T (2002) Polynomial-time cellular automata in the hyperbolic plane accept accept exactly the PSPACE Languages. SCI'2002, Orlando, pp 411–416
7. Margenstern M (2000) New tools for cellular automata in the hyperbolic plane. J Univers Comput Sci 6(12):1226–1252
8. Margenstern M (2002) A contribution of computer science to the combinatorial approach to hyperbolic geometry. In: SCI'2002, Orlando, USA, 14–19 July 2002
9. Margenstern M (2002) Revisiting Poincaré's theorem with the splitting method. In: Bolyai'200, International Conference on Geometry and Topology, Cluj-Napoca, Romania, 1–3 October 2002
10. Margenstern M (2003) Implementing Cellular Automata on the Triangular Grids of the Hyperbolic Plane for New Simulation Tools. ASTC'2003, Orlando, 29 March–4 April
11. Margenstern M (2004) The Tiling of the Hyperbolic 4D Space by the 120-cell is Combinatoric. J Univers Comput Sci 10(9):1212–1238
12. Margenstern M (2006) A new way to implement cellular automata on the penta- and heptagrids. J Cell Autom 1(1): 1–24
13. Margenstern M (2007) A universal cellular automaton with five states in the 3D hyperbolic space. J Cell Autom 1(4):317–351
14. Margenstern M (2007) On the communication between cells of a cellular automaton on the penta- and heptagrids of the hyperbolic plane. J Cell Autom 1(3):213–232
15. Margenstern M (2007) Cellular Automata in Hyperbolic Spaces, vol 1: Theory. Old City Publishing, Philadelphia, p 422
16. Margenstern M (2008) A Uniform and Intrinsic Proof that there are Universal Cellular Automata in Hyperbolic Spaces. J Cell Autom 3(2):157-180
17. Margenstern M (2008) The domino problem of thehyperbolic plane is undecidable. Theor Comput Sci (to appear)
18. Margenstern M, Morita K (1999) A Polynomial Solution for 3-SAT in the Space of Cellular Automata in the Hyperbolic Plane. J Univers Comput Syst 5:563–573
19. Margenstern M, Morita K (2001) NP problems are tractable in the space of cellular automata in the hyperbolic plane. Theor Comput Sci 259:99–128
20. Margenstern M, Skordev G (2003) The tilings $\{p, q\}$ of the hyperbolic plane are combinatoric. In: SCI'2003, V, pp 42–46
21. Margenstern M, Skordev G (2003) Tools for devising cellular automata in the hyperbolic 3D space. Fundamenta Informaticae 58(2):369–398
22. Margenstern M, Song Y (2008) A new universal cellular automaton on the pentagrid. AUTOMATA'2008, Bristol, UK, 12–14 June 2008
23. Martin B (2005) VirHKey: a VIRtual Hyperbolic KEYboard with gesture interaction and visual feedback for mobile devices. In: MobileHCI'05, September, Salzburg, Austria
24. Morgenstein D, Kreinovich V (1995) Which algorithms are feasible and which are not depends on the geometry of space-time. Geombinatorics 4(3):80–97
25. Róka Z (1994) One-way cellular automata on Cayley Graphs. Theor Comput Sci 132:259–290
26. Stewart I (1994) A Subway Named Turing. Math Recreat Sci Am 90–92

Books and Reviews

Alekseevskij DV, Vinberg EB, Solodovnikov AS (1993) Geometry of spaces of constant curvature. In: Vinberg EB (ed) Geometry II, Encyclopedia of Mathematical Sciences, vol 29. Springer, Berlin

Berlekamp ER, Conway JH, Guy RK (1982) Winning Ways for Your Mathematical Plays. Academic Press

Bonola R (1912) Non-Euclidean Geometry. Open Court Publishing Company (also, (1955) Dover, New York)

Codd EF (1968) Cellular Automata. Academic Press, New York

Coxeter HSM (1969) Introduction to Geometry. Wiley, New York

Coxeter HSM (1974) Regular Complex Polytopes. Cambridge University Press, Cambridge

Delorme M, Mazoyer J (eds) (1999) Cellular automata, a parallel model. Kluwer, p 460

Epstein DBA, Cannon JW, Holt DF, Levi SVF, Paterson MS, Thurston WP (1992) Word Processing in Groups. Jones and Barlett, Boston

Grünbaum B, Shephard GS (1987) Tilings and Patterns. Freeman, New York

Gruska J (1997) Foundations of computing. International Thomson Computer Press

Knuth DE (1998) The Art of Computer Programming, vol II: Seminumerical algorithms. Addison-Wesley

Meschkowski H (1964) Noneuclidean Geometry. Translated by Shenitzer A. Academic Press, New York

Millman RS, Parker GD (1981) Geometry, a metric approach with models. Springer

Minsky ML (1967) Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs

Ramsay A, Richtmyer RD (1995) Introduction to Hyperbolic Geometry. Springer

Toffoli T, Margolus N (1987) Cellular automata machines. MIT Press, Cambridge

von Neuman J (1966) Theory of self-reproducing automata. Edited and completed by A.W. Burks. The University of Illinois Press, Urbana

Wolfram S (1994) Cellular Automata and Complexity. Addison-Wesley

Wolfram S (2002) A New Kind of Science. Wolfram Media

Cellular Automata and Language Theory

Martin Kutrib
Institut für Informatik, Universität Giessen,
Giessen, Germany

Article Outline

Glossary

Cellular automaton A (one-dimensional) cellular automaton is a linear array of cells which are connected to both of their nearest neighbors. The total number of cells in the array is determined by the input data. They are exactly in one of a finite number of states, which is changed according to local rules depending on the current state of a cell itself and the current states of its neighbors. The state changes take place simultaneously at discrete time steps. The input mode for cellular automata is called parallel. One can suppose that all cells fetch their input symbol during a pre-initial step.

Iterative array Basically, iterative arrays are cellular automata whose leftmost cell is distinguished. This so-called communication cell is connected to the input supply and fetches the input sequentially. The cells are initially empty, that is, in a special quiescent state.

Formal language The data on which the devices operate are strings built from input symbols of a finite set or alphabet. A subset of strings over a given alphabet is a formal language.

Signal Signals are used to transmit and encode information in cellular automata. If a cell changes to the state of its neighbor after some k time steps, and if subsequently its neighbors and their neighbors do the same, then the basic signal moves with speed $\frac{1}{k}$ through the array. With the help of auxiliary signals, rather complex signals can be established.

Closure property Closure properties of families of formal languages indicate their robustness under certain operations. A family of formal languages is closed under some operation, if any application of the operation on languages from the family yields again a language from the family.

Turing machine A Turing machine is the simplest form of a universal computer. It captures the idea of an effective procedure or algorithm. At any time the machine is in any one of a finite number of states. It is equipped with an infinite tape divided into cells and a read-write head scanning a single cell. Each cell may contain a symbol from a finite set or alphabet. Initially, the finite input is written in successive cells. All other cells are empty. Dependent on a list of instructions, which serve as the program for the machine, the action is determined completely by the current state and the symbol currently scanned by the head. The action comprises the symbol to be written on the current cell, the new state of the machine, and the information of whether the head should move left or right.

Decidability A formal problem with two alternatives is decidable, if there is an algorithm or a Turing machine that solves it and halts on all inputs. That is, given an encoding of some instance of the problem, the algorithm or Turing machine returns the correct answer *yes* or *no*. The problem is semidecidable, if the algorithm halts on all instances for which the answer is *yes*.

Definition of the Subject

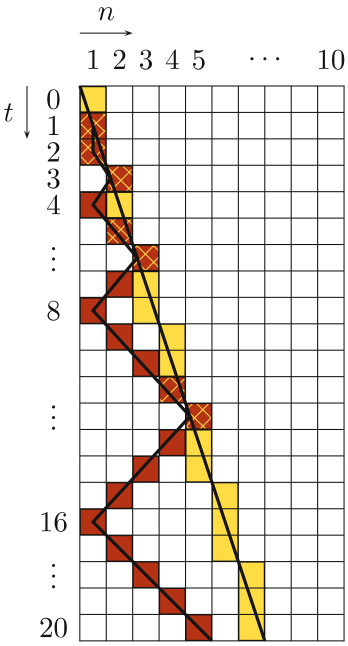
One of the cornerstones in the theory of automata is the early result of John von Neumann, who solved the logical problem of nontrivial self-reproduction. He employed a mathematical device which is a multitude of interconnected identical finite-state machines operating in parallel to form a larger machine. He showed that it is logically possible for such a nontrivial computing device to replicate itself ad infinitum [97]. Such devices are commonly called cellular automata (abbreviated, CA), and can be considered as homogeneously structured models for massively parallel computing systems. The global behavior of cellular automata is achieved by local interactions only. While the underlying rules are quite simple, the global behavior may be rather complex. In general, it is unpredictable.

The data supplied to CAs can be arranged as strings of symbols. Instances of problems to solve can be encoded as strings with a finite number of different symbols. Furthermore, complex answers to problems can be encoded as binary sequences such that the answer is computed bit by bit. In order to compute one piece of the answer, the set of possible inputs is split into two sets associated with the binary outcome. From this point of view, the computational capabilities of CAs are studied in terms of string acceptance, that is, the determination to which of the two sets a given string belongs. These investigations are with respect to and with the methods of language theory. They originated in Stephen N. Cole [18,19] and Alvy R. Smith [76,80]. Over the years substantial progress has been achieved, but there are still some basic open problems with deep relations to other fields. So, exploring the capabilities of cellular automata may benefit the understanding of the nature of parallelism and nondeterminism.

Introduction

In general, the specification of a cellular automaton includes the type and specification of the cells, their interconnection scheme (which can imply a dimension of the system), the local rules which are formalized as local transition function, and the input and output modes. With an eye towards language acceptance, we consider one-dimensional synchronous devices with nearest neighbor connections whose cells are deterministic finite-state machines. They are commonly called cellular automata in case of parallel input mode, and iterative arrays (abbreviated, IA) if the input mode is sequential. If each cell is connected to only one of its neighbors, say to the right one, then the flow of information through the array is from right to left. The corresponding device is a one-way cellular automaton (abbreviated, OCA). In any case the number of cells is determined by the length of the input string; there is one cell per symbol. If the input is in parallel, then all cells fetch their input symbol during a pre-initial step. To this end, the set of symbols has to be a subset of the set of states. Sometimes for practical reasons and for the design of systolic algorithms, a sequential input mode is more convenient than the parallel one. In iterative arrays the leftmost cell is distinguished to be the communication cell. It is equipped with a one-way read-only input tape.

In order to obtain the binary answer of a system we have to overcome a problem with the end of the computation. It follows from the definitions that the machines never halt. A way to cope with the situation is to define a predicate on configurations. The answer depends on whether a configuration satisfies the predicate or not. Here we apply the common predicate that requires a border cell or the communication cell to be in some state designated to be an accepting state. Further predicates are studied, e. g., in [38,82], while more general input modes are considered in [57]. Due to the bounded number of cells in a computation, the time complexity is exponentially bounded. After exceeding the bound, the computation runs into a loop and is rather useless. With respect to the wide language classes obeying an exponential or polynomial time bound, parallel devices cannot take advantage of their large number of processing elements. They are just a factor to be multiplied with the size of the input. So, there is a particular interest in fast computations, that is, in real-time and linear-time computations. Real time is determined by the shortest time necessary for nontrivial computations, whereas linear time is real time multiplied by an arbitrary but fixed constant greater than or equal to one. In addition, we consider general computations without time bounds which, actually, are exponentially time bounded. The following well-known example from [17] joins several notions. It uses signals in order to construct the mapping $n \longmapsto 2^n$ in time; i. e., the leftmost cell recognizes the time steps 2^n, $n \geq 1$. The time constructor is then extended to a real-time CA and IA.



Cellular Automata and Language Theory, Figure 1
Space-time diagram showing signals of a real-time two-way acceptor of the language $\{a^{2^n} \mid n \geq 1\}$

Example 1 The unary language consisting of the strings $\{a^{2^n} \mid n \geq 1\}$ is accepted by IAs as well as by CAs in real time.

At initial time the leftmost cell emits a signal which moves with speed $\frac{1}{3}$ to the right; i. e., the signal alternates from moving one cell to the right and staying for two time steps in a cell (see Fig. 1). In addition, another signal is emitted which moves with maximal speed (speed 1). It bounces between the slow signal and the leftmost cell. It is easy to see that the signal passes through the leftmost cell exactly at the time steps 2^n, $n \geq 1$. Finally, an iterative array can accept its input when the last input symbol is read at one of these time steps. Similarly, in a CA computation the rightmost cell can emit a third signal which moves with maximal speed to the left. This signal arrives at the leftmost cell at the time step which corresponds to the length of the input. If it meets the bouncing signal at its arrival, the input is accepted. □

More results about mappings that are constructible in the above sense can be found in [13,63,93]. In [27,94] the series of prime numbers is constructed in real-time devices.

The investigations of iterative arrays and cellular automata as cellular language acceptors originated in [18,19] and [76,80]. In [18,19], it is shown that the family of languages accepted by real-time IAs is closed under intersection, union, and complementation, but is not closed under concatenation and reversal; and in [76,80], where among other results the identity of the sequential complexity class DSPACE(n) (i. e., the class of languages accepted by deterministic Turing machines whose tape is bounded by the length of the input n) and the family of languages accepted by CAs without time bound is shown. A long-standing open problem is the question whether or not one-way information flow is a strict weakening of two-way information flow for unbounded time. In [16,32] it is proved that a PSPACE-complete language is accepted by OCAs, from which one can draw conclusions about the hardness of sequential OCA simulations. Furthermore, in the same papers strong closure properties are derived for the family of OCA languages. In addition, it is a proper superset of the context-free languages which, in turn, are of great practical relevance.

The proofs are based on characterizations of the parallel language families by certain types of customized sequential machines. Such machines have been developed for all classes of acceptors which are here under consideration [32,36,37]. In particular, speed-up theorems are given that allow to speed up the time beyond real time linearly. Therefore, linear-time computations can be sped up close to real time. Nevertheless, for OCAs and IAs linear time is strictly more powerful than real time. The problem is still open for CAs. In fact, it is an open question whether real-time CAs are strictly weaker than unbounded time CAs. If both classes coincide, then a PSPACE-complete language would be accepted in polynomial time! Apart from that it is known that linear-time CAs can be simulated by unbounded time OCAs [16,32].

The rest of the article is organized as follows. In the following section (Sect. "Cellular Language Acceptors"), some basic notions and formal definitions are given. The problem in connection with the end of computations is discussed in more detail, and honest time complexities are derived informally. Then, in Sect. "Tools and Techniques" selected tools and techniques are presented that can be applied to prove or to disprove that certain languages are accepted by certain devices. In particular, it is shown that two-way devices can simulate the data structure *stack* without loss of time. It is often much harder to disprove the acceptance of languages than to prove it, since the technique of a suitable construction is trivially not applicable. Some techniques based on counting and pumping arguments are shown and applied in order to obtain witness languages. In Sect. "Computational Capacities" computational capacity aspects are investigated. A basic hierarchy of language families defined by cellular automata and iterative arrays is established. The levels are compared with well-known linguistic families. Next, Sect. "Closure Properties" is devoted to exploring the closure properties of the language families in question. For example, it turns out that the languages accepted by unbounded time one-way and two-way cellular automata as well as iterative arrays share the strong closure properties of the class DSPACE(n) which characterizes the deterministic context-sensitive languages. Decidability problems are considered in Sect. "Decidability Problems". By reductions of Turing machine problems, it follows that almost all interesting properties are undecidable. In fact, they are not even semidecidable for the weakest devices in question. This emphasizes once more the power gained in parallelism even if the number of cells is bounded. In general, their behavior is unpredictable. Finally, in Sect. "Future Directions", some future directions of cellular language acceptors are discussed.

Cellular Language Acceptors

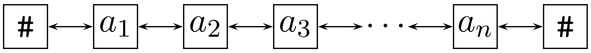
We denote the set of nonnegative integers by $\mathbb{N}$. In connection with formal languages, strings are called *words*. Let A^* denote the set of all words over a finite alphabet A. The *empty word* is denoted by λ, and we set $A^+ = A^* - \{\lambda\}$. For the *reversal of a word w* we write w^R and for its *length* we write $|w|$. For the *number of occurrences* of a symbol a in w we use the notation $|w|_a$. We use $\subseteq$ for *inclusions* and $\subset$ for *strict inclusions*. In order to avoid technical overloading in writing, two languages L and L' are considered to be equal, if they differ at most by the empty word, i. e., $L - \{\lambda\} = L' - \{\lambda\}$. Throughout the article, two automata or grammars are said to be *equivalent* if and only if they accept or generate the same language.

The cells of a cellular automaton are identified by positive integers. In order to handle the application of the local transition function to the border cells, we assume that the missing neighbors are in a permanent so-called border state. A formal definition is:

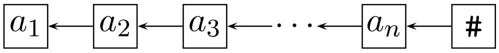
Definition 2 *A* two-way cellular automaton *(CA) is a system* $\langle S, \delta, \#, A, F\rangle$, *where*

1. *S is the finite, nonempty set of* cell states,
2. *$\# \notin S$ is the permanent* boundary state,
3. *$A \subseteq S$ is the nonempty set of* input symbols,
4. *$F \subseteq S$ is the set of* accepting states, *and*
5. *$\delta : (S \cup \{\#\}) \times S \times (S \cup \{\#\}) \to S$ is the* local transition function.

If the flow of information is restricted to one-way, the resulting device is a one-way cellular automaton (abbreviated, OCA). In such devices, the next state of each cell depends on the state of the cell itself and the state of its immediate neighbor to the right.



Cellular Automata and Language Theory, Figure 2
A (two-way) cellular automaton



Cellular Automata and Language Theory, Figure 3
A one-way cellular automaton

A *configuration* of a cellular automaton $\langle S, \delta, \#, A, F\rangle$ at time $t \geq 0$ is a description of its global state, which is formally a mapping $c_t\{1, \ldots, n\} \to S$, for $n \geq 1$. The configuration at time 0 is defined by the given input $w = a_1 \cdots a_n \in A^+$. We set $c_0(i) = a_i$, for $1 \leq i \leq n$. Configurations may be represented as words over the set of cell states in their natural ordering. For example, the initial configuration for w is represented by $\#a_1 a_2 \cdots a_n\#$. Successor configurations are computed according to the global transition function Δ. Let c_t, $t \geq 0$, be a configuration with $n \geq 2$, then its successor c_{t+1} is defined as follows:

$$c_{t+1} = \Delta(c_t)$$

$$\iff \begin{cases} c_{t+1}(1) = \delta(\#, c_t(1), c_t(2)) \\ c_{t+1}(i) = \delta(c_t(i-1), c_t(i), c_t(i+1)) , \\ \quad i \in \{2, \ldots, n-1\} \\ c_{t+1}(n) = \delta(c_t(n-1), c_t(n), \#) \end{cases}$$

for CAs, and

$$c_{t+1} = \Delta(c_t)$$

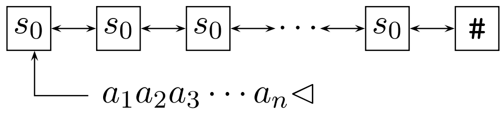
$$\iff \begin{cases} c_{t+1}(i) = \delta(c_t(i), c_t(i+1)) , \\ \quad i \in \{1, \ldots, n-1\} \\ c_{t+1}(n) = \delta(c_t(n), \#) \end{cases}$$

for OCAs. For $n = 1$, the next state of the sole cell is $\delta(\#, c_t(1), \#)$. Thus, Δ is induced by δ.

A computation can be represented as a space-time diagram, where each row is a configuration and the rows appear in chronological ordering.

In order to define iterative arrays formally we have to provide an initial (quiescent) state for the cells. We assume that once the whole input is consumed an end-of-input symbol is supplied permanently.

Definition 3 *An* iterative array *(IA) is a system* $\langle S, \delta, \delta_0, s_0, \#, \triangleleft, A, F\rangle$, *where*



Cellular Automata and Language Theory, Figure 4
An iterative array

1. *S is the finite, nonempty set of* cell states,
2. $s_0 \in S$ *is the* quiescent state,
3. $\# \notin S$ *is the permanent* boundary state,
4. $\triangleleft \notin A$ *is the* end-of-input symbol,
5. *A is the finite, nonempty set of* input symbols,
6. $F \subseteq S$ *is the set of* accepting states,
7. $\delta : S \times S \times (S \cup \{\#\}) \to S$ *is the* local transition function for non-communication cells *satisfying* $\delta(s_0, s_0, s_0) = \delta(s_0, s_0, \#) = s_0$,
8. $\delta_0 : (A \cup \{\triangleleft\}) \times S \times (S \cup \{\#\}) \to S$ *is the* local transition function for the communication cell.

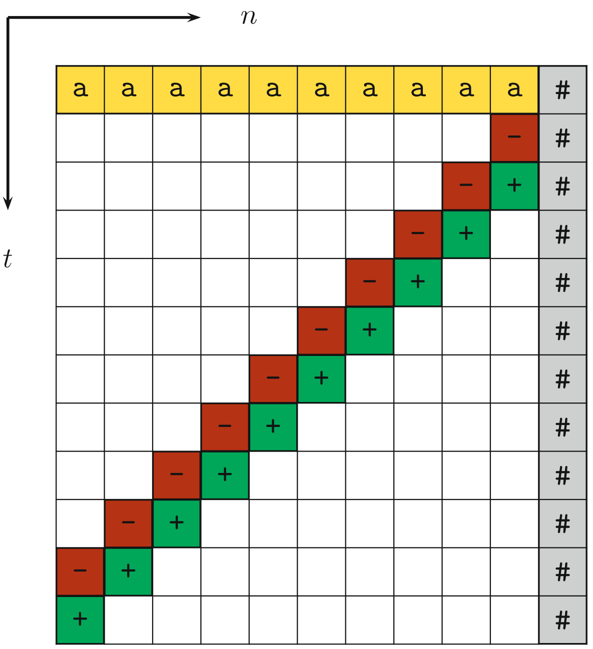
A configuration of an iterative array $\langle S, \delta, \delta_0, s_0, \#, \triangleleft, A, F\rangle$ at time $t \geq 0$ is a pair (w_t, c_t), where $w_t \in A^*$ is the remaining input sequence and $c_t : \{1, \dots, n\} \to S$, for $n \geq 1$, is a mapping that maps the single cells to their current states. The configuration (w_0, c_0) at time 0 is defined by the input word w_0 and the mapping $c_0(i) = s_0$, $1 \leq i \leq n$. The global transition function Δ is induced by δ and δ_0 as follows: Let (w_t, c_t), $t \geq 0$, be a configuration and $i \in \{2, \dots, n-1\}$. Then

$$(w_{t+1}, c_{t+1}) = \Delta((w_t, c_t))$$
$$\iff \begin{cases} c_{t+1}(1) = \delta_0(a, c_t(1), c_t(2)) \\ c_{t+1}(i) = \delta(c_t(i-1), c_t(i), c_t(i+1)) \\ c_{t+1}(n) = \delta(c_t(n-1), c_t(n), \#) \end{cases}$$

where $a = \triangleleft$, $w_{t+1} = \lambda$ if $w_t = \lambda$, and $a = a_1$, $w_t + 1 = a_2 \cdots a_n$ if $w_t = a_1 \cdots a_n$.

An input w is accepted by a CA, OCA, or an IA $\mathcal{M}$ if at some time i during its course of computation the leftmost cell enters an accepting state. The *language accepted by* $\mathcal{M}$ is denoted by $L(\mathcal{M})$. Let $t : \mathbb{N} \to \mathbb{N}$, $t(n) \geq n$ ($t(n) \geq n+1$ for IAs) be a mapping. If all $w \in L(\mathcal{M})$ are accepted with at most $t(|w|)$ time steps, then $L(\mathcal{M})$ is said to be of time complexity t.

Observe that time complexities do not have to meet any further conditions. This general treatment is made possible by the way of acceptance. An input w is accepted if the leftmost cell enters an accepting state at some time $i \leq t(|w|)$. But what if afterwards, a final configuration has been reached? Subsequent states of the leftmost cell are not relevant.



Cellular Automata and Language Theory, Figure 5
An OCA accepting any unary language. Here + is an accepting and — a non-accepting state

Following the different approach to gather the result of a computation at time step $t(|w|)$ by the outside world does not yield the desired outcome in general. In this case, the intrinsic computation may be hidden in the determination of the time step $t(|w|)$. That is, computational power may be added from the outside world. For example, let $L \in \{a\}^+$ be an arbitrary language. Then L is accepted by some OCA with time complexity

$$t(n) = \begin{cases} n & \text{if} \quad a^n \notin L \\ n+1 & \text{if} \quad a^n \in L \end{cases},$$

where the local transition function is easily designed to realize the behavior depicted in the space-time diagram of Fig. 5.

So, it is reasonable to consider only such time complexities t that allow the leftmost cell to recognize the time step $t(n)$. For example, the identity $t(n) = n$ is an honest time complexity for OCAs and CAs. A signal which is initially emitted by the rightmost cell and moves with maximal speed arrives at the leftmost cell exactly at time step n. By slowing down the signal to speed $\frac{x}{y}$, i. e., the signal alternating moves x cells to the left and stays for $y-x$ time steps in a cell, it is seen that the time complexities $\frac{x}{y} \cdot n$, for any positive integers x,y, are also honest. Another example are exponential time complexities $t(n) = k^n$, for any integer $k \geq 2$. Without going too deep into technical details, a corresponding device can be set up as a $k$ary counter. The rightmost cell simulates the least significant digit and adds one to the counter at every time step. The neighboring cell to the left observes when a carry-over appears, increases its own digit and so on. Then, the leftmost cell produces a first carry-over exactly at time step k^n.

The family of languages that are accepted by IAs (and CAs, OCAs) with time complexity t is denoted by $\mathcal{L}_t(\text{IA})$ (and $\mathcal{L}_t(\text{CA})$, $\mathcal{L}_t(\text{OCA})$, respectively). The index is omitted for arbitrary time. Actually, arbitrary time is exponential time due to the space bound. If t is the function $n+1$ (the function n), acceptance is said to be in *real time* and we write $\mathcal{L}_{\text{rt}}(\text{IA})$ ($\mathcal{L}_{\text{rt}}(\text{CA})$, $\mathcal{L}_{\text{rt}}(\text{OCA})$). Since for nontrivial computations an IA has to read at least one end-of-input symbol, real time has to be defined as $(n+1)$-time. The *linear-time* languages $\mathcal{L}_{\text{lt}}(\text{IA})$ are defined according to $\mathcal{L}_{\text{lt}}(\text{IA}) = \bigcup_{k \in \mathbb{Q}, k \geq 1} \mathcal{L}_{k \cdot n}(\text{IA})$, and similarly for CAs and OCAs.

Tools and Techniques

An elementary technique in automata theory is the usage of multiple tracks. Basically, this means to consider the state set as Cartesian product of some smaller sets. Each component of a state is called a *register*, and the same register of all cells together forms a *track*.

The first goal of this section is to show how to simulate pushdown stores, i. e., stores obeying the principle *last in first out*, by IAs and CAs in real time. Assume without loss of generality that at most one symbol is pushed onto or popped from the stack at each time step. We distinguish one cell that simulates the top of the pushdown store. It suffices to use three additional tracks for the simulation. Let the three pushdown registers of each cell be numbered one, two and three from top to bottom. Each cell prefers to have only the first two registers filled. The third register is used as a buffer. In order to reach that charge it obeys the following rules (cf. Fig. 6).

a) If all three registers of its left (upper) neighbor are filled, it takes over the symbol from the third register of the neighbor and stores it in its first register. The old contents of the first and second registers are shifted to the second and third register.

b) If there is only the first register of its left (upper) neighbor filled, the cell erases its first register and shifts the contents of the second and third registers to the first and second register. Observe that the erased symbol is taken over by the left neighbor.

c) Possibly, more than one of these actions are superimposed.

From the simulation, it follows immediately that real-time IAs as well as real-time CAs accept all languages accepted by sequential pushdown automata as long as they work in real time.

Theorem 4 *Given some real-time deterministic pushdown automaton, an equivalent real-time IA and CA can effectively be constructed, i. e., every real-time deterministic context-free language belongs to the families $\mathcal{L}_{rt}$(IA) and $\mathcal{L}_{rt}$(CA).*

Now we turn to a technique for disproving that languages are accepted. In general, the method is based on equivalence classes which are induced by formal languages. If some language induces a number of equivalence classes which exceeds the number of classes distinguishable by a certain device, then the language is not accepted by that device. First we give the definition of an equivalence relation which applies to real-time IAs.

Definition 5 *Let $L \subseteq A^*$ be a language and $l \geq 1$ be a constant. Two words $w \in A^*$ and $w' \in A^*$ are l-equivalent with respect to L if and only if*

$$wu \in L \iff w'u \in L$$

for all $u \in A^*$, $|u| \leq l$. The number of l-equivalence classes with respect to L is denoted by $E(L, l)$.

In [19] the following upper bound for the number of equivalence classes distinguishable by real-time IA is derived.

Lemma 6 *If $L \in \mathcal{L}_{rt}$(IA), then there exists a constant $p \geq 1$ such that $E(l, L) \leq p^l$.*

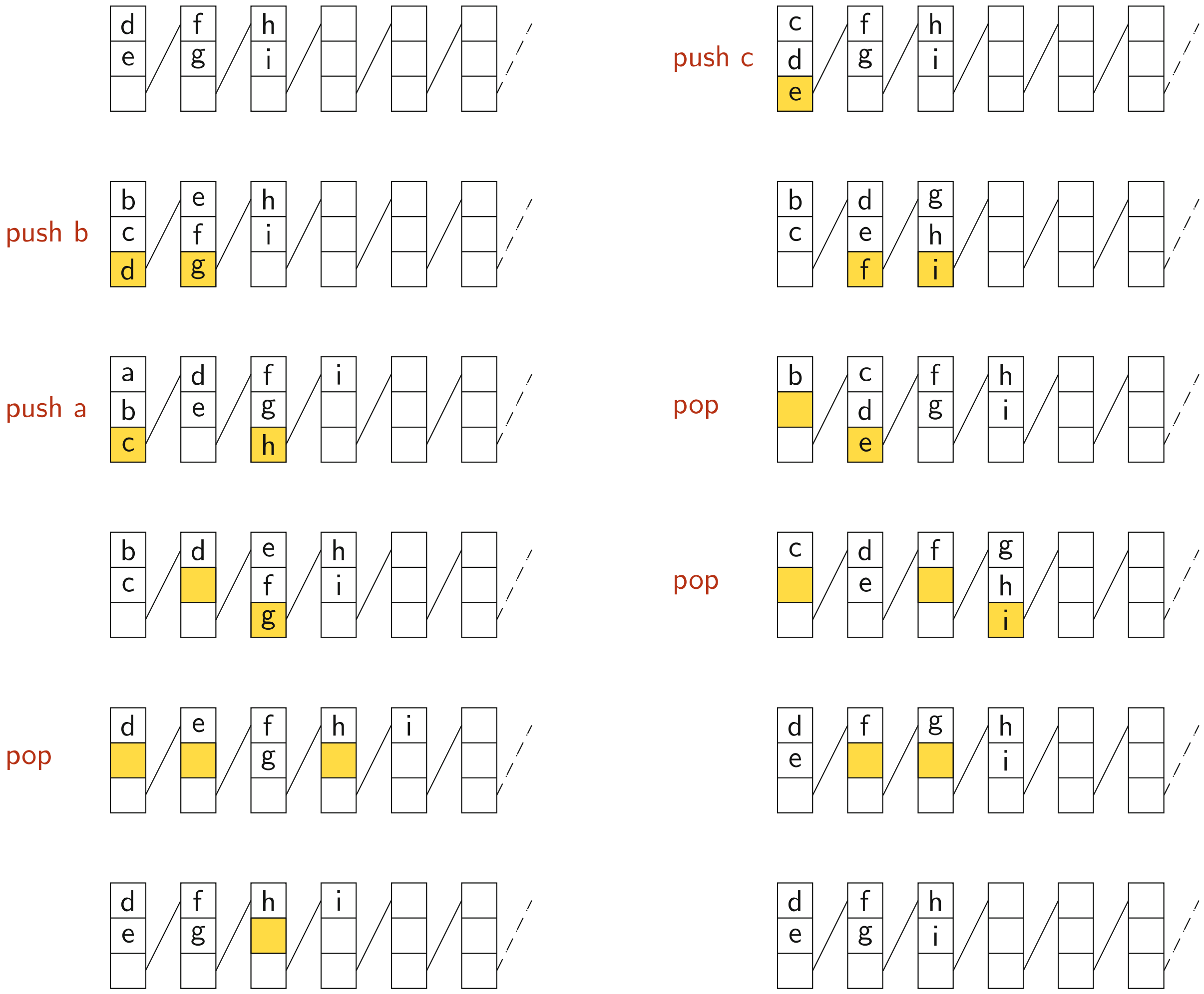
Proof Let $\mathcal{M}$ be a real-time IA with state set S. In order to determine an upper bound for the number of l-equivalence classes with respect to $L(\mathcal{M})$, we consider the possible configurations of $\mathcal{M}$ after reading all but l input symbols. The remaining computation depends on the last l input symbols and the states of the cells $1, \ldots, l+2$. For the $l+2$ states there are at most $|S|^{l+2}$ different possibilities. Setting $p = |S|^3$, we derive $|S|^{l+2} \leq |S|^{3l} = p^l$, and obtain at most p^l different possibilities. Since the number of equivalence classes is not affected by the last l input symbols, there are at most p^l equivalence classes. □

Example 7 The language

$$L = \{\&x_k\& \cdots \&x_1 ? y_1\& \cdots \&y_k\& \mid k \geq 1, x_i^R = y_i z_i \\ \text{and} \quad x_i, y_i, z_i \in \{a, b\}^*\}$$

does not belong to $\mathcal{L}_{\text{rt}}(\text{IA})$.

For a pair of different prefixes $w = \&x_k \& \cdots \& x_1 ?$ and $w' = \& x'_k\& \cdots \& x'_1 ?$ with $|x_i| = |x'_i| = k$, $1 \leq i \leq k$, there exists at least one $1 \leq j \leq k$ such that $x_j \neq x'_j$. This



Cellular Automata and Language Theory, Figure 6
Principle of a pushdown store simulation. Subfigures are in row-major order

implies $w\,\&^{j-1}\,x_j^R\,\&^{k-j+1} \in L$ and $w'\,\&^{j-1}\,x_j^R\,\&^{k-j+1} \notin L$. Since there are 2^{k^2} different prefixes of the given form, language L induces at least 2^{k^2} classes.

On the other hand, if L would be accepted by some real-time IA, then by Lemma 6 there is a constant $p \geq 1$ such that $E(L,2k) \leq p^{2k}$. Since L is infinite, we may choose k large enough such that $2^{k^2} > p^{2k}$, which is a contradiction. □

Now we change to real-time OCAs. The next result is a tool which allows us to show that languages do not belong to the family $\mathscr{L}_{\mathrm{rt}}(\mathrm{OCA})$. It is based on pumping arguments for cyclic strings [68].

Lemma 8 *Let L be a real-time OCA language. Then there exists a constant $p \geq 1$ such that any pair of a word w and an integer k that meets the condition $w^k \in L$ and $k > p^{|w|}$ implies that there is some $1 \leq q \leq p^{|w|}$ such that $w^{k+jq} \in L$, for all $j \geq 0$.*

Proof For a given real-time OCA $\mathcal{M} = \langle S, \delta, \#, A, F \rangle$, we set $p = |S|^2$. Let $w^k \in L(\mathcal{M})$, where $k > p^{|w|}$. Then we consider an accepting computation of $\mathcal{M}$ on input w^k. The initial configuration is represented by $\#w^k\#$. Clearly, a cyclic left part of some configuration leads again to a cyclic left part, though the new left part gets one cell shorter at any time step. Therefore, after $|w|$ time steps the left part of the configuration which still may influence the overall computation result is represented by $\#w_1^{k-1}s_1$, where $|w_1| = |w|$ and $s_1 \in S$. After another $|w|$ time steps, we obtain $\#w_2^{k-2}s_2$, where $|w_2| = |w|$ and $s_2 \in S$. In general, the relevant part of a configuration at time $i \cdot |w|$, $1 \leq i \leq k$, is represented by

$\# w_i^{k-i} s_i$, where $|w_i| = |w|$ and $s_i \in S$. In addition, state s_k is an accepting one.

Since the number of different words w_i is bounded by $|S|^{|w|}$, for $w_i s_i$ there are at most

$$|S|^{|w|+1} \le |S|^{2\left\lceil \frac{|w|+1}{2} \right\rceil} \le p^{|w|}$$

different possibilities. Now $k > p^{|w|}$ implies that $w_i s_i = w_l s_l$ for some $1 \le i < l \le k$. Therefore, there is a loop and, for $q = l - i$, the word w^{k+jq} is accepted, for any $j \ge 0$. □

Example 9 The language $L = \{a^{2^n} \mid n \ge 1\}$ is not accepted by any real-time OCA.

Contrarily assume there is a real-time OCA accepting L. Then we set $w = a, k = 2^p$ and observe that the conditions of Lemma 8 are met. Therefore, a^{2^p+q} as well as a^{2^p+2q} belong to L. If $2^p + q$ is not a power of two, we obtain a contradiction. So, let $2^p + q = 2^{p+r}$, for some $r \ge 1$. We derive $2^{p+r} < 2^{p+r} + q = 2^{p+r} + 2^{p+r} - 2^p = 2^{p+r+1} - 2^p < 2^{p+r+1}$, and conclude that $2^p + 2q$ is strictly in between two consecutive powers of two. Hence, a^{2^p+2q} does not belong to L. □

Next we come back to equivalence classes. By a similar idea as for iterative arrays an equivalence relation can be defined such that the number of equivalence classes distinguishable by real-time OCAs is bounded. But due to the nature of OCA computations, both the prefixes as well as the suffixes of inputs have to be regarded [84,85]. We continue with the equivalence relation:

Definition 10 *Let $L \subseteq A^*$ be a language and $X \subseteq A^*$ and $Y \subseteq A^*$ be two sets of words. Two words $w \in A^*$ and $w' \in A^*$ are (L, X, Y)-equivalent if and only if*

$$xwy \in L \iff xw'y \in L$$

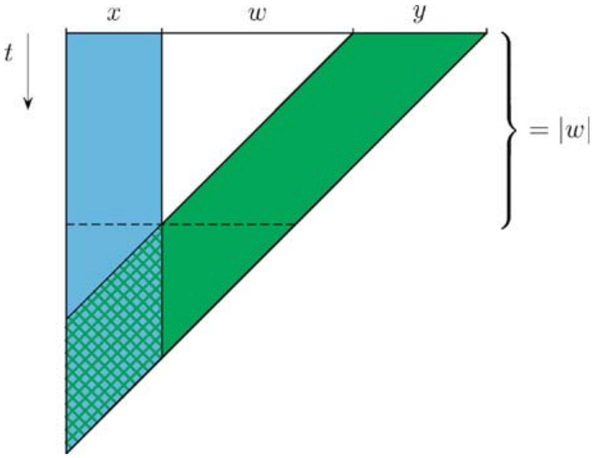
for all $x \in X$ and $y \in Y$.

The next step is to derive an upper bound for the number of equivalence classes which can be distinguished by some real-time OCA.

Theorem 11 *Let $L \subseteq A^*$ be a real-time OCA language and $X = A^{m_1}$ and $Y = A^{m_2}$ be two sets of words for positive integers m_1 and m_2. Then there exists a constant $p \ge 1$ such that the number N of (L, X, Y)-equivalence classes is bounded by*

$$N \le p^{|X|} p^{(m_2+1)|Y|} .$$

Proof For any word w and any $x \in X$ and $y \in Y$ the input xwy implies exactly one real-time OCA configuration at time $|w|$ (see Fig. 7). At this time only the leftmost $|xy| + 1$ cells can still influence the overall computation result, i. e., the cells $1, \dots, |xy| + 1$. The cells $1, \dots, |x|$ are not affected by the cells $|xw| + 1, \dots, |xwy|$ up to time $|w|$. So, they are not affected by y up to that time. Furthermore, the cells $|x| + 1, \dots, |xy| + 1$ are not affected by the cells $1, \dots, |x|$, that is, not by x. On the other hand, different w may cause different configurations for x and y, where x and y are independent of each other.



Cellular Automata and Language Theory, Figure 7
Dependencies in the proof of Theorem 11

It follows that w and w' are equivalent, if the cells $1, \dots, |xy| + 1$ at time step $|w|$ respectively $|w'|$ are in the same states for all $x \in X$ and $y \in Y$. For any $y \in Y$, there are $|S|^{m_2} + 1$ different configurations for the cells $|x| + 1, \dots, |xy| + 1$. So, there are altogether at most $|S|^{(m_2+1)|Y|}$ different possibilities to distinguish the words w.

The possibilities for the cells $1, \dots, |x|$ are as follows. For a word $x = x_{m_1} \dots x_1$ the state s_i of cells $1 \le i \le m_1$ at time $|w|$ depends only on $x_i, \dots, x_1$ and w. So, for s_i there are at most $|S|^{|A|^i}$ different possibilities to distinguish the words w. Together we obtain

$$\prod_{i=1}^{m_1} |S|^{|A|^i} < |S|^{|A|^{m_1+1}}$$

possibilities. For $p = |S|^{|A|}$ this implies $|S|^{|A|^{m_1+1}} = p^{|A|^{m_1}} = p^{|X|}$. Thus, the number of equivalence classes is bounded by $p^{|X|} p^{(m_2+1)|Y|}$. □

As an example we show that the language $L_d \subset \{0, 1, (,), |\}^+$ whose words are of the form

$$x(x_1|y_1) \cdots (x_n|y_n)y ,$$

where $x,x_i,y,y_i \in \{0,1\}^*$, for $1 \le i \le$ n, and $(x|y) = (x_i|y_i)$ for at least one $i \in \{1, \dots, n\}$, is not a real-time OCA language.

It can be thought of as a dictionary. The task for the OCA is to check whether the pair $(x|y)$ occurs in the dictionary or not.

Example 12 Let $X = Y = \{0,1\}^*$. Two words $w = (x_1|y_1)\cdots(x_n|y_n)$ and $w' = (x'_1|y'_1)\cdots(x'_m|y'_m)$ are (L_d, X, Y)-equivalent if and only if $\{(x_1|y_1), \dots, (x_n|y_n)\} = \{(x'_1|y'_1), \dots, (x'_m|y'_m)\}$.

First assume that the two sets are equal. Let $x \in X$ and $y \in Y$, then $xwy \in L_d$ implies $(x|y) = (x_i|y_i)$, for some i. Since the two sets are equal, we have $(x|y) = (x'{}_j|y'{}_j)$, for some j. Therefore, $xwy \in L_d$ implies $xw'y \in L_d$ and vice versa, i. e., w and w' are (L_d, X, Y)-equivalent.

Now assume the two sets are different. Without loss of generality, we can assume that there exist $x \in X$ and $y \in Y$ with $(x|y) = (x_i|y_i)$, for some i, but $(x|y) \neq (x'_j|y'_j)$, for all $j = 1, \dots, m$. Then $xwy \in L_d$ but $xw'y \notin L_d$ and, thus, w and w' are not (L_d,X,Y)-equivalent.

In order to derive a lower bound for the number of (L_d, X, Y) equivalence classes induced by L_d, let $m_1, m_2 \geq 1$, $X = \{0,1\}^{m_1}$ and $Y = \{0,1\}^{m_2}$. Then the number N of (L_d,X,Y)-equivalence classes is at least $2^{2^{m_1+m_2}}$.

For fixed $m_1 \geq 1$ and $m_2 \geq 1$, we consider all words of the form $(x_1|y_1)\cdots(x_k|y_k)$ with $x_i \in \{0,1\}^{m_1}$ and $y_i \in \{0,1\}^{m_2}$, for all $i \in \{1, \dots, k\}$, and $(x_i|y_i) \neq (x_j|y_j)$, for $i \neq j$. These words are said to be of type (m_1,m_2). Following the argumentation above, two words are equivalent if and only if the sets of subwords are equal. There are $2^{2^{m_1+m_2}}$ words of type (m_1,m_2) which belong to different equivalence classes with respect to L_d, $X = \{0,1\}^{m_1}$ and $Y = \{0,1\}^{m_2}$.

Now assume that L_d is accepted by some real-time OCA. For all $m_1,m_2 \geq 1$ there is a constant p such that the number of equivalence classes is at most $p^{2^{m_1}} p^{(m_2+1)2^{m_2}}$. For large enough m_1 and m_2 we obtain a contradiction to the lower bound $2^{2^{m_1+m_2}}$. □

Helpful tools of a different nature are speed-up theorems. Strong results are obtained in [34,36], where the parallel language families are characterized by certain types of customized sequential machines. Such machines have been developed for all classes of acceptors which are here under consideration. In particular, speed-up theorems are given that allow to speed up the time beyond real time linearly. Therefore, linear-time computations can be sped up close to real time. The question whether or not real time is strictly weaker than linear time is discussed in detail later.

Theorem 13 *Let $\mathcal{M}$ be a CA, OCA, or IA obeying time complexity $rt + r(n)$, where $r\colon \mathbb{N} \to \mathbb{N}$ is a mapping and rt denotes real time. Then for all $k \geq 1$ an equivalent device $\mathcal{M}'$ of the same type obeying time complexity $rt + \lfloor \frac{r(n)}{k} \rfloor$ can effectively be constructed.*

The next two examples are frequently used applications of the linear speed-up theorem.

Example 14 Let $k_0 \geq 1$ and $\mathcal{M}$ be a device in question with time complexity $rt + k_0$. Then there is an equivalent real-time device $\mathcal{M}'$ of the same type. It suffices to set $k = k_0 + 1$ and to apply Theorem 13 in order to obtain $rt + \lfloor \frac{k_0}{k} \rfloor = rt + \lfloor \frac{k_0}{k_0+1} \rfloor = rt$ for the time complexity of $\mathcal{M}'$. □

Example 15 Let $k_0 \geq 1$ and $\mathcal{M}$ be a device in question with time complexity $rt + k_0 \cdot rt$. Then for all rational numbers $\varepsilon > 0$ there is an equivalent device $\mathcal{M}'$ of the same type with time complexity $\lfloor (1+\epsilon) \cdot rt \rfloor$. We set $k = \lceil \frac{k_0}{\epsilon} \rceil$ and apply Theorem 13 in order to obtain $rt + \left\lfloor \frac{k_0 \cdot rt}{\lceil k_0/\epsilon \rceil} \right\rfloor \leq rt + \left\lfloor \frac{k_0 \cdot rt}{k_0/\epsilon} \right\rfloor = rt + \lfloor \epsilon \cdot rt \rfloor = \lfloor (1+\epsilon) \cdot rt \rfloor$. □

Computational Capacities

In this section we explore the computational capacities of real-time, linear-time, and unbounded time devices. The goal is to establish a hierarchy of language families and to compare the levels with well-known linguistic families of the Chomsky hierarchy language family. The properness of some inclusions are long-standing open problems with deep relations to sequential complexity problems. In order to establish the hierarchy we start at the upper and lower end. Straightforward constructions of linearly space-bounded Turing machines from IAs, of IAs from CAs, and of CAs from linearly space-bounded Turing machines show the following lemma [80].

Lemma 16 *The families $\mathcal{L}(CA)$ and (IA) are identical with the deterministic context-sensitive languages, i. e., with the complexity class DSPACE(n).*

At the lower end we consider the regular languages. Since already the communication cell is a deterministic finite-state machine, clearly, the regular languages are a subset of $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. In Theorem 4 the stronger result that any real-time deterministic context-free language belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ has been shown. So, the properness of the following inclusion is obvious.

Corollary 17 *The regular languages are a proper subset of $\mathcal{L}_{rt}(IA)$.*

The question arises whether the real-time condition of Theorem 4 can be relaxed in order to obtain a larger sub-family of the context-free languages which is accepted by real-time IAs. The answer is negative. The following lemma marks a sharp boundary between context-free languages acceptable and non-acceptable by real-time IAs. Recall that linear context-free languages are accepted by nondeterministic one-turn pushdown automata. Since here we deal with deterministic devices, the alternative grammar characterization is more suitable for our purposes.

A grammar $\mathcal{G} = \langle N,T,S,P \rangle$ is said to be *linear context free*, if the productions in P are of the forms $(X \to a)$, $(X \to Ya)$, or $(X \to aY)$, where $X,Y \in N$ are nonterminals and $a \in T$ is a terminal symbol (e. g. [72]).

Lemma 18 *There exists a deterministic, linear context-free language that does not belong to $\mathcal{L}_{rt}$(IA).*

Proof Clearly, language

$$L = \{\&x_k\& \cdots \&x_1?y_1\& \cdots \&y_k\& \mid k \geq 1, x_i^R = y_i z_i \\ \text{and} \quad x_i, y_i, z_i \in \{a,b\}^*\}$$

is deterministic and linear context free. By Example 7 it does not belong to $\mathcal{L}_{\text{rt}}$(IA). □

We turn to real-time OCAs. Similar as above, the linear context-free languages serve as sub-family of the context-free languages which is contained in $\mathcal{L}_{\text{rt}}$(OCA) [76].

Theorem 19 *Given some linear context-free grammar, an equivalent real-time OCA can effectively be constructed, i. e., every linear context-free language belongs to $\mathcal{L}_{rt}$(OCA).*

Proof Let $\mathcal{G} = \langle N, T, S, P \rangle$ be a linear context-free grammar, and $w = a_1 a_2 \cdots a_n$ be a word in $L(\mathcal{G})$. If $X \Rightarrow^* a_i \cdots a_j$, $1 \leq i < j \leq n$, then there exists a $Y \in N$ such that either $(Y \Rightarrow^* a_i \cdots a_{j-1} \wedge X \Rightarrow Ya_j)$ or $(Y \Rightarrow^* a_{i+1} \cdots a_j \wedge X \Rightarrow a_i Y)$. Based on this fact we define sets of nonterminals for the given word w as follows:

$$N(i,i) = \{X \in N \mid (X \to a_i) \in P\}\,, \quad 1 \leq i \leq n$$
$$N(i,j) = N_1(i,j) \cup N_2(i,j)\,, \quad 1 \leq i < j \leq n\,,$$
$$\text{where } N_1(i,j) = \{X \in N \mid (X \to Ya_j) \in P \\ \wedge\ Y \in N(i,j-1)\}$$
$$\text{and } N_2(i,j) = \{X \in N \mid (X \to a_i Y) \in P \\ \wedge\ Y \in N(i+1,j)\}\,.$$

So, the word $a_1 \cdots a_n$ is generated by $\mathcal{G}$ if and only if $S \in N(1,n)$.

A real-time OCA $\mathcal{M} = \langle S', \delta, \#, T, F \rangle$ accepting $L(\mathcal{G})$ behaves as follows. It analyzes its input $a_1 \cdots a_n$ in such a way that the cells successively compute certain sets $N(i,j)$. In the first step, all cells $1 \leq i \leq n$ compute $N(i,i)$ and (a_i, a_i). In subsequent steps $k \geq 2$, they compute $N(i, i+k-1)$ and (a_i, a_{i+k-1}) if $k \leq n+1-i$, and keep their state otherwise. The new values can be determined by the state of the cell itself and the state of its right neighbor, i. e., $N(i, i+k-2)$, (a_i, a_{i+k-1}) and $N(i+1, i+k-1)$, (a_{i+1}, a_{i+k}). Thus, the leftmost cell computes $N(1,n)$ at time step n. □

Corollary 20 *The regular languages are a proper subset of $\mathcal{L}_{rt}$(OCA).*

Again, the question arises whether the condition of Theorem 19 can be relaxed in order to obtain a larger sub-family of the context-free languages that is accepted by real-time OCAs. Again, the answer is negative.

Lemma 21 *There exists a two-linear context-free language that does not belong to $\mathcal{L}_{rt}$(OCA).*

Proof Consider the linear language $L = \{a^n b^n \mid n \geq 1\} \cup \{a^n bvab^n \mid n \geq 1, v \in \{a,b\}^*\}$. By counting arguments, in [85] it is shown that the two-linear concatenation LL is not accepted by any real-time OCA. □

Beginning with the regular languages at the bottom of the hierarchy, next we deal with the proper supersets given by real-time deterministic and linear context-free languages. Since both families are known to be incomparable with respect to inclusion, the hierarchy splits into two strands. Unfortunately we have to continue with these two strands for a moment.

Theorem 22 *The families $\mathcal{L}_{rt}$(OCA) and $\mathcal{L}_{rt}$(IA) are incomparable.*

Proof By Lemma 18 there exists a linear context-free language not accepted by any real-time IA, but by Theorem 19 it belongs to a real-time OCA.

Conversely, one can show that the two-linear language of Lemma 21 belongs to $\mathcal{L}_{\text{rt}}$(IA). □

The next step is to go beyond $\mathcal{L}_{\text{rt}}$(OCA) and $\mathcal{L}_{\text{rt}}$(IA). Structurally this concerns CAs, but first we increase the time complexity. For the proof of infinite hierarchies in between real time and linear time, it is necessary to control the lengths of words with respect to some internal substructures. The following notion of constructibility expresses the idea that the length of a word relative to the length of a subword should be computable. To this end, a function $f\colon \mathbb{N} \to \mathbb{N}$ is said to be *OCA-constructible*, if there exist an λ-free homomorphism h and a language $L \in \mathcal{L}_{\text{rt}}$(OCA) such that $h(L) = \{a^{f(n)-n} b^n \mid n \geq 1\}$. Since constructible functions describe the length of the whole word dependent on the length of a subword, it is obvious that each constructible function must be greater than or equal to the identity. At a glance, this notion of constructibility might look somehow unusual or restrictive. However, λ-free homomorphisms are very powerful, so that the family of (in this sense) constructible functions is very rich. Examples and the next theorem are presented in [45].

Theorem 23 *Let $r_1, r_2 \colon \mathbb{N} \to \mathbb{N}$ be two increasing functions. If $r_2 \cdot \log(r_2) \in o(r_1)$ and r_1^{-1} is OCA constructible, then $\mathcal{L}_{n+r_2(n)}(OCA) \subset \mathcal{L}_{n+r_1(n)}(OCA)$.*

Example 24 Let $0 \leq p < q \leq 1$ be two rational numbers. Clearly, $n^p \cdot \log(n^p)$ is of order $o(n^q)$. Moreover, the inverse of n^q is OCA-constructible. Thus, an application of Theorem 23 yields the strict inclusion

$$\mathcal{L}_{n+n^p}(\mathrm{OCA}) \subset \mathcal{L}_{n+n^q}(\mathrm{OCA})\,. \qquad \square$$

Example 25 Let $i < j$ be two positive integers, then $\log^{[j]}(n) \cdot \log^{[j+1]}(n)$ is of order $o(\log^{[i]}(n))$. Since the inverse of $\log^{[i]}(n)$ is OCA-constructible, we obtain the strict inclusion

$$\mathcal{L}_{n+\log^{[j]}(n)}(\mathrm{OCA}) \subset \mathcal{L}_{n+\log^{[i]}(n)}(\mathrm{OCA})\,. \qquad \square$$

Similar results are known for iterative arrays. The infinite strict hierarchies in the range between real time and linear time are established in [9]. The constructibility is defined differently. A strictly increasing function $f \colon \mathbb{N} \to \mathbb{N}$ is *IA-constructible*, if there exists an IA with infinitely many cells to the right, such that the leftmost cell enters some state from a distinguished subset of states exactly at all time steps $f(i)$, $1 \leq i$. Example 1 shows that the function 2^n is IA-constructible.

Theorem 26 *Let $r_1, r_2 \colon \mathbb{N} \to \mathbb{N}$ be two increasing functions. If $r_2 \in o(r_1)$ and r_1^{-1} is IA-constructible, then $\mathcal{L}_{n+r_2(n)}(IA) \subset \mathcal{L}_{n+r_1(n)}(IA)$.*

The following example is based on natural functions.

Example 27 Since the family of IA-constructible functions is closed under composition and contains 2^n and n^k, $k \geq 1$, the functions $\log^{[i]}(n)$, $i \geq 1$, and $\sqrt[k]{n}$ are inverses of constructible functions. Actually, the inverses of 2^n and n^k are $\lceil \log(n) \rceil$ and $\lceil n^{\frac{1}{k}} \rceil$ but for convenience we simplify the notation. Therefore, an application to the hierarchy theorem yields

$$\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{n+\log^{[i+1]}(n)}(\mathrm{IA}) \subset \mathcal{L}_{n+\log^{[i]}(n)}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{\mathrm{lt}}(\mathrm{IA})$$

and

$$\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{n+n^{\frac{1}{i+1}}}(\mathrm{IA}) \subset \mathcal{L}_{n+n^{\frac{1}{i}}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{\mathrm{lt}}(\mathrm{IA})\,,$$

or in combination, e. g.,

$$\begin{aligned}\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) &\subset \cdots \subset \mathcal{L}_{n+(\log^{[j+1]}(n))^{\frac{1}{i+1}}}(\mathrm{IA}) \\ &\subset \mathcal{L}_{n+(\log^{[j+1]}(n))^{\frac{1}{i}}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{n+(\log^{[j]}(n))^{\frac{1}{i+1}}}(\mathrm{IA}) \\ &\subset \mathcal{L}_{n+(\log^{[j]}(n))^{\frac{1}{i}}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{\mathrm{lt}}(\mathrm{IA})\,.\end{aligned}$$

□

Example 9 reveals a unary language not accepted by any real-time OCA. Note that the witness language $\{a^{2^n} \mid n \geq 1\}$ is not context free. Next, we generalize this observation and show that even massively parallel real-time OCAs cannot accept more unary languages than a single deterministic finite-state machine [73].

Theorem 28 *Each unary real-time OCA language is regular.*

Proof Let $A = \{a\}$ be an alphabet, $L \subseteq A^+$ a language, and $\mathcal{M} = \langle S, \delta, \#, A, F \rangle$ be an OCA accepting L in real time. We construct an equivalent deterministic finite-state machine $\mathcal{E}$ with state set $S \times S$, initial state s_0, set of accepting states F', and transition function δ' as follows.

$$\begin{aligned}&s_0 = (\#, a)\,, \qquad F' = \{(s_1, s_2) \in S \times S \mid s_1 \in F\}\,,\\ &\quad \delta'((s_1, s_2), a) = (\delta(s_2, s_1), \delta(s_2, s_2))\,,\\ &\quad \text{for all } (s_1, s_2) \in S \times S\end{aligned}$$

In order to give evidence that the construction is correct, we depict two correspondent computations in Fig. 8. □

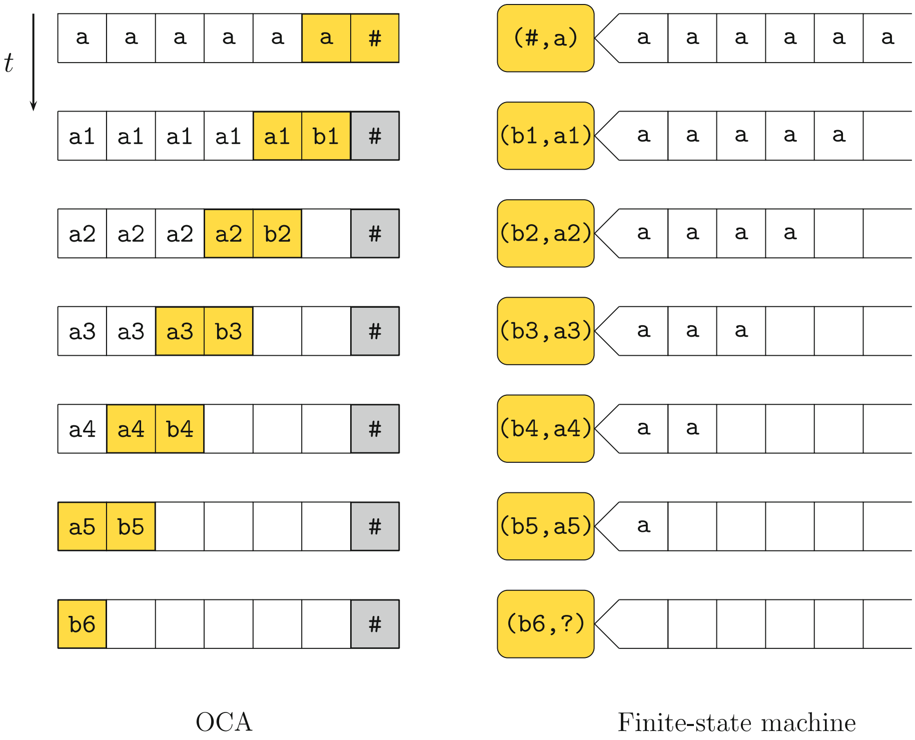
Next we can join the two strands of the hierarchy again. The superfamily is $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$.

Theorem 29 *The family $\mathcal{L}_{rt}(OCA)$ is properly included in $\mathcal{L}_{rt}(CA)$.*

Proof The inclusion follows for structural reasons. For the properness we argue as follows. Since the language $L = \{a^{2^n} \mid n \geq 1\}$ is not regular, it does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. On the other hand, Example 1 shows that it belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$. □

Theorem 30 *The family $\mathcal{L}_{rt}(IA)$ is properly included in $\mathcal{L}_{rt}(CA)$.*

Proof First we give evidence of the inclusion $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) \subseteq \mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$. A real-time CA can be set up in such a way that it shifts its input successively to the left on an additional track. So, the leftmost cell receives the input symbol by symbol. Therefore, the CA can simulate the iterative array, where the leftmost cell simulates the communication cell.



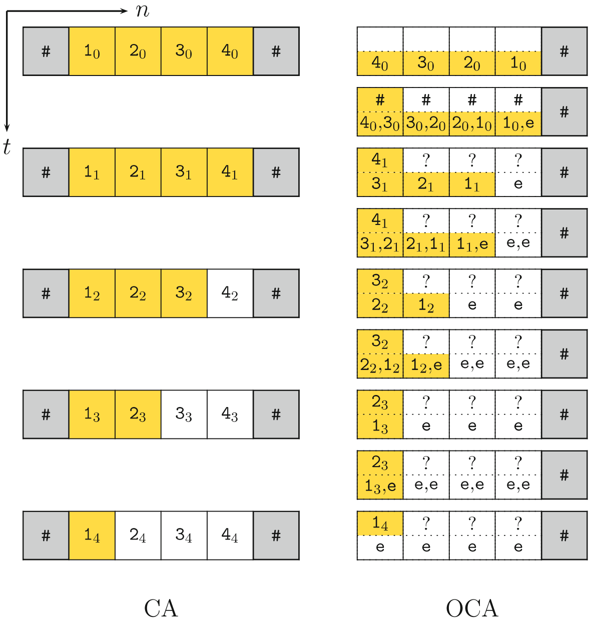
Cellular Automata and Language Theory, Figure 8
A finite-state machine simulating a real-time OCA on unary input

The properness of the inclusion follows by the inclusion $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA}) \subset \mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ and the incomparability of $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ and $\mathcal{L}_{\mathrm{rt}}(IA)$. □

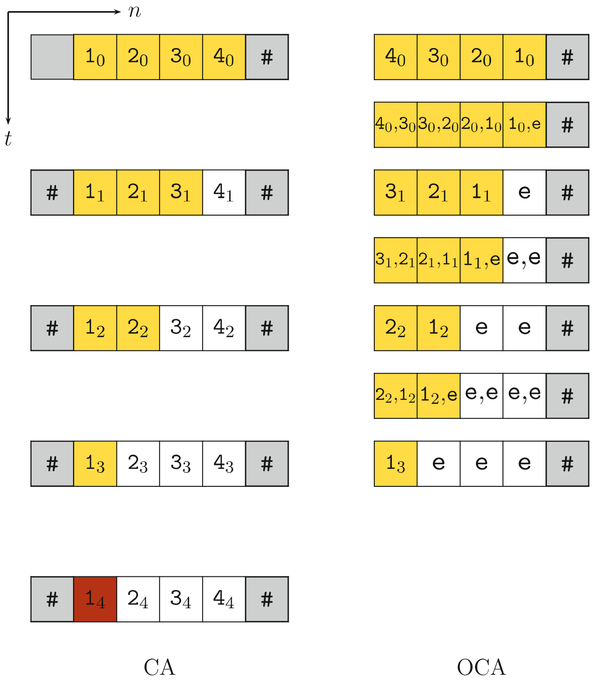
Once we know that, in general, a real-time CA language cannot be accepted by any real-time OCA, the question arises of how much time is necessary for that, if possible at all. The next result gives an upper bound for the time [17,95]. Admittedly, to this end the input has to be reversed. Alternatively, one could reverse the neighborhood of the cells in an OCA. Then the rightmost cell indicates the result of the computation. In this case the input could remain as it is. In any case, the condition cannot be relaxed since it is an open problem whether the corresponding language families are closed under reversal.

Theorem 31 *A language is accepted by a linear-time OCA if and only if its reversal is accepted by a CA in real time.*

Proof Let $\mathcal{M}$ be a real-time CA. The cells of a linear-time OCA $\mathcal{M}'$ accepting $L(\mathcal{M})^R$ collect the information necessary to simulate one transition of $\mathcal{M}$, in an intermediate step. Therefore, the first step of $\mathcal{M}$ is simulated in the second step of $\mathcal{M}'$. We obtain a behavior as depicted in Fig. 9.



Cellular Automata and Language Theory, Figure 9
Intermediate steps in the construction of the proof of Theorem 31



Cellular Automata and Language Theory, Figure 10
Example of a linear-time OCA simulation of a real-time CA computation on reversed input

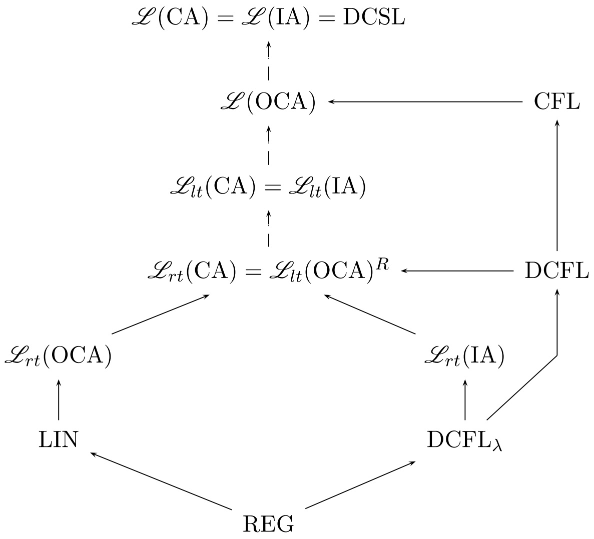
Altogether, $\mathcal{M}'$ cannot simulate the last step of $\mathcal{M}$. So, the construction has to be extended slightly. Each cell has an extra register that is used to simulate transitions of $\mathcal{M}$ under the assumption that the cell is the leftmost one (see Fig. 10). The transitions of the real leftmost cell now correspond to the missing transitions of the previous simulation. □

Climbing up one level, the hierarchy continues with two-way linear-time devices.

Theorem 32 *A language is accepted by a linear-time IA if and only if it is accepted by a linear-time CA.*

Proof The inclusion $\mathcal{L}_{\mathrm{lt}}(\mathrm{IA}) \subseteq \mathcal{L}_{\mathrm{lt}}(\mathrm{CA})$ follows by the proof of Theorem 30.

Conversely, a linear-time iterative array can simulate a linear-time CA as follows. In the first phase, it reads the input and stores it successively in its cells. Next, the iterative array starts a time optimal firing squad synchronization algorithm (see for example [62,98]). That is, an algorithm which is initiated by the communication cell, and which synchronizes the n cells within $2n - 2$ time steps. Finally, all cells start the simulation of the CA at the same time. Clearly, the iterative array obeys a linear time bound if the cellular automaton does. □


$\mathscr{L}(\mathrm{CA}) = \mathscr{L}(\mathrm{IA}) = \mathrm{DCSL}$
$\mathscr{L}(\mathrm{OCA})$
CFL
$\mathscr{L}_{lt}(\mathrm{CA}) = \mathscr{L}_{lt}(\mathrm{IA})$
$\mathscr{L}_{rt}(\mathrm{CA}) = \mathscr{L}_{lt}(\mathrm{OCA})^R$
DCFL
$\mathscr{L}_{rt}(\mathrm{OCA})$
$\mathscr{L}_{rt}(\mathrm{IA})$
LIN
DCFL_λ
REG

Cellular Automata and Language Theory, Figure 11
Basic hierarchy of language families. A solid arrow indicates a proper inclusion and a dashed arrow an inclusion. In addition, the linear languages (LIN) are properly included in the context-free languages (CFL). Deterministic and real-time deterministic context-free languages are denoted by DCFL and DCFL_λ, regular languages by REG, and deterministic context-sensitive languages by DCSL

The next inclusion does not follow for structural reasons. It is proved in [16,32] in terms of simulations of equivalent sequential machines. The properness is an open problem.

Theorem 33 *Each linear-time CA language belongs to the family $\mathcal{L}(OCA)$.*

The family $\mathcal{L}(\mathrm{OCA})$ is very powerful. It contains the context-free languages as well as a PSPACE-complete language [16,32]. Altogether we obtained the hierarchy depicted in Fig. 11, where the only known proper inclusions are at the lower end. Nevertheless, even the real-time OCA languages contain important families, e. g., the Dyck languages [73] and the bracketed context-free languages [26]. Furthermore, the non-semilinear language $\{(a^i b)^* \mid i \geq 0\}$ [73] and the inherently ambiguous language $\{a^i b^j c^k \mid i = j \text{ or } j = k \text{ for } i, j, k \geq 1\}$ [80] belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. On the other hand, the context-free languages have been shown to be incomparable with $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. Whether or not they are a subset of the family $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ is an open question raised in [80].

We conclude this section with a real-time OCA based characterization of the recursively enumerable languages and its implication to incomparability [58].

Lemma 34 *A language is recursively enumerable if and only if there exists a homomorphism h and a language $L' \in \mathcal{L}_{\mathrm{rt}}(OCA)$ such that $L = h(L')$. The same holds for $\mathcal{L}_{\mathrm{rt}}(IA)$.*

Proof It is known that the recursively enumerable languages can be represented by $h(L_1 \cap L_2)$, where h is a homomorphism and L_1, L_2 are either linear context-free languages [4] or real-time deterministic context-free languages [29]. In fact, in the latter paper deterministic pushdown automata are constructed that use λ-moves. But the constructions can easily be modified to obtain real-time pushdown automata. So, the assertions follow since $\mathcal{L}_{\text{rt}}(\text{OCA})$ contains the linear context-free languages, $\mathcal{L}_{\text{rt}}(\text{IA})$ contains the real-time deterministic context-free languages, and both families are closed under intersection (Theorem 39). □

The homomorphic characterization of the recursively enumerable languages reveals the following general incomparability result.

Theorem 35 *Both families $\mathcal{L}_{rt}(OCA)$ and $\mathcal{L}_{rt}(IA)$ are incomparable to each language family $\mathcal{L}$ that contains the context-free languages, is itself properly contained in the recursively enumerable languages, and is closed under homomorphism.*

Proof Since there is a context-free language not belonging to $\mathcal{L}_{\text{rt}}(\text{OCA})$, we obtain $\mathcal{L} \not\subseteq \mathcal{L}_{\text{rt}}(\text{OCA})$. Conversely, if $\mathcal{L}_{\text{rt}}(\text{OCA}) \subseteq \mathcal{L}$, then the recursively enumerable languages are a subset of $\mathcal{L}$ by Lemma 34. This contradicts the proper containment in the recursively enumerable languages. The same argumentation applies to $\mathcal{L}_{\text{rt}}(\text{IA})$. □

Example 36 The families $\mathcal{L}_{\text{rt}}(\text{OCA})$ and $\mathcal{L}_{\text{rt}}(\text{IA})$ are incomparable to the languages of indexed grammars, certain grammars with regulated rewriting, certain contextual grammars, and certain Lindenmayer systems, e. g., ET0L systems. □

Closure Properties

Closure properties of families of formal languages indicate their robustness under certain operations. A family of languages is *closed under some operation*, if any application of the operation on languages from the family yields again a language from the family. It is *effectively closed* if the result of the operation can be constructed from the given language(s). The knowledge of closure properties often reveals insights in particular features, structures, and capabilities. Moreover, positive closure properties are a useful tool for modular constructions and decompositions in a natural way. Negative properties may serve, for example, as valuable basis for extensions. That is, for building the smallest family of languages which is closed under the operation and contains the family in question. In any case closure properties are filigree tools for dealing with language families.

Boolean Operations

The operations union, intersection, complementation and set difference are commonly called *Boolean operations* or *elementary operations*.

For deterministic devices, the closure under complementation is often shown by interchanging accepting and non-accepting states. But, in general, this requires halting computations.

Theorem 37 *For $X \in \{CA, OCA, IA\}$, the families $\mathcal{L}_{rt}(X)$ and $\mathcal{L}_{lt}(X)$ are effectively closed under complementation.*

Proof We show the closure exemplarily for linear-time OCAs. The other proofs are similar. The reason that the complementary device cannot be constructed by simply interchanging accepting and non-accepting states is that an input is accepted when the leftmost cell enters an accepting state at some arbitrary time step. So, in general, the leftmost cell will enter accepting as well as non-accepting states during a computation. To cope with this problem we modify a given linear-time OCA $\mathcal{M}$ in the following way. At initial time a signal is emitted by the rightmost cell. The signal moves on an extra track with speed $\frac{1}{2}$ to the left. It will arrive at the leftmost cell at twice real time, i. e., linear time. This is the time step at which we wish to make the final decision whether to accept or to reject the input. To this end, the leftmost cell has to remember if it has entered an accepting state at some time before. So, we use a copy S' of the state set S of $\mathcal{M}$, and modify the local transition function to drive the leftmost cell into a state of S' when it enters an accepting state. Subsequently, the normal behavior of the leftmost cell is simulated, except that states of S' are used instead of states of S. Now the modified automaton $\mathcal{M}'$ accepts input if and only if the leftmost cell is in some state of S' when the signal arrives. In order to accept the complement of $L(\mathcal{M}) = L(\mathcal{M}')$, it suffices to let the automaton accept input if and only if the leftmost cell is in some state of S when the signal arrives. □

Devices without any time bounds are, in fact, exponentially time bounded due to the space limitation.

Lemma 38 *For $X \in$ {CA, OCA, IA}, the families $\mathcal{L}(X)$ are effectively closed under complementation.*

Proof The construction of Theorem 37 is modified as follows. Some device in question with n cells and state set S may run through at most $|S|^n$ different configurations until its behavior becomes cyclic. So, it suffices to check whether the input is accepted during the first $|S|^n$ time steps. As discussed briefly in Sect. "Cellular Language Acceptors", exponential time complexities are honest. That is, the leftmost cell can recognize the time step $|S|^n$. To this end, an

$|S|$ary counter is set up on an extra track. The rightmost cell simulates the least significant digit and adds one to the counter at every time step. The neighboring cell to the left observes when a carry-over appears, increases its own digit and so on. The time step at which the first carry-over appears in the leftmost cell is the desired one. □

Now we turn to intersection, union, and set difference.

Theorem 39 *For $X \in \{CA, OCA, IA\}$, the families $\mathscr{L}_{rt}(X)$, $\mathscr{L}_{lt}(X)$, and (X) are effectively closed under intersection, union, and set difference.*

Proof In order to prove the closures, the well-known two-track technique is applicable. That is, on two different tracks acceptors for the languages in question are simulated independently of each other. By the same construction as in the previous proof, the leftmost cell can remember whether the single tracks accept or not. The accepting states are composed by the accepting and non-accepting components for the single tracks in the usual way. For example, to show intersection both components have to be accepting ones. □

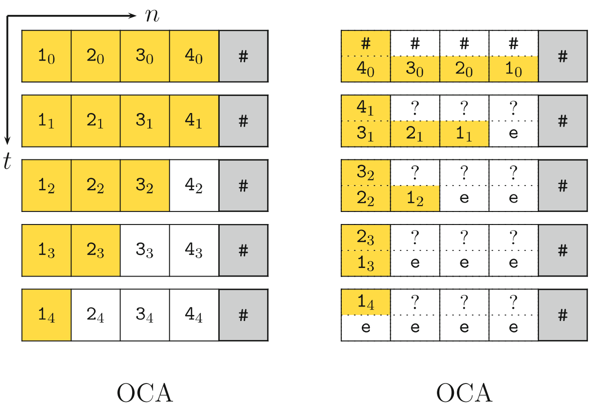
Reversal

The closure under reversal is of crucial importance. It is an open problem for $\mathscr{L}_{\mathrm{rt}}(\mathrm{CA})$ and, equivalently, for $\mathscr{L}_{\mathrm{lt}}(\mathrm{OCA})$. Moreover, it is linked with the open closure property under concatenation for the same family and, hence, with the question whether linear-time CAs are more powerful than real-time CAs. So, it remains open whether the computational capacities of CAs differ if the rightmost or the leftmost cell indicates acceptance.

Theorem 40 *The family $\mathscr{L}_{rt}(OCA)$ is effectively closed under reversal.*

Proof Let $\mathcal{M}$ be some real-time OCA. In order to obtain a real-time OCA $\mathcal{M}'$ for the language $L(\mathcal{M})^R$, the arguments of the local transition function are interchanged. That is, δ’$(s_2, s_1) = s_3$ if $\delta(s_1, s_2) = s_3$. In addition, we have to pay special attention to the boundary state. The further construction is as in the proof of Theorem 31, with the exception that we do not need to collect information in intermediate steps, since the information flow is one-way in both devices. (see Fig. 12). □

As mentioned above, the closure under reversal can be interpreted in two different ways. On one hand, one can construct an OCA that accepts the reversal of a given OCA language. On the other hand, one can construct an OCA that accepts the same language with the rightmost cell. The latter point of view yields immediately the closure under reversal of two-way linear-time cellular automata.



Cellular Automata and Language Theory, Figure 12
Construction showing the closure of real-time OCA languages under reversal

Theorem 41 *The family $\mathscr{L}_{lt}(CA) = \mathscr{L}_{lt}(IA)$ is effectively closed under reversal.*

Proof Let $\mathcal{M}$ be some linear-time CA. In order to obtain a linear-time CA $\mathcal{M}'$ for the language $L(\mathcal{M})^R$, the first and third arguments of the local transition function are interchanged. That is, δ’$(s_3, s_2, s_1) = s_4$ if $\delta(s_1, s_2, s_3) = s_4$. The resulting device accepts $L(\mathcal{M})^R$ with the rightmost cell. Then the result is sent as a signal to the leftmost cell. Altogether, $\mathcal{M}'$ still obeys a linear time bound. □

The closure under reversal of the devices without time bounds follows from the known closures of the characterizing linguistic language family, i. e., the deterministic context-sensitive languages DSPACE(n).

Corollary 42 *The family $\mathscr{L}(CA) = \mathscr{L}(IA)$ is effectively closed under reversal.*

In [16,32] unbounded time OCAs are simulated by a variant of unbounded time one-way iterative arrays and vice versa. Moreover, it is shown that the family of accepted languages forms an AFL, i. e., an *abstract family of languages*, (e. g. [72]) which is in addition closed under reversal.

Lemma 43 *The family $\mathscr{L}(OCA)$ is effectively closed under reversal.*

In order to show negative closure properties it is sometimes convenient to have witness languages not belonging to the family in question. By Example 7 the language

$$L = \{\&x_k\& \cdots \&x_1 ? y_1 \& \cdots \& y_k \& \mid k \geq 1, x_i^R = y_i z_i \\ \text{and} \quad x_i, y_i, z_i \in \{a, b\}^*\}$$

is known not to belong to $\mathscr{L}_{\mathrm{rt}}(\mathrm{IA})$.

Theorem 44 *The family $\mathcal{L}_{rt}(IA)$ is not closed under reversal.*

Proof It suffices to show that L^R belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. To this end, a real-time deterministic pushdown automaton accepting L^R is easily constructed. Then by Theorem 4 the containment $L^R \in \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ follows. □

If the answer to the open reversal closure of $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ is negative, we have to deal with two different language families. Since the properness of the inclusion $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA}) \subseteq \mathcal{L}_{\mathrm{lt}}(\mathrm{CA})$ is also open, the problem gains in importance. A negative answer of the former problem would imply a proper inclusion. A language $L \in \mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ whose reversal does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ may serve as witness since $\mathcal{L}_{\mathrm{lt}}(\mathrm{CA})$ is closed under reversal by Theorem 41. In fact, the following stronger relation is shown in [33] by a long proof.

Theorem 45 *The family $\mathcal{L}_{rt}(CA)$ is closed under reversal if and only if $\mathcal{L}_{rt}(CA)$ and $\mathcal{L}_{lt}(CA)$ are identical.*

Concatenation

Concerning the closure properties under concatenation, the situation is similar to reversal. Either the properties are trivial due to the characterizations by well-known language families, or they are negative, or open problems. For devices without time bounds we have, again, the closures of the deterministic context-sensitive languages DSPACE(n).

Corollary 46 *The family $\mathcal{L}(CA) = \mathcal{L}(IA)$ is effectively closed under concatenation.*

Since any AFL is closed under concatenation and, as mentioned before, $\mathcal{L}(\mathrm{OCA})$ is an AFL [16,32] whose closure properties are shown by simulations, the next lemma follows immediately.

Lemma 47 *The family $\mathcal{L}(OCA)$ is effectively closed under concatenation.*

In order to show that $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ is not closed under concatenation, once more the witness is

$$L = \{\&x_k\& \cdots \&x_1?y_1\& \cdots \&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } x_i, y_i, z_i \in \{a,b\}^*\} .$$

Theorem 48 *The family $\mathcal{L}_{rt}(IA)$ is not closed under concatenation.*

Proof In contrast to the assertion, assume that $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ is closed under concatenation. The language

$$L_1 = \{\&x(\&\{a,b\}^+)^k ?\&^k y\& \mid x^R = yz,\ x, y, z \in \{a,b\}^*\}$$

is clearly a real-time deterministic context-free and, thus, a real-time IA language. The language

$$L_2 = (\&\{a,b\}^+)^*$$

is regular and, therefore, accepted by some real-time IA, too. Due to the assumption, the language

$$L_3 = L_2 L_1 \&^*$$

belongs also to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. Next, the language

$$L_4 = \{(\&\{a,b\}^+)^k ?(\{a,b\}^* \&)^k \mid k \geq 1\}$$

is a real-time deterministic context-free language and therefore accepted by some real-time IA. Finally, from the closure under intersection we obtain

$$L_5 = L_3 \cap L_4 \in \mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) .$$

However, the language $L_5 \subset L$ contains the words used to show $L \notin \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. We conclude $L_5 \notin \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$, a contradiction. □

The question whether or not the family $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ is closed under concatenation was open for a long time. It has been solved negatively in [84]. Here we utilize the language L_d of Example 12. Recall that $L_d \subset \{0,1,(,),|\}^+$ is the language whose words are of the form

$$x(x_1|y_1)\cdots(x_n|y_n)y ,$$

where $x, x_i, y, y_i \in \{0,1\}^*$, for $1 \leq i \leq n$, and $(x|y) = (x_i|y_i)$, for at least one $i \in \{1,\ldots,n\}$.

The following example will be helpful.

Example 49 The language $L_c = \{w \bullet w \mid w \in \{0,1\}^+\}$ is a real-time OCA language.

An acceptor for L_c has several tracks. On one track the input is preserved. On two different tracks the input is successively shifted to the left. The shifting is in such a way that a symbol moves to the left one cell per time step until it passes through the center cell with input •. Subsequently, it moves to the left every other time step. In order to achieve this slow-down, the second track is used. Symbols are received in the first register, then shifted to the second one and, finally, send to the left neighbor. In addition, at initial time a signal is emitted at the right border. When the signal has passed through the center cell, it starts to compare the original input symbol with the symbol to be shifted out of the cell next.

Let the input be $a_m \cdots a_1 \bullet a'_m \cdots a'_1$. At time step $m+1+i$, $1 \leq i \leq m$, the signal arrives in the cell with input a_i. The symbol $a'{}_i$ takes $m+1-i$ time steps to arrive at the center, and is to be shifted out of the cell carrying the original input a_i at time $m+1-i+2i$. So, the signal compares a'_i with a_i as required. □

Theorem 50 *The family $\mathscr{L}_{rt}(OCA)$ is not closed under concatenation.*

Proof Consider the language L_1 whose words are of the form

$$x(x_1|y_1)\cdots(x_n|y_n)(x| ,$$

where $x, x_i, y_i \in \{0,1\}^*$, for $1 \le i \le n$.

Similar to Example 49 we obtain a simple construction of a real-time OCA accepting L_1. For symmetry reasons the language L_2 whose words are of the form

$$y)(x_1|y_1)\cdots(x_n|y_n)y ,$$

where $y, x_i, y_i \in \{0,1\}^*$, for $1 \le i \le n$, is a real-time OCA language, too.

However, the concatenation L_1L_2 equals L_d, which does not belong to $\mathscr{L}_{\mathrm{rt}}(\mathrm{OCA})$. □

The question whether or not one of the families $\mathscr{L}_{\mathrm{rt}}(\mathrm{CA}) = \mathscr{L}_{\mathrm{lt}}^{R}(\mathrm{OCA})$ or $\mathscr{L}_{\mathrm{lt}}(\mathrm{CA}) = \mathscr{L}_{\mathrm{lt}}(\mathrm{IA})$ is closed under concatenation is another famous open problem in this field. Nevertheless, it is shown in [33] that the closure of $\mathscr{L}_{\mathrm{rt}}(\mathrm{CA})$ under reversal implies its closure under concatenation. Since in this case we obtain $\mathscr{L}_{\mathrm{rt}}(\mathrm{CA}) = \mathscr{L}_{\mathrm{lt}}(\mathrm{CA})$, the family of linear-time CA languages were also closed under concatenation.

Theorem 51 *If the family $\mathscr{L}_{rt}(CA)$ is closed under reversal, then it is closed under concatenation.*

Proof If $\mathscr{L}_{\mathrm{rt}}(\mathrm{CA})$ is closed under reversal, then by Theorem 45 we have the identity $\mathscr{L}_{\mathrm{rt}}(\mathrm{CA}) = \mathscr{L}_{\mathrm{lt}}(\mathrm{CA})$. So, it suffices to construct a linear-time CA $\mathcal{M}$ for the concatenation of two real-time CA languages L_1 and L_2. By the closure under reversal, Theorem 31, and the speed-up theorems, there are $2n$-time OCAs $\mathcal{M}_1$ and $\mathcal{M}_2$ for L_1^R and L_2.

During a first phase, automaton $\mathcal{M}$ reverses its input, say $a_1 \cdots a_n$, on an extra track. This takes n time steps.

During a second phase, automaton $\mathcal{M}$ simulates automaton $\mathcal{M}_1$ on $a_n \cdots a_1$ and $\mathcal{M}_2$ on $a_1 \cdots a_n$ in parallel. When some cell enters an accepting state during the simulations, the cell is marked on the corresponding track. If a cell at position $n+1-i$ carrying the input symbol a_i is marked by the simulation of $\mathcal{M}_1$, we have $a_i \cdots a_1 \in L_1^R$ and, thus, $a_1 \cdots a_i \in L_1$. If a cell at position i carrying the input symbol a_i is marked by the simulation of $\mathcal{M}_2$, we have $a_i \cdots a_n \in L_2$. So, the input $a_1 \cdots a_n$ belongs to the concatenation $L_1\,L_2$ if and only if $\mathcal{M}_1$ marks a cell at position $n+1-i$ and $\mathcal{M}_2$ a cell at position $i+1$, for $1 \le i < n$. In order to check this condition, $\mathcal{M}$ reverses the result of the simulation of $\mathcal{M}_1$. Now a cell at position i is marked if and only if previously a cell at position $n+1-i$ was marked. Therefore, it suffices to verify by a signal whether two adjacent cells are marked. □

Cellular Automata and Language Theory, Table 1
Summary of closure properties. Concatenation REG denotes the concatenation with regular languages at the right, REG concatenation at the left, *hom* denotes homomorphisms, *gsm* generalized sequential machine mappings, and inj. length-pres. abbreviates injective length-preserving. A + indicates closure, a — non-closure, and a question mark an open problem

	$\mathscr{L}_{\mathrm{rt}}$ (OCA)	$\mathscr{L}_{\mathrm{rt}}$ (IA)	$\mathscr{L}_{\mathrm{rt}}$ (CA)	$\mathscr{L}_{\mathrm{lt}}$ (CA)	$\mathscr{L}$ (OCA)	$\mathscr{L}$ (CA)
∪, ∩	+	+	+	+	+	+
complementation, —	+	+	+	+	+	+
reversal	+	—	?	+	+	+
concatenation	—	—	?	?	+	+
λ-free iteration	—	—	?	?	+	+
concatenation REG	+	+	+	?	+	+
REG concatenation	+	—	?	?	+	+
marked concatenation	+	+	+	+	+	+
marked λ-free iteration	+	+	+	+	+	+
hom^{-1}	+	+	+	+	+	+
deterministic *gsm*$^{-1}$	+	+	+	+	+	+
gsm^{-1}	—	?	?	?	+	+
inj. length-pres. *hom*	+	+	+	+	L_2	+
λ-free *hom*	—	—	?	?	+	+
λ-free *gsm*	—	—	?	?	+	+
λ-free substitution	—	—	?	?	+	+
hom	—	—	—	—	—	—

The concatenation closure for *unary* real-time CA languages has been solved in the affirmative [33].

Table 1 summarizes some closure properties of the language families in question.

Decidability Problems

It is well known that all nontrivial decidability problems for Turing machines are undecidable [69]. Moreover, many of them are not even semidecidable, e. g., neither finiteness nor infiniteness. Now we turn to explore undecidable properties for cellular automata and iterative arrays. Most of the early results are shown in [73] by reductions of the the Post Correspondence Problem. In terms of trellis automata the undecidability of emptiness, equivalence, and universality is derived in [22]. Here we present improved results that show the *non-semidecidability* of the properties. Almost all results in this section are obtained by Andreas Malcher in [58,59].

In [30] large Turing machine computations have been encoded in small grammars. These encodings and variants thereof are of tangible advantage for our purposes. To this end, we consider *valid computations of Turing machines*. Roughly speaking, these are histories of accepting Turing machine computations. It suffices to consider deterministic Turing machines with a single tape and a single read-write head. Without loss of generality and for technical reasons, one can assume that any accepting computation has at least three and, in general, an odd number of steps. Therefore, it is represented by an even number of configurations. Moreover, it is assumed that the Turing machine cannot print blanks, and that a configuration is halting if and only if it is accepting.

Let S be the state set of some Turing machine $\mathcal{M}$, where s_0 is the initial state, $T \cap S = \emptyset$ is the tape alphabet containing the blank symbol, $A \subset T$ is the input alphabet, and $F \subseteq S$ is the set of accepting states. Then a configuration of $\mathcal{M}$ can be written as a word of the form T^*ST^* such that $t_1 \cdots t_i\, s\, t_{i+1} \cdots t_n$ is used to express that $\mathcal{M}$ is in state s, scanning tape symbol t_{i+1}, and t_1 to t_n is the support of the tape inscription. The set of valid computations VALC($\mathcal{M}$) is now defined to be the set of words of the form $w_1 \$ w_3 \$ \cdots \$ w_{2k-1} ¢\, w_{2k}^R \$ \cdots \$ w_4^R \$ w_2^R$, where w_i are configurations, \$ and ¢ are symbols not appearing in w_i, w_1 is an initial configuration of the form s_0A^*, w_{2k} is an accepting configuration of the form $T^*\ FT^*$, and w_{i+1} is the successor configuration of w_i, for $1 \le i \le 2k$. The set of *invalid computations* INVALC($\mathcal{M}$) is the complement of VALC($\mathcal{M}$) with respect to the coding alphabet $\{\$,¢\} \cup T \cup S$. The following lemma shows some of the important properties of valid computations.

Lemma 52 *Let $\mathcal{M}$ be some Turing machine.*

1. *$L(\mathcal{M})$ is empty if and only if VALC($\mathcal{M}$) is empty.*
2. *$L(\mathcal{M})$ is finite if and only if VALC($\mathcal{M}$) is finite.*
3. *$L(\mathcal{M})$ is finite if and only if VALC($\mathcal{M}$) is context free.*
4. *$L(\mathcal{M})$ is finite if and only if INVALC($\mathcal{M}$) is regular.*
5. *VALC($\mathcal{M}$) can be represented by the intersection of two real-time deterministic, linear context-free languages, such that both deterministic pushdown automata and both linear context-free grammars can effectively be constructed from $\mathcal{M}$.*
6. *INVALC($\mathcal{M}$) is a linear context-free language, such that its grammar can effectively be constructed from $\mathcal{M}$.*

Proof Assertions 1 and 2 are immediate observations. In order to show assertion 3 assume that $L(\mathcal{M})$ is finite. Then VALC($\mathcal{M}$) is finite and, clearly, context free. If conversely $L(\mathcal{M})$ is infinite, then an application of the pumping lemma shows that VALC($\mathcal{M}$) is not context free [30]. Now, assertion 4 is shown as follows. If $L(\mathcal{M})$ is finite, then VALC($\mathcal{M}$) is finite. Therefore, INVALC($\mathcal{M}$) is co-finite and, thus, regular. Conversely, if INVALC($\mathcal{M}$) is regular, then VALC($\mathcal{M}$) is regular since the regular languages are closed under complementation. Therefore, VALC($\mathcal{M}$) is context free which implies that $L(\mathcal{M})$ is finite. The two deterministic, linear context-free languages for assertion 5 are constructed in [4]. Assertion 6 has been shown in [30] for a similar definition of invalid computations. The proof can easily be adapted. □

Lemma 53 *Let $\mathcal{M}$ be a Turing machine. Then both languages VALC($\mathcal{M}$) and INVALC($\mathcal{M}$) are accepted by real-time OCAs as well as by real-time IAs.*

Proof Given some Turing machine $\mathcal{M}$, by Lemma 52 item 6 we can effectively construct a linear context-free grammar for the language INVALC($\mathcal{M}$). Due to Theorem 19, a real-time OCA accepting INVALC($\mathcal{M}$) can be constructed from the grammar. Since $\mathcal{L}_{rt}$(OCA) is effectively closed under complementation, we obtain a real-time OCA that accepts the valid computations of $\mathcal{M}$.

Similarly, by Lemma 52 item 5 we can effectively construct two real-time deterministic pushdown automata whose intersection represents the language VALC($\mathcal{M}$). Due to Theorem 4, two real-time IAs can be constructed from the pushdown automata. Since the family $\mathcal{L}_{rt}$(IA) is effectively closed under intersection and complementation, we obtain real-time IAs that accept the valid and the invalid computations of $\mathcal{M}$. □

Now we are prepared to reduce the finiteness and infiniteness problems of Turing machines to some of the decidability problems in question.

Theorem 54 *For any language family that effectively contains $\mathcal{L}_{rt}$(OCA) or $\mathcal{L}_{rt}$(IA), emptiness, universality, finiteness, infiniteness, context-freeness, and regularity are not semidecidable.*

Proof Given some Turing machine $\mathcal{M}$, by Lemma 53 we can effectively construct a real-time OCA and a real-time IA for the language VALC($\mathcal{M}$). Therefore, by Lemma 51 item 1, if emptiness were semidecidable for real-time OCAs or IAs, then emptiness is semidecidable for Turing machines, too.

Since $L(\mathcal{M}) = A^*$ if and only if the complement of $L(\mathcal{M})$ is empty, the non-semidecidability of universality follows from the effective closure under complementation and the non-semidecidability of emptiness.

In the same way as emptiness and universality the non-semidecidability of finiteness and infiniteness follows.

Since we have constructed real-time OCAs and IAs for INVALC($\mathcal{M}$) as well as for VALC($\mathcal{M}$), the finiteness problem for Turing machines immediately reduces to the context-freeness and to the regularity problem for $\mathcal{L}_{\mathrm{rt}}$(OCA) and $\mathcal{L}_{\mathrm{rt}}$(IA) by Lemma 52 items 3 and 4. □

Theorem 55 *For any language family that effectively contains $\mathcal{L}_{rt}$(OCA) or $\mathcal{L}_{rt}$(IA) equivalence and inclusion are not semidecidable.*

Proof It is easy to construct a real-time OCA and a real-time IA that accept the empty language. So, the semidecidability of equivalence would imply the semidecidability of emptiness. Since $L(\mathcal{M}) = L(\mathcal{M}')$ if and only if $L(\mathcal{M}) \subseteq L(\mathcal{M}')$ and $L(\mathcal{M}') \subseteq L(\mathcal{M})$, inclusion is not semidecidable either. □

Next the question arises whether some structural properties of cellular language acceptors are (semi)decidable. For example, whether or not a real-time two-way language is a real-time one-way language. We also compare sequential input mode and two-way information flow with parallel input mode and one-way information flow from a decidability point of view. The questions turn out to be not even semidecidable. So the resources inherent in the structures of cellular automata seem to be fairly different.

Let $\mathcal{M}$ be some Turing machine. We consider the language

$$L_{\mathcal{M}} = \{w^{|w|!} \mid w \in \triangleright\mathrm{VALC}(\mathcal{M})\triangleleft\}\,,$$

where $\triangleright$ and $\triangleleft$ are new symbols not appearing in VALC($\mathcal{M}$), and deal with its acceptance.

Lemma 56 *Given some Turing machine $\mathcal{M}$, a real-time IA accepting $L_{\mathcal{M}}$ can effectively be constructed from $\mathcal{M}$, i. e., the language $L_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}$(IA).*

Proof We set B to be the alphabet of VALC($\mathcal{M}$), and represent $L_{\mathcal{M}}$ as intersection of the three languages

$$L_{\mathcal{M},1} = \{w \mid w \in (\triangleright\mathrm{VALC}(\mathcal{M})\triangleleft)^*\}\,,$$
$$L_{\mathcal{M},2} = \{w^i \mid w \in \triangleright B^*\triangleleft,\ i \geq 2,\ i \quad \text{even}\}, \text{ and}$$
$$L_{\mathcal{M},3} = \{wu \mid w \in \triangleright B^*\triangleleft,\ u \in (\triangleright B^*\triangleleft)^*, |wu|_{\triangleright} = |w|!\}.$$

Since $\mathcal{L}_{\mathrm{rt}}$(IA) is closed under intersection, it remains to be shown that each of the languages is accepted by some real-time IA. Language $L_{\mathcal{M},1}$ is the marked iteration of $\triangleright$VALC($\mathcal{M}$) followed by a single $\triangleleft$, where $\triangleleft$ is the marking symbol. Since VALC($\mathcal{M}$) $\in \mathcal{L}_{\mathrm{rt}}$(IA) and due to its closure under concatenation with single symbols and under marked iteration [73], we obtain $L_{\mathcal{M},1} \in \mathcal{L}_{\mathrm{rt}}$(IA).

The copy language $\{vv \mid v \in B^*\}$ belongs to $\mathcal{L}_{\mathrm{rt}}$(IA) [19]. So, $L = \{vv \mid v \in \triangleright B^*\triangleleft\}$ is accepted by some real-time IA. Since $\mathcal{L}_{\mathrm{rt}}$(IA) is closed under marked iteration and intersection, the languages L^* and $\triangleright B^* \triangleleft L^* \triangleright B^*\triangleleft$ as well as their intersection are accepted by some real-time IA. Here the marking is hidden in a regular structure. The new word starts after the second symbol $\triangleleft$, respectively. Clearly, the intersection equals $L_{\mathcal{M},2}$.

The construction of a real-time IA accepting the language $L_{\mathcal{M},3}$ makes use of the IA-constructibility of the factorials (cf. Example 1 and the discussion before Theorem 26). Recall, that this means that there is an IA which indicates by states of the leftmost cell the time steps $n!$, for $n \geq 1$. For further results on IA-constructibility we refer to [13,27,63].

Since all constructions and closures are effective, the assertion follows. □

Lemma 57 *Let $\mathcal{M}$ be some Turing machine. Then $L_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}$(OCA) if and only if $L(\mathcal{M})$ is finite.*

Proof We observe that a finite language is regular and that the regular languages are accepted by real-time OCAs. Conversely, if $L(\mathcal{M})$ is infinite, then we apply Lemma 8 in order to show $L_{\mathcal{M}} \notin \mathcal{L}_{\mathrm{rt}}$(OCA). Assume the contrary, and let p be the constant of Lemma 8. Clearly, due to the infinity of $L(\mathcal{M})$ there is some $w \in \triangleright\mathrm{VALC}(\mathcal{M})\triangleleft$ such that $|w|! > p^{|w|}$. We conclude $w^{|w|!} \in L_{\mathcal{M}}$, and the conditions of Lemma 8 are met with $k = |w|!$. Therefore, there is some $1 \leq q \leq p^{|w|}$ such that $w^{|w|!+q} \in L_{\mathcal{M}}$. But $|w|! < |w|!+q < (|w|+1)!$ and, thus, $|w|!+q$ is not a factorial which implies the contradiction $w^{|w|!+q} \notin L_{\mathcal{M}}$. □

Now we may obtain the next (un)decidability result.

Theorem 58 *For any language family $\mathcal{L}$ that effectively contains $\mathcal{L}_{rt}$(IA), it is not semidecidable whether $L \in \mathcal{L}$ is a real-time OCA language.*

Proof If the problem in question were semidecidable, then the finiteness for Turing machines is also semidecidable. To this end, given some Turing machine $\mathcal{M}$ we construct a real-time IA for the language $L_{\mathcal{M}}$ according to Lemma 56. If $L_{\mathcal{M}}$ is accepted by some real-time OCA, then $L_{\mathcal{M}}$ is finite by Lemma 57. This implies the finiteness of VALC($\mathcal{M}$) and, thus, the finiteness of $L(\mathcal{M})$. □

Since the families $\mathcal{L}_{\mathrm{rt}}$(OCA) and $\mathcal{L}_{\mathrm{rt}}$(IA) are incomparable with respect to set inclusion, there is a natural interest to know whether the incomparability is also with respect to the decidability in question. Therefore, we turn to the converse question of Theorem 58, i. e., whether it is (semi)decidable that a real-time OCA language is accepted by some real-time IA. First we need some preliminaries.

A Turing machine $\mathcal{M}$ is converted into a Turing machine $\mathcal{M}'$ such that the input alphabet A' of $\mathcal{M}'$ contains at least two symbols and, furthermore, $\mathcal{M}'$ accepts any input of length n if and only if there is at least one input of length n accepted by $\mathcal{M}$. Clearly, this conversion is always effectively possible. Extending a frequently used witness language, we set

$$L'_{\mathcal{M}} = \{\&x_k\& \cdots \&x_1 ? y_1\& \cdots \&y_k\& \mid k \geq 1, x_i^R = y_i z_i \\ \text{and} \quad y_i z_i \in (A')^* \quad \text{and} \quad x_i \in \mathrm{VALC}(\mathcal{M}')^R\},$$

where $\&$ and $?$ are new symbols not appearing in $\mathrm{VALC}(\mathcal{M}')$.

Lemma 59 *Let $\mathcal{M}$ be some Turing machine. Then $L'_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}(IA)$ if and only if $L(\mathcal{M})$ is finite.*

Proof If $L(\mathcal{M})$ is finite, then so is $L(\mathcal{M}')$ and, thus, $\mathrm{VALC}(\mathcal{M}')^R$ is finite, say $\mathrm{VALC}(\mathcal{M}')^R = \{v_1, v_2, \ldots, v_r\}$. A real-time deterministic pushdown automaton accepting $L'_{\mathcal{M}}$ has r different stack symbols representing the elements in $\{v_1, v_2, \ldots, v_r\}$. It reads the input until the $?$ appears. For any occurring $v \in \mathrm{VALC}(\mathcal{M}')^R$ the corresponding stack symbol is pushed onto the stack. After reading the $?$, the pushdown automaton matches each y_i with the suffix of the $v \in \mathrm{VALC}(\mathcal{M}')^R$ which is identified by the symbol at the top of the stack. By Theorem 4, we obtain $L'_{\mathcal{M}} \in \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$.

Now let $L(\mathcal{M})$ be infinite. Then $L(\mathcal{M}')$, $\mathrm{VALC}(\mathcal{M}')^R$, and $L'_{\mathcal{M}}$ are infinite, as well. In order to show that in this case $L'_{\mathcal{M}}$ does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ we apply Lemma 6 as follows. Assume in contrast to the assertion that $L'_{\mathcal{M}}$ is accepted by some real-time IA with state set S. Every $v \in \mathrm{VALC}(\mathcal{M}')^R$ has a suffix of the form $\$\, u\, s_0$, where s_0 is the initial state and $u = \mathit{input}(v)$ is the reversal of the input. Let $|u| = k$. Moreover, due to the construction of $\mathcal{M}'$, for every input word with length k there is an element in $\mathrm{VALC}(\mathcal{M}')^R$. We denote the set of these elements by $V(k)$ and conclude $|V(k)| = |A'|^k$. For two different prefixes $w = \&\, x_k\, \& \cdots \&\, x_1\ ?$ and $w' = \&\, x'_k\, \& \cdots \&\, x'_1\ ?$ with $x_i, x'_i \in V(k)$, $1 \leq i \leq k$, there exists at least one $1 \leq j \leq k$ such that $x_j \neq x'_j$. Therefore, $w\, \&^{j-1}\, s_0\, \mathit{input}(x_j)^R\, \&^{k-j+1} \in L'_{\mathcal{M}}$ and $w'\, \&^{j-1}\, s_0\, \mathit{input}(x_j)^R\, \&^{k-j+1} \notin L'_{\mathcal{M}}$. Since the number of such prefixes is $|A'|^{k^2}$ and $|A'| \geq 2$, we obtain at least 2^{k^2} different $2k$-equivalence classes with respect to $L'_{\mathcal{M}}$. On the other hand, there is a constant $p \geq 1$ such that $E(L'_{\mathcal{M}}, 2k) \leq p^{2k}$. Since $L(\mathcal{M})$ is infinite, we may choose k large enough such that $2^{k^2} > p^{2k}$, which is a contradiction. □

Lemma 60 *Given some Turing machine $\mathcal{M}$, a real-time OCA accepting $L'_{\mathcal{M}}$ can effectively be constructed from $\mathcal{M}$, i. e., the language $L'_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}(OCA)$.*

Proof The language $L'_{\mathcal{M}}$ can be represented as the intersection of L_1 and L_2, where $L_1 = \{\&x_k\& \cdots \&x_1 ? y_1\& \cdots \&y_k\& \mid k \geq 1, x_i^R = y_i\, z_i \text{ and } x_i, y_i, z_i \in (A')^*\}$ and $L_2 = (\&\mathrm{VALC}(\mathcal{M}')^R)^* ?((A')^*\&)^*$. Since L_1 is a linear context-free language, it belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. The family $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ contains $\mathrm{VALC}(\mathcal{M}')$, is closed under reversal, marked iteration and right concatenation with regular sets [73]. Therefore, L_2 belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$, as well. From its closure under intersection we derive $L'_{\mathcal{M}} \in \mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. □

Similarly to Theorem 58 we obtain the next undecidability of a structural property.

Theorem 61 *For any language family $\mathcal{L}$ that effectively contains $\mathcal{L}_{rt}(OCA)$ it is not semidecidable whether $L \in \mathcal{L}$ is a real-time IA language.*

Proof If the problem in question were semidecidable, then also the finiteness for Turing machines. To this end, given some Turing machine $\mathcal{M}$, we construct a real-time OCA for the language $L'_{\mathcal{M}}$ according to Lemma 60. If $L'_{\mathcal{M}}$ is accepted by some real-time IA, then $L(\mathcal{M})$ is finite by Lemma 59. □

In general, a family $\mathcal{L}$ of languages possesses a *pumping lemma in the narrow sense* if for each $L \in \mathcal{L}$ there exists a constant $n \geq 1$ computable from L such that each $z \in L$ with $|z| > n$ admits a factorization $z = uvw$, where $|v| \geq 1$ and $u' v^i\, w' \in L$, for infinitely many $i \geq 0$. The prefix u' and the suffix w' depend on u, w and i.

Theorem 62 *Any language family whose word problem is semidecidable and that effectively contains $\mathcal{L}_{rt}(OCA)$ or $\mathcal{L}_{rt}(IA)$ does not possess a pumping lemma (in the narrow sense).*

Proof Let $\mathcal{M}$ be a real-time OCA or IA and assume there is a pumping lemma. Clearly, $L(\mathcal{M})$ is infinite if and only if it contains some w with $|w| > n$. So, we can semidecide infiniteness by first computing n and then verifying for all words longer than n whether they belong to $L(\mathcal{M})$. If at least for one word the answer is in the affirmative, then by pumping infinitely many words belong to $L(\mathcal{M})$. □

Example 63 The families $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ and $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ itself as well as, e. g., the families $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$, $\mathcal{L}_{\mathrm{lt}}(\mathrm{OCA})$, $\mathcal{L}_{\mathrm{lt}}(\mathrm{CA}) = \mathcal{L}_{\mathrm{lt}}(\mathrm{IA})$, and $\mathcal{L}(\mathrm{CA}) = \mathcal{L}(\mathrm{IA}) = \mathrm{DSPACE}(n)$ do not possess a pumping lemma. □

Theorem 64 *There is no minimization algorithm converting some CA, OCA or IA with arbitrary time complexity to an equivalent automaton of the same type with a minimal number of states.*

Proof For a given input alphabet A, we consider a minimal CA or OCA accepting the empty language. It has $|A|$ states and no accepting states. Assume there is a minimization algorithm. Then we can minimize an arbitrary CA and OCA and check whether the result has $|A|$ states and no accepting states. In this case the accepted language is empty. If the minimal automaton has $|A|$ states and at least one accepting state, there is an input such that the leftmost cell is initially accepting. So, the input is accepted and the accepted language is not empty. Hence emptiness is decidable, which is a contradiction to Theorem 54. Similar arguments apply for IAs. □

It remains to be mentioned that there is a nontrivial decidable property of (unbounded) cellular automata. It is known that injectivity of the global transition function is equivalent to the reversibility of the automaton. It is shown in [3] that global reversibility is decidable for one-dimensional CAs, whereas the problem is undecidable for higher dimensions [41].

Future Directions

The investigation of cellular language acceptors obeying a linear space bound reveals the hierarchy of language families in between the regular and the deterministic context-sensitive languages established in Sect. "Computational Capacities" (see Fig. 11). If the space bound is omitted, that is, if there is a potentially unlimited number of cells, then computation universality is achieved by direct simulation of Turing machines [78]. In particular, the universality can be achieved in spite of additional structural and computational limitations [2,60,64,66]. Similarly, some space bound supposed for cellular language acceptors does not yield to new language families. A Turing machine sweeping back and forth over the nonempty part of the tape can simulate the parallel device obeying the same space bound.

On the other hand, if the cellular language acceptor is simultaneously $s(n)$ space and $t(n)$ time bounded, a Turing machine simulation takes $s(n) \cdot t(n)$ time. A challenging question for further investigations is to identify languages and language classes for which homogeneously structured massive parallelism can significantly decrease the time complexity of sequential devices. Of particular interest are languages which allow a maximal saving. That is, for a sequential time complexity $t(n)$, the parallel time complexity is bounded by $t(n)/s(n)$, where $s(n)$ is the parallel space complexity. For example, in case of unary real-time languages, OCAs cannot do better than deterministic finite-state machine. Conversely, it is well known that any one-tape Turing machine takes at least $O(n^2)$ time to accept the language of palindromes $\{w \mid w = w^R, w \in \{a,b\}^*\}$. Since it is a linear context-free language, it is accepted by some real-time OCA, achieving the maximal saving in time.

From a more general point of view, central questions for future studies concern the power of additional limited resources at the disposal of time or space bounded computations. For example, nondeterminism, dimensions, the number of bits communicated to neighboring cells, or the restriction to reversible computations, all these can be seen as limited resources. We discuss some approaches in more detail.

Traditionally, nondeterministic devices have been viewed as having as many nondeterministic guesses as time steps. The studies of this concept of unlimited nondeterminism led, for example, to the famous open LBA-problem or the unsolved question whether or not P equals NP. In order to gain further understanding of the nature of nondeterminism, in [28,44] it has been viewed as an additional limited resource. In [11,46,52] cellular automata started to be considered from this point of view.

In classical computations the states of the neighboring cells are communicated in one time step. That is, the number of bits exchanged is determined by the number of states. A natural and interesting restriction is to limit the number of bits to some constant being independent of the number of states. Iterative arrays with restricted inter-cell communication have been investigated in [93,94], where algorithmic design techniques for sequence generation are shown. In particular, several important infinite, non-regular sequences such as exponential or polynomial, Fibonacci and prime sequences can be generated in real time. Connectivity recognition problems are dealt with in [92], whereas in [99] the computational capacity of one-way cellular automata with restricted inter-cell communication is considered. First results concerning formal language aspects of IAs with restricted inter-cell communication are shown in [53,54].

Finally, we turn to reversibility. Reversibility in the context of computing devices means that deterministic computations are also backward deterministic. Roughly speaking, in a reversible device no information is lost and every configuration occurring in any computation has at most one predecessor. Many different formal models have been studied in connection with reversibility. An early result on general reversible CAs is the possibility to make any CA, possibly irreversible, reversible by increasing the dimension. In detail, in [91] it is shown that any k-dimensional CA can be embedded into a $(k+1)$-dimensional reversible CA. This result has significantly been improved by showing how to make irreversible one-dimensional CAs reversible without increasing the dimension [65]. Furthermore, it is known that even reversible one-dimensional one-way CAs are computationally universal [64,66]. These results concern cellular automata with unbounded space. Moreover, in order to obtain a reversible device the neighborhood as well as the time complexity may be increased. In [23] it is shown that the neighborhood of a reversible CA is at most $n-1$ when the given reversible CA has n states. Additionally, this upper bound is shown to be tight. Cellular language acceptors with bounded space that are reversible on the core of computation, that is, from initial configuration to the configuration given by the time complexity, are introduced in [55,56]. At first glance, such a setting should simplify matters. However, it is quite the contrary, and such real-time reversibility is undecidable. There are many properties and relations still to be discovered in this setting.

Bibliography

Primary Literature

1. Achilles AC, Kutrib M, Worsch T (1996) On relations between arrays of processing elements of different dimensionality. In: Parallel processing by cellular automata and arrays (PARCELLA 1996). Mathematical Research 96. Akademie Verlag, Berlin, pp 13–20
2. Albert J, Čulik II K (1987) A simple universal cellular automaton and its one-way and totalistic version. Complex Systems 1: 1–16
3. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tesselation structures. J Comput System Sci 6:448–464
4. Baker BS, Book RV (1974) Reversal-bounded multipushdown machines. J Comput System Sci 8:315–332
5. Bleck B, Kröger H (1992) Cellular algorithms. In: Advances in parallel computing, vol 2. JAI Press, London, pp 115–143
6. Bucher W, Čulik II K (1984) On real time and linear time cellular automata. RAIRO Inform Théor 18:307–325
7. Buchholz T, Klein A, Kutrib M (1998) On time reduction and simulation in cellular spaces. Int J Comput Math 71:459–474
8. Buchholz T, Klein A, Kutrib M (1999) Iterative arrays with a wee bit alternation. In: Fundamentals of Computation Theory (FCT 1999). LNCS, vol 1684. Springer, Berlin, pp 173–184
9. Buchholz T, Klein A, Kutrib M (2000) Iterative arrays with small time bounds. In: Mathematical Foundations of Computer Science (MFCS 1998). LNCS, vol 1893. Springer, Berlin, pp 243–252
10. Buchholz T, Klein A, Kutrib M (2000) Real-time language recognition by alternating cellular automata. In: Theoretical Computer Science (TCS 2000). LNCS, vol 1872. Springer, Berlin, pp 213–225
11. Buchholz T, Klein A, Kutrib M (2002) On interacting automata with limited nondeterminism. Fund Inform 52:15–38
12. Buchholz T, Klein A, Kutrib M (2003) Iterative arrays with limited nondeterministic communication cell. In: Words, Languages and Combinatorics III (WLC 2000), World Scientific Publishing, Singapore, pp 73–87
13. Buchholz T, Kutrib M (1997) Some relations between massively parallel arrays. Parallel Comput 23(11):1643–1662
14. Buchholz T, Kutrib M (1998) On time computability of functions in one-way cellular automata. Acta Inf 35:329–352
15. Chang JH, Ibarra OH, Palis MA (1987) Parallel parsing on a one-way array of finite-state machines. IEEE Trans Comp C-36: 64–75
16. Chang JH, Ibarra OH, Vergis A (1988) On the power of one-way communication. J ACM 35:697–726
17. Choffrut C, Čulik II K (1984) On real-time cellular automata and trellis automata. Acta Inf 21:393–407
18. Cole SN (1966) Real-time computation by n-dimensional iterative arrays of finite-state machines. In: IEEE Symposium on Switching and Automata Theory (SWAT 1966). IEEE Press, New York, pp 53–77
19. Cole SN (1969) Real-time computation by n-dimensional iterative arrays of finite-state machines. IEEE Trans Comp C-18(4): 349–365
20. Čulik II K, Fris I (1985) Topological transformations as a tool in the design of systolic networks. Theoret Comp Sci 37:183–216
21. Čulik II K, Gruska J, Salomaa A (1984) Systolic trellis automata I. Int J Comp Math 15:195–212
22. Čulik II K, Gruska J, Salomaa A (1986) Systolic trellis automata: Stability, decidability and complexity. Inf Control 71:218–230
23. Czeizler E, Kari J (2005) A tight linear bound on the neighborhood of inverse cellular automata. In: Automata, Languages and Programming (ICALP 2005). LNCS, vol 3580. Springer, Berlin, pp 410–420
24. Delorme M, Mazoyer J (2004) Real-time recognition of languages on an two-dimensional archimedean thread. Theor Comp Sci 322:335–354
25. Dubacq JC, Terrier V (2002) Signals for cellular automata in dimension 2 or higher. In: Theoretical Informatics (LATIN 2002). LNCS, vol 2286. Springer, Berlin, pp 451–464
26. Dyer CR (1980) One-way bounded cellular automata. Inf Control 44:261–281
27. Fischer PC (1965) Generation of primes by a one-dimensional real-time iterative array. J ACM 12:388–394
28. Fischer PC, Kintala CMR (1979) Real-time computations with restricted nondeterminism. Math. Syst Theory 12:219–231
29. Ginsburg S, Greibach SA, Harrison MA (1967) One-way stack automata. J ACM 14:389–418
30. Hartmanis J (1967) Context-free languages and Turing machine computations. Proc Symp App Math 19:42–51
31. Höllerer WO, Vollmar R (1975) On forgetful cellular automata. J Comp Syst Sci 11:237–251
32. Ibarra OH, Jiang T (1987) On one-way cellular arrays. SIAM J Comp 16:1135–1154
33. Ibarra OH, Jiang T (1988) Relating the power of cellular arrays to their closure properties. Theor Comp Sci 57:225–238
34. Ibarra OH, Kim SM, Moran S (1985) Sequential machine characterizations of trellis and cellular automata and applications. SIAM J Comp 14:426–447
35. Ibarra OH, Kim SM (1984) Characterizations and computational complexity of systolic trellis automata. Theor Comp Sci 29:123–153
36. Ibarra OH, Palis MA (1985) Some results concerning linear iterative (systolic) arrays. J Paral Distrib Comp 2:182–218
37. Ibarra OH, Palis MA (1988) Two-dimensional iterative arrays: Characterizations and applications. Theor Comp Sci 57:47–86
38. Ibarra OH, Palis MA, Kim SM (1985) Fast parallel language

recognition by cellular automata. Theor Comp Sci 41(2–3):231–246

39. Imai K, Morita K (1996) Firing squad synchronization problem in reversible cellular automata. Theor Comp Sci 165(2):475–482
40. Iwamoto C, Hatsuyama T, Morita K, Imai K (2002) Constructible functions in cellular automata and their applications to hierarchy results. Theor Comp Sci 270:797–809
41. Kari J (1994) Reversibility and surjectivity problems of cellular automata. J Comp Syst Sci 48(1):149–182
42. Kasami T, Fuji M (1968) Some results on capabilities of one-dimensional iterative logical networks. Elect Commun Japan 51-C:167–176
43. Kim S, McCloskey R (1990) A characterization of constant-time cellular automata computation. Phys D 45:404–419
44. Kintala CMR (1977) Computations with a Restricted Number of Nondeterministic Steps. Ph D thesis, Pennsylvania State University, University Park
45. Klein A, Kutrib M (2003) Fast one-way cellular automata. Theor Comp Sci 1–3:233–250
46. Klein A, Kutrib M (2007) Cellular devices and unary languages. Fund Inf 78:343–368
47. Kosaraju SR (1975) Speed of recognition of context-free languages by array automata. SIAM J Comp 4:331–340
48. Krithivasan K, Mahajan M (1995) Nondeterministic, probabilistic and alternating computations on cellular array models. Theor Comp Sci 143:23–49
49. Kutrib M (1999) Pushdown cellular automata. Theor Comp Sci 215(1–2):239–261
50. Kutrib M (2001) Efficient universal pushdown cellular automata and their application to complexity. In: Machines, Computations, and Universality (MCU 2001). LNCS, vol 2055. Springer, Berlin, pp 252–263
51. Kutrib M, Löwe JT (2002) Massively parallel fault tolerant computations on syntactical patterns. Fut Gener Comp Syst 18:905–919
52. Kutrib M, Löwe JT (2003) Space- and time-bounded nondeterminism for cellular automata. Fund Inf 58(2003):273–293
53. Kutrib M, Malcher A (2006) Fast cellular automata with restricted inter-cell communication: Computational capacity. In: Theoretical Computer Science (IFIP TCS2006). IFIP 209. Springer, Berlin, pp 151–164
54. Kutrib M, Malcher A (2006) Fast iterative arrays with restricted inter-cell communication: Constructions and decidability. In: Mathematical Foundations of Computer Science (MFCS 2006). LNCS, vol 4162. Springer, Berlin, pp 634–645
55. Kutrib M, Malcher A (2008) Fast reversible language recognition using cellular automata. Inform Comput 206(9–10):1142–1151
56. Kutrib M, Malcher A (2007) Real-time reversible iterative arrays. In: Fundamentals of Computation Theory 2007 (FCT 2007). LNCS, vol 4693. Springer, Berlin, pp 376–387
57. Kutrib M, Worsch T (1994) Investigation of different input modes for cellular automata. In: Parallel Processing by Cellular Automata and Arrays (PARCELLA 1994). Mathematical Research 81, Akademie Verlag, Berlin, pp 141–150
58. Malcher A (2002) Descriptional complexity of cellular automata and decidability questions. J Autom Lang Comb 7:549–560
59. Malcher A (2004) On the descriptional complexity of iterative arrays. IEICE Trans Inf Syst E87-D(3):721–725
60. Martin B (1994) A universal cellular automaton in quasi-linear time and its S-m-n form. Theor Comp Sci 123(2):199–237
61. Matamala M (1997) Alternation on cellular automata. Theor Comp Sci 180:229–241
62. Mazoyer J (1987) A six-state minimal time solution to the firing squad synchronization problem. Theor Comp Sci 50:183–238
63. Mazoyer J, Terrier V (1999) Signals in one-dimensional cellular automata. Theor Comp Sci 217:53–80
64. Morita K (1992) Computation-universality of one-dimensional one-way reversible cellular automata. Inf Proc Lett 42:325–329
65. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comp Sci 148:157–163
66. Morita K, Harao M (1989) Computation universality of one dimensional reversible injective cellular automata. Trans IEICE E72 pp 758–762
67. Morita K, Ueno S (1994) Parallel generation and parsing of array languages using reversible cellular automata. Int J Pattern Recog Artif Int 8:543–561
68. Nakamura K (1999) Real-time language recognition by one-way and two-way cellular automata. In: Mathematical Foundations of Computer Science (MFCS 1999). LNCS, vol 1672. Springer, Berlin, pp 220–230
69. Rice HG (1953) Classes of recursively enumerable sets and their decision problems. Trans Amer Math Soc 89:25–59
70. Rosenfeld A (1979) Picture Languages. Academic Press, New York
71. Rosenstiehl P, Fiksel JR, Holliger A (1972) Intelligent graphs: Networks of finite automata capable of solving graph problems. In: Graph Theory and Computing. Academic Press, New York, pp 219–265
72. Salomaa A (1973) Formal Languages. Academic Press, Orlando
73. Seidel SR (1979) Language recognition and the synchronization of cellular automata. Technical Report 79–02, Department of Computer Science, University of Iowa, Iowa City
74. Seiferas JI (1977) Iterative arrays with direct central control. Acta Inf 8:177–192
75. Seiferas JI (1977) Linear-time computation by nondeterministic multidimensional iterative arrays. SIAM J Comp 6:487–504
76. Smith III AR (1970) Cellular automata and formal languages. In: IEEE Symposium on Switching and Automata Theory (SWAT 1970). IEEE Press, New York, pp 216–224
77. Smith III AR (1971) Cellular automata complexity trade-offs. Inf Control 18:466–482
78. Smith III AR (1971) Simple computation–universal cellular spaces. J ACM 18:339–353
79. Smith III AR (1971) Two-dimensional formal languages and pattern recognition by cellular automata. In: IEEE Symposium on Switching and Automata Theory (SWAT 1971). IEEE Press, New York, pp 144–152
80. Smith III AR (1972) Real-time language recognition by one-dimensional cellular automata. J Comp Syst Sci 6:233–253
81. Smith III AR (1976) Introduction to and survey of polyautomata theory. In: Automata, Languages, Development. North-Holland, Amsterdam, pp 405–422
82. Sommerhalder R, van Westrhenen SC (1983) Parallel language recognition in constant time by cellular automata. Acta Inf 19:397–407
83. Terrier V (1994) Language recognizable in real time by cellular automata. Complex Syst 8:325–336
84. Terrier V (1995) On real time one-way cellular array. Theor Comp Sci 141:331–335

85. Terrier V (1996) Language not recognizable in real time by one-way cellular automata. Theor Comp Sci 156:281–287
86. Terrier V (1999) Two-dimensional cellular automata recognizer. Theor Comp Sci 218:325–346
87. Terrier V (2003) Two-dimensional cellular automata and deterministic on-line tesselation automata. Theor Comp Sci 301:167–187
88. Terrier V (2004) Two-dimensional cellular automata and their neighborhoods. Theor Comp Sci 312:203–222
89. Terrier V (2006) Closure properties of cellular automata. Theor Comp Sci 352:97–107
90. Terrier V (2006) Low complexity classes of multidimensional cellular automata. Theor Comp Sci 369:142–156
91. Toffoli T (1977) Computation and construction universality of reversible cellular automata. J Comp Syst Sci 15:213–231
92. Umeo H (2001) Linear-time recognition of connectivity of binary images on 1-bit inter-cell communication cellular automaton. Parallel Comp 27:587–599
93. Umeo H, Kamikawa N (2002) A design of real-time non-regular sequence generation algorithms and their implementations on cellular automata with 1-bit inter-cell communications. Fund Inf 52:257–275
94. Umeo H, Kamikawa N (2003) Real-time generation of primes by a 1-bit-communication cellular automaton. Fund Inf 58:421–435
95. Umeo H, Morita K, Sugata K (1982) Deterministic one-way simulation of two-way real-time cellular automata and its related problems. Inf Proc Lett 14:158–161
96. Vollmar R (1981) On cellular automata with a finite number of state changes. Computing 3:181–191
97. von Neumann J (1966) Theory of Self-Reproducing Automata. In: Burks AW (ed) University of Illinois Press, Champaign
98. Waksman A (1966) An optimum solution to the firing squad synchronization problem. Inf Control 9:66–78
99. Worsch T (2000) Linear time language recognition on cellular automata with restricted communication. In: Theoretical Informatics (LATIN 2000). LNCS, vol 1776. Springer, Berlin, pp 417–426

Books and Reviews

Delorme M, Mazoyer J (eds) (1999) Cellular Automata – A Parallel Model. Kluwer, Dordrecht

Cellular Automata with Memory

Ramón Alonso-Sanz
ETSI Agrónomos (Estadística), Universidad Politécnica de Madrid, Madrid, Spain

Article Outline

Glossary

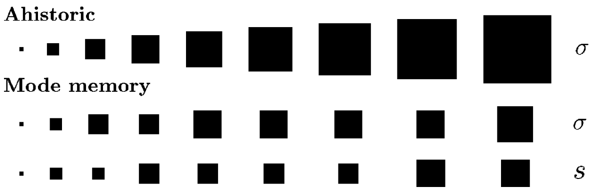
Cellular automata Cellular Automata (CA) are discrete, spatially explicit extended dynamic systems composed of adjacent cells characterized by an internal state whose value belongs to a finite set. The updating of these states is made simultaneously according to a common local transition rule involving only a neighborhood of each cell.

Memory Standard CA are ahistoric (memoryless): i. e., the new state of a cell depends on the neighborhood configuration only at the preceding time step. The standard framework of CA can be extended by the consideration of all past states (history) in the application of the CA rules by implementing memory capabilities in cells and links when topology is dynamic.

Definition

Cellular Automata (CA) are discrete, spatially explicit extended dynamic systems. A CA system is composed of adjacent cells characterized by an internal state whose value belongs to a finite set. The updating of these states is made simultaneously according to a common local transition rule involving only a neighborhood of each cell. Thus, if $\sigma_i^{(T)}$ is taken to denote the value of cell i at time step T, the site values evolve by iteration of the mapping: $\sigma_i^{(T+1)} = \phi\big(\{\sigma_j^{(T)}\} \in \mathcal{N}_i\big)$, where ϕ is an arbitrary function which specifies the cellular automaton *rule* operating on $\mathcal{N}_i$, i. e. the set of cells in the neighborhood of the cell i. Last but not least, standard CA are ahistoric (memoryless): i. e., the new state of a cell depends on the neighborhood configuration only at the preceding time step.

The standard framework of CA can be extended by the consideration of all past states (history) in the application of the CA rules by implementing memory capabilities in cells: $\sigma_i^{(T+1)} = \phi\big(\{s_j^{(T)}\} \in \mathcal{N}_i\big)$, with $s_j^{(T)} = s\big(\sigma_j^{(1)}, \ldots, \sigma_j^{(T-1)}, \sigma_j^{(T)}\big)$ being a state function of the series of states of the cell j up to time-step T. Thus, in CA with memory here: while the transition functions φ of the CA remain unaltered, historic memory of all past iterations is retained by featuring each cell by a summary of its past states.



Cellular Automata with Memory, Figure 1
The *speed of light* starting from a single active cell

The memory mechanism considered here is different from that of other CA with memory reported in the literature (e. g., p. 7 in [1], p. 43 in [35], p. 118 in [56]). Typically, these are higher-order-in-time rules that incorporate memory into the transition rule, determining the new configuration in terms of the configurations at previous time-steps. Thus, in *second order in time* rules: $\sigma_i^{(T+1)} = \Phi\big(\{\sigma_j^{(T)}\} \in \mathcal{N}_i, \{\sigma_j^{(T-1)}\} \in \mathcal{N}_i\big)$. CA with memory in cells are cited in [58], but just to state that "CA with memory in cells would result in a qualitatively different behavior". Some authors [55], define rules with *memory* as those with dependence in ϕ on the state of the cell to be updated. So, one-dimensional rules with no *memory*, take the form: $\sigma_i^{(T+1)} = \phi\big(\sigma_{i-1}^{(T)}, \sigma_{i+1}^{(T)}\big)$. Our use of the term memory is not any of these.

Introduction

As a simple example, in the two-dimensional, two-state automaton labeled ahistoric in Fig. 1, a cell becomes (or remains) alive if any cell in its nearest neighborhood is alive, but becomes (or remains) dead on the contrary case. The initial single perturbation in Fig. 1 spreads as fast as possible, i. e. at the *speed of light*.

The lower series of patterns in Fig. 1 shows the effect of featuring cells by their most frequent state, i. e. *mode* memory: $s_i^{(T)} = \text{mode}(\sigma_i^{(1)}, \dots, \sigma_i^{(T)})$ (with $s_i^{(T)} = \sigma_i^{(T)}$ in case of a tie) on the *speed of light*. Memory has a characteristic inertial effect.

Average Memory

Cells can be featured by a weighted mean value of all their previous states through powers of some parameter $\alpha \in [0,1]$ acting as a memory factor. Thus, at every time-step T, the weighted mean of the states up to T is computed for every cell i:

$$m_i^{(T)}(\sigma_i^{(1)}, \dots, \sigma_i^{(T)}) = \frac{\sigma_i^{(T)} + \sum_{t=1}^{T-1} \alpha^{T-t} \sigma_i^{(t)}}{1 + \sum_{t=1}^{T-1} \alpha^{T-t}} \equiv \frac{\omega_i^{(T)}}{\Omega(T)}$$

and then, the featuring states s are obtained by rounding the m ones to 1 for $m > 0.5$ and to 0 for $m < 0.5$. If m is exactly 0.5, then the last state is assigned ($s_i^{(T)} = \sigma_i^{(T)}$). This memory mechanism is *accumulative* in their demand of knowledge of past history: to calculate the memory charge $\omega_i^{(T)}$ it is not necessary to know the whole $\{\sigma_i^{(t)}\}$ series, while it can be sequentially calculated as: $\omega_i^{(T)} = \alpha\omega_i^{(T-1)} + \sigma_i^{(T)}$. It is, $\Omega(T) = (\alpha^T - 1)/(\alpha - 1)$.

The choice of the memory factor α simulates the long-term or remnant memory effect: the limit case $\alpha = 1$ corresponds to memory with equally weighted records (*full* memory, equivalent to *mode* if $k = 2$), whereas $\alpha \ll 1$ intensifies the contribution of the most recent states and diminishes the contribution of the past ones (short term working memory). The choice $\alpha = 0$ leads to the ahistoric model.

In the most unbalanced scenario up to T, i. e.: $\sigma_i^{(1)} = \dots = \sigma_i^{(T-1)} \neq \sigma_i^{(T)}$, it is:

$$m(0,0,\dots,0,1) = \frac{1}{2} \Rightarrow \frac{\alpha - 1}{\alpha^T - 1} = \frac{1}{2}$$
$$m(1,1,\dots,1,0) = \frac{1}{2} \Rightarrow \frac{\alpha^T - \alpha}{\alpha^T - 1} = \frac{1}{2}\,.$$

Thus, memory is only operative if α is greater than a critical α_T that verifies:

$$\alpha_T^T - 2\alpha_T + 1 = 0\,, \tag{1}$$

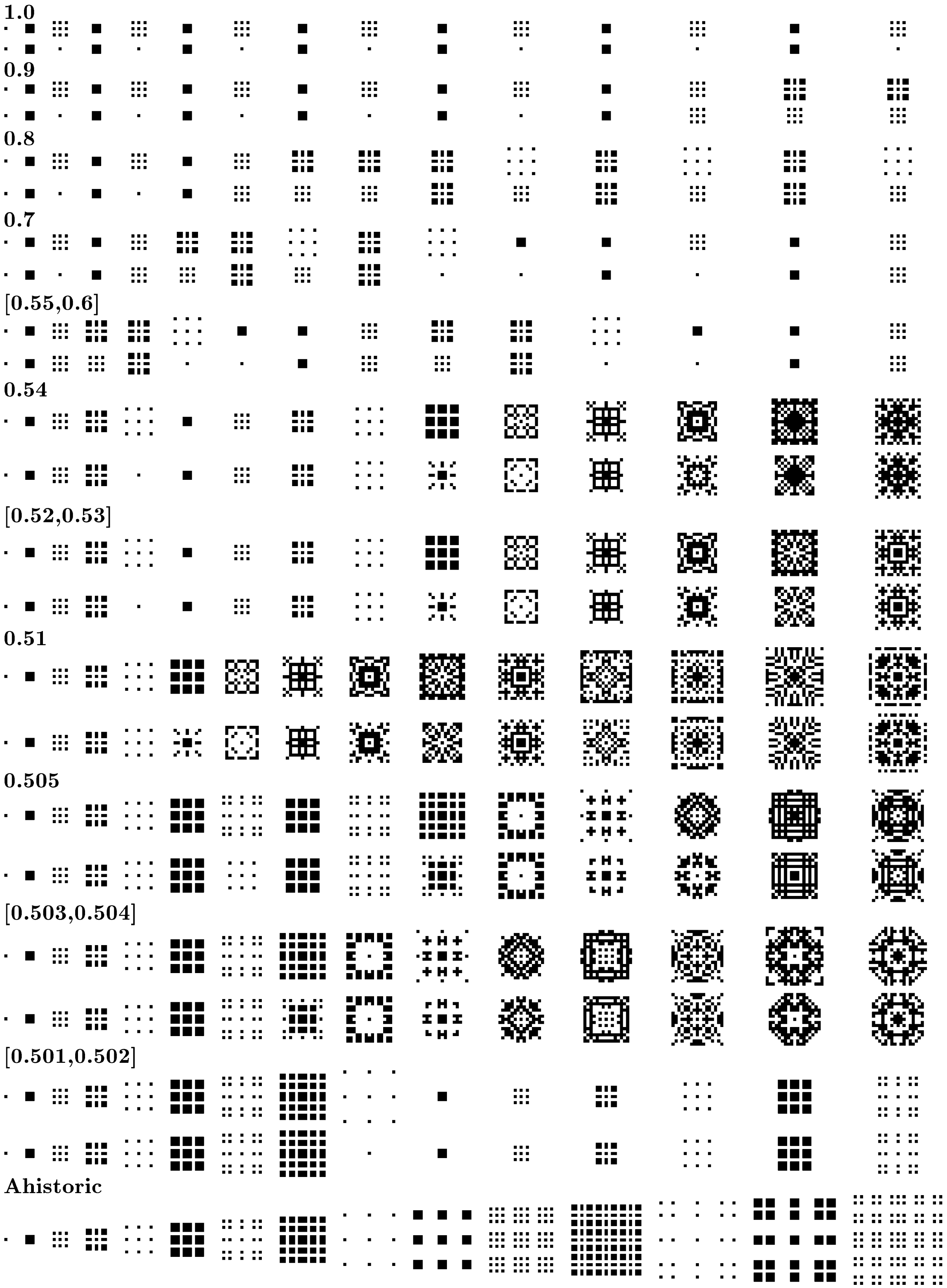
in which case cells will be featured at T with state values different to the last one. Initial operative values are: $\alpha_3 = 0.61805$, $\alpha_4 = 0.5437$. When $T \to \infty$, Eq. 1 becomes: $-2\alpha_\infty + 1 = 0$, thus, in the $k = 2$ scenario, α-memory is not effective if $\alpha \leq 0.5$.

A Worked Example: The Parity Rule

The so-called *parity* rule states: cell alive if the number of neighbors is odd, dead on the contrary case. Figure 2 shows the effect of memory on the parity rule starting from a single live cell in the Moore neighborhood. In accordance with the above given values of α_3 and α_4: (*i*) The pattern at $T = 4$ is the ahistoric one if $\alpha \leq 0.6$, altered when $\alpha \geq 0.7$, and (*ii*) the patterns at $T = 5$ for $\alpha = 0.54$ and $\alpha = 0.55$ differ.

Not low levels of memory tend to freeze the dynamics since the early time-steps, e. g. over 0.54 in Fig. 2. In the particular case of full memory small oscillators of short

► Cellular Automata with Memory, Figure 2
The 2D parity rule with memory up to $T = 15$


1.0
0.9
0.8
0.7
[0.55,0.6]
0.54
[0.52,0.53]
0.51
0.505
[0.503,0.504]
[0.501,0.502]
Ahistoric

range in time are frequently generated, such as the period-two oscillator that appears as soon as at $T = 2$ in Fig. 2. The group of evolution patterns shown in the [0.503,0.54] interval of α variation of Fig. 2, is rather unexpected to be generated by the parity rule, because they are too *sophisticated* for this simple rule. On the contrary, the evolution patterns with very small memory, $\alpha = 0.501$, resemble those of the ahistoric model in Fig. 2. But this similitude breaks later on, as Fig. 3 reveals: from $T = 19$, the parity rule with minimal memory evolves producing patterns notably different to the ahistoric ones. These patterns tend to be framed in squares of size not over $T \times T$, whereas in the ahistoric case, the patterns tend to be framed in $2T \times 2T$ square regions, so even minimal memory induces a very notable reduction in the affected cell area in the scenario of Fig. 2. The patterns of the featured cells tend not to be far to the actual ones, albeit examples of notable divergence can be traced in Fig. 2. In the particular case of the minimal memory scenario of Fig. 2, that of $\alpha = 0.501$, memory has no effect up to $T = 9$, when the pattern of featured live cells reduces to the initial one; afterward both evolutions are fairly similar up to $T = 18$, but at this time step both kinds of patterns notably differs, and since then the evolution patterns in Fig. 3 notably diverge from the ahistoric ones.

To give consideration to previous states (historic memory) in two-dimensional CA tends to confine the disruption generated by a single live cell. As a rule, full memory tends to generate oscillators, and less historic information retained, i. e. smaller α value, implies an approach to the ahistoric model in a rather smooth form. But the transition which decreases the memory factor from $\alpha = 1.0$ (full memory) to $\alpha = 0.5$ (ahistoric model), is not always regular, and some kind of *erratic* effect of memory can be traced.

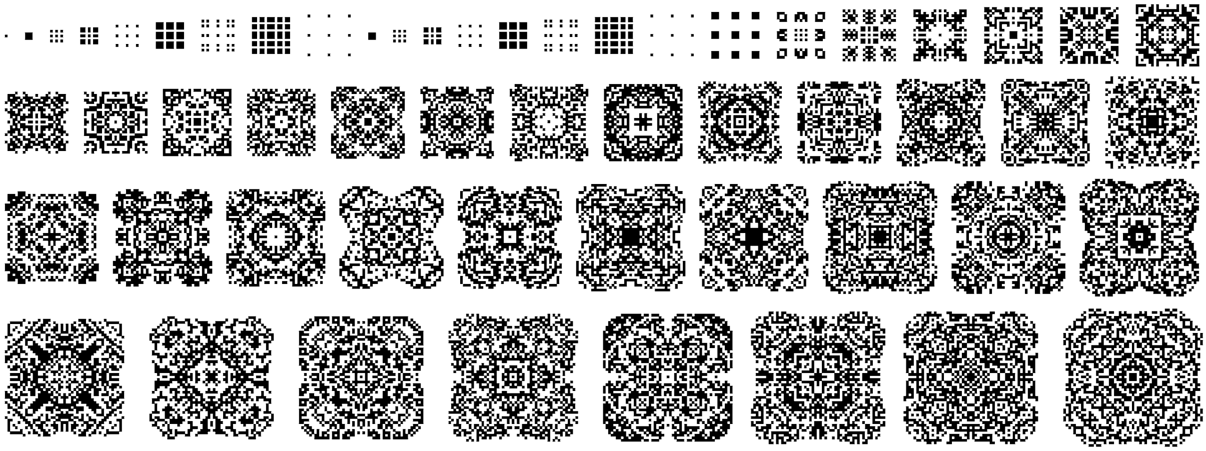
The inertial (or conservating) effect of memory dramatically changes the dynamics of the semitotalistic LIFE rule. Thus, (*i*) the *vividness* that some small clusters exhibit in LIFE, has not been detected in LIFE with memory. In particular, the *glider* in LIFE does not *glide* with memory, but stabilizes very close to its initial position as the *tub* , (*ii*) as the size of a configuration increases, often live clusters tend to persist with a higher number of live cells in LIFE with memory than in the ahistoric formulation, (*iii*) a single *mutant* appearing in a stable *agar* can lead to its destruction in the ahistoric model, whereas its effect tends to be restricted to its proximity with memory [26].

One-Dimensional CA *Elementary* rules are one-dimensional, two-state rules operating on nearest neighbors. Following Wolfram's notation, these rules are characterized by a sequence of binary values (β) associated with each of the eight possible triplets

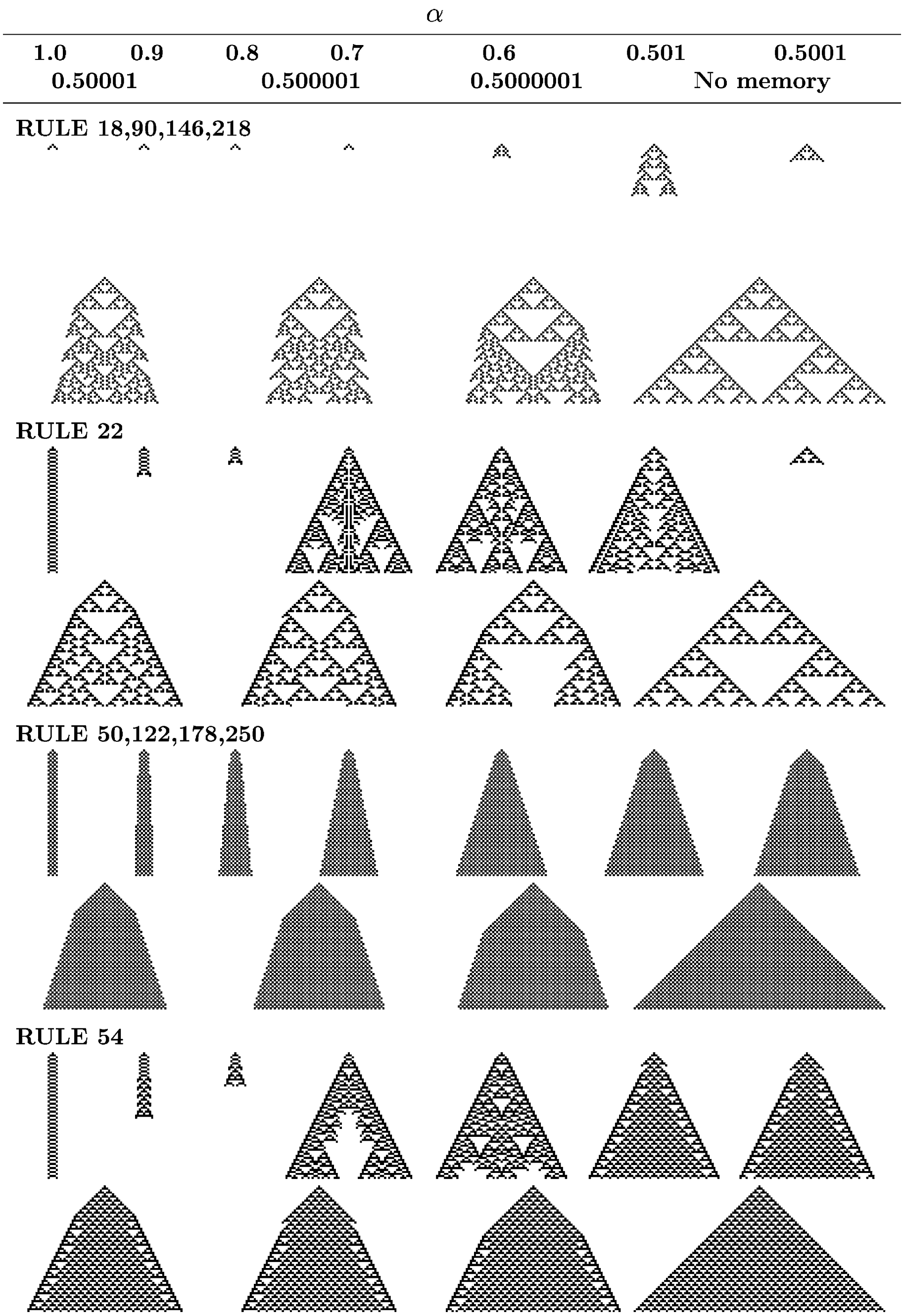
$$\left(\sigma_{i-1}^{(T)}, \sigma_{i}^{(T)}, \sigma_{i+1}^{(T)}\right) :$$

$$\begin{array}{cccccccc} 111 & 110 & 101 & 100 & 011 & 010 & 001 & 000 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 \end{array}$$

The rules are conveniently specified by their *rule number* $\mathcal{R} = \sum_{i=1}^{8} \beta_i 2^{8-i}$. *Legal* rules are *reflection symmetric* ($\beta_5 = \beta_2, \beta_7 = \beta_4$), and *quiescent* ($\beta_8 = 0$), restrictions that leave 32 possible *legal* rules.

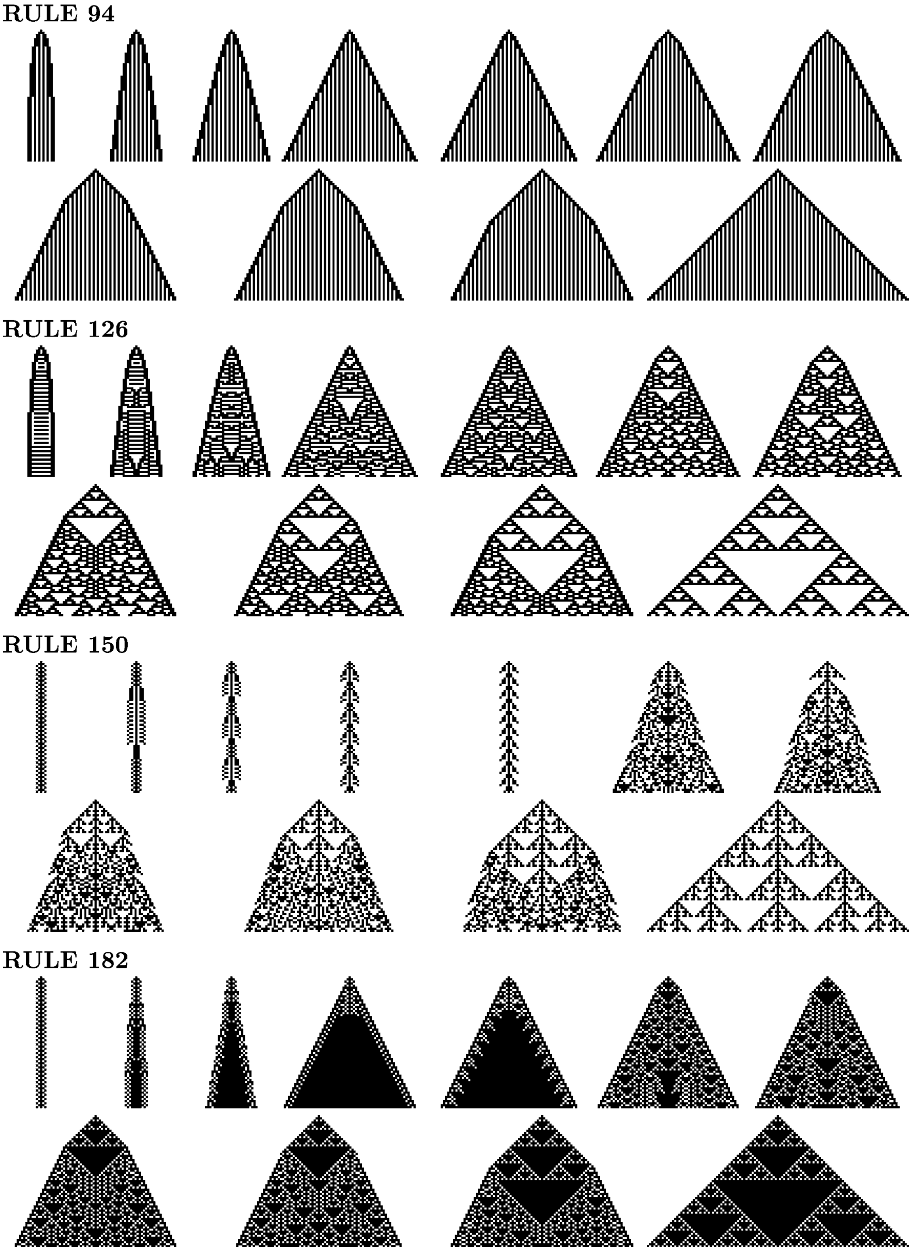


Cellular Automata with Memory, Figure 3
The 2D parity rule with $\alpha = 0.501$ memory starting from a single site live cell up to $T = 55$



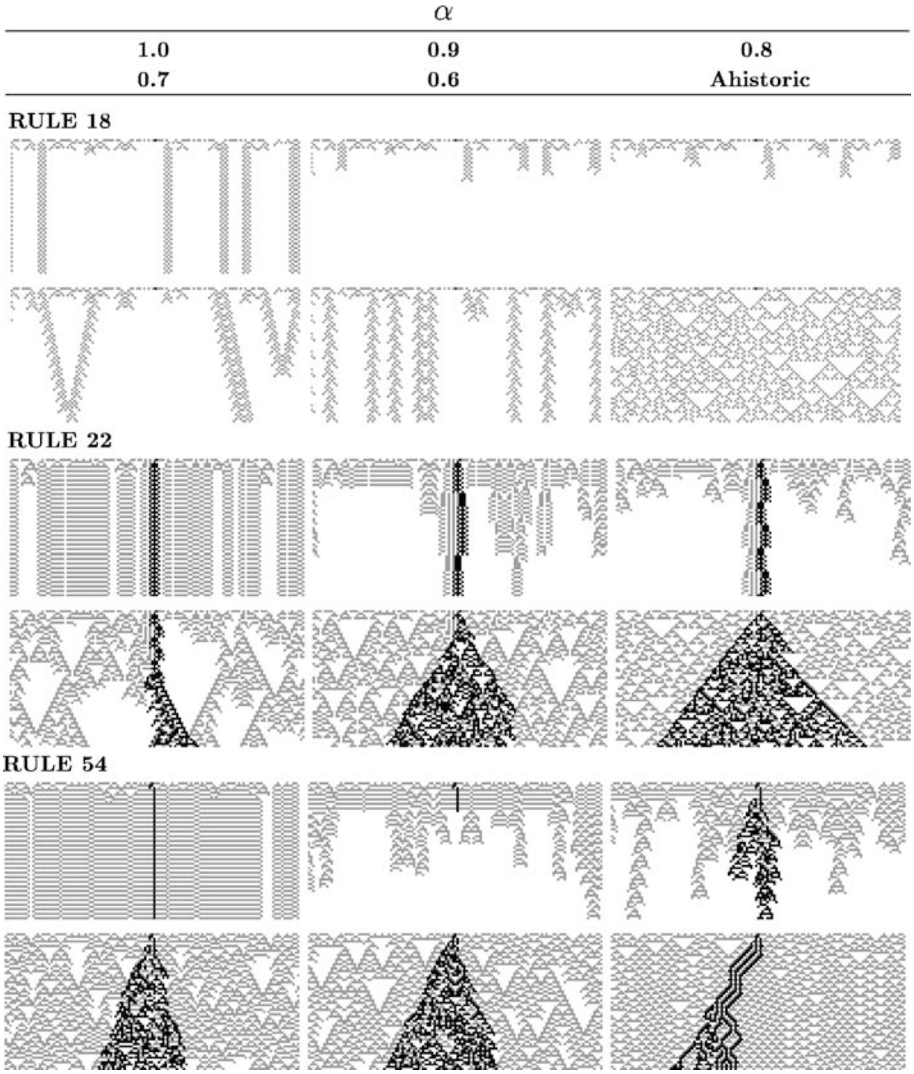
Cellular Automata with Memory, Figure 4
Elementary, legal rules with memory from a single site live cell

Figure 4 shows the spatio-temporal patterns of legal rules affected by memory when starting from a single live cell [17]. Patterns are shown up to $T = 63$, with the memory factor varying from 0.6 to 1.0 by 0.1 intervals, and adopting also values close to the limit of its effectivity: 0.5. As a rule, the transition from the $\alpha = 1.0$ (fully historic) to the ahistoric scenario is fairly gradual, so that the patterns become more expanded as less historic memory is retained (smaller α). Rules 50, 122, 178,250, 94, and 222,254 are paradigmatic of this smooth evolution. Rules 222 and



Cellular Automata with Memory, Figure 4
(***continued***)

254 are not included in Fig. 4 as they evolve as rule 94 but with the *inside* of patterns full of active cells. Rules 126 and 182 also present a gradual evolution, although their patterns with high levels of memory models hardly resemble the historic ones. Examples without a smooth effect of memory are also present in Fig. 4: (*i*) rule 150 is sharply restrained at $\alpha = 0.6$, (*ii*) the important rule 54 extinguish in [0.8,0.9], but not with full memory, (*iii*) the rules in the group {18,90,146,218} become extinct from $\alpha = 0.501$. Memory kills the evolution for these rules already at $T = 4$ for α values over α_3 (thus over 0.6 in Fig. 4): after $T = 3$ all the cells, even the two outer cells alive at $T = 3$, are featured as dead, and (*iv*) rule 22 becomes extinct for $\alpha = 0.501$, not in 0.507, 0.6, and 0.7, again extinguish at 0.8 and 0.9, and finally generate an oscillator with full memory. It has been argued that rules

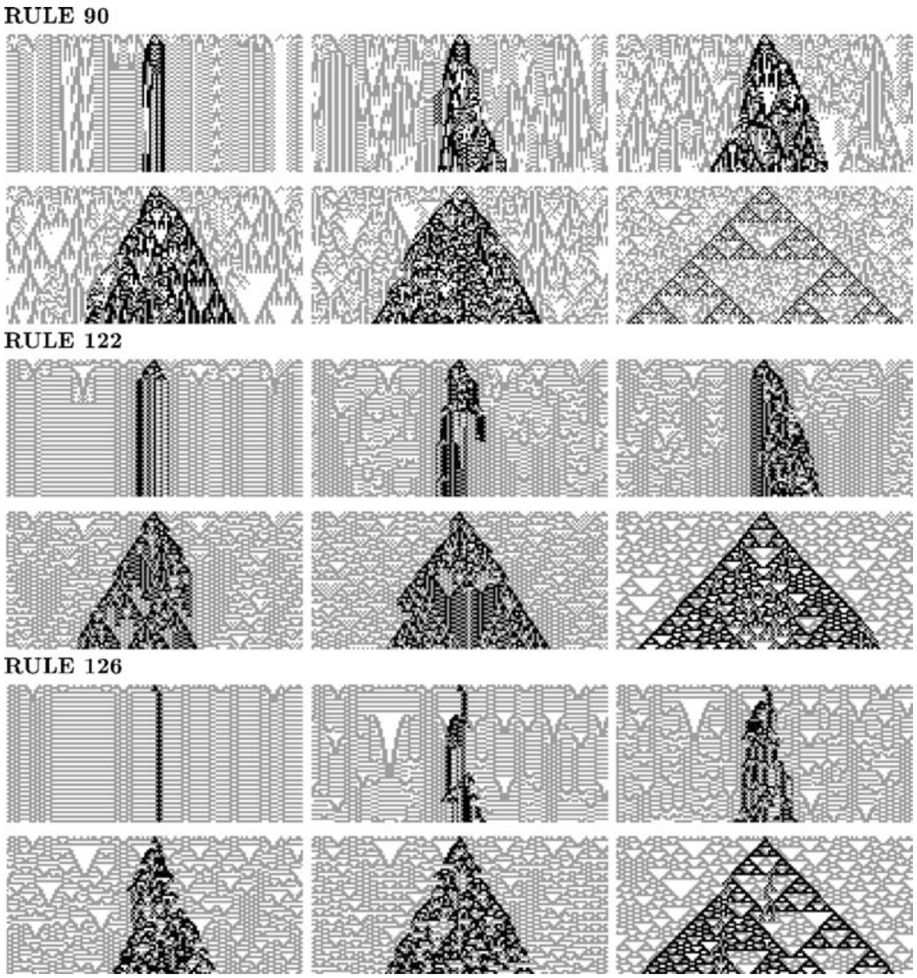


Cellular Automata with Memory, Figure 5
Elementary, legal rules with memory starting at random

18, 22, 122, 146 and 182 *simulate* Rule 90 in that their behavior coincides when restricted to certain spatial subsequences. Starting with a single site live cell, the coincidence fully applies in the historic model for rules 90, 18 and 146. Rule 22 shares with these rules the extinction for high α values, with the notable exception of no extinction in the fully historic model. Rules 122 and 182 diverge in their behavior: there is a gradual decrease in the width of evolving patterns as α is higher, but they do not reach extinction.

Figure 5 shows the effect of memory on legal rules when starting at random: the values of sites are initially uncorrelated and chosen at random to be 0 (*blank*) or 1 (*gray*) with probability 0.5. Differences in patterns resulting from reversing the center site value are shown as *black* pixels. Patterns are shown up to $T = 60$, in a line of size 129 with periodic boundary conditions imposed on the edges. Only the nine legal rules which generate non-periodic patterns in the ahistoric scenario are significantly affected by memory. The patterns with inverted triangles dominate the scene in the ahistoric patterns of Fig. 5, a common appearance that memory tends to eliminate.

History has a dramatic effect on Rule 18. Even at the low value of $\alpha = 0.6$, the appearance of its spatio-temporal pattern fully changes: a number of isolated periodic structures are generated, far from the distinctive inverted triangle world of the ahistoric pattern. For $\alpha = 0.7$, the live structures are fewer, advancing the extinction found in



Cellular Automata with Memory, Figure 5
(*continued*)

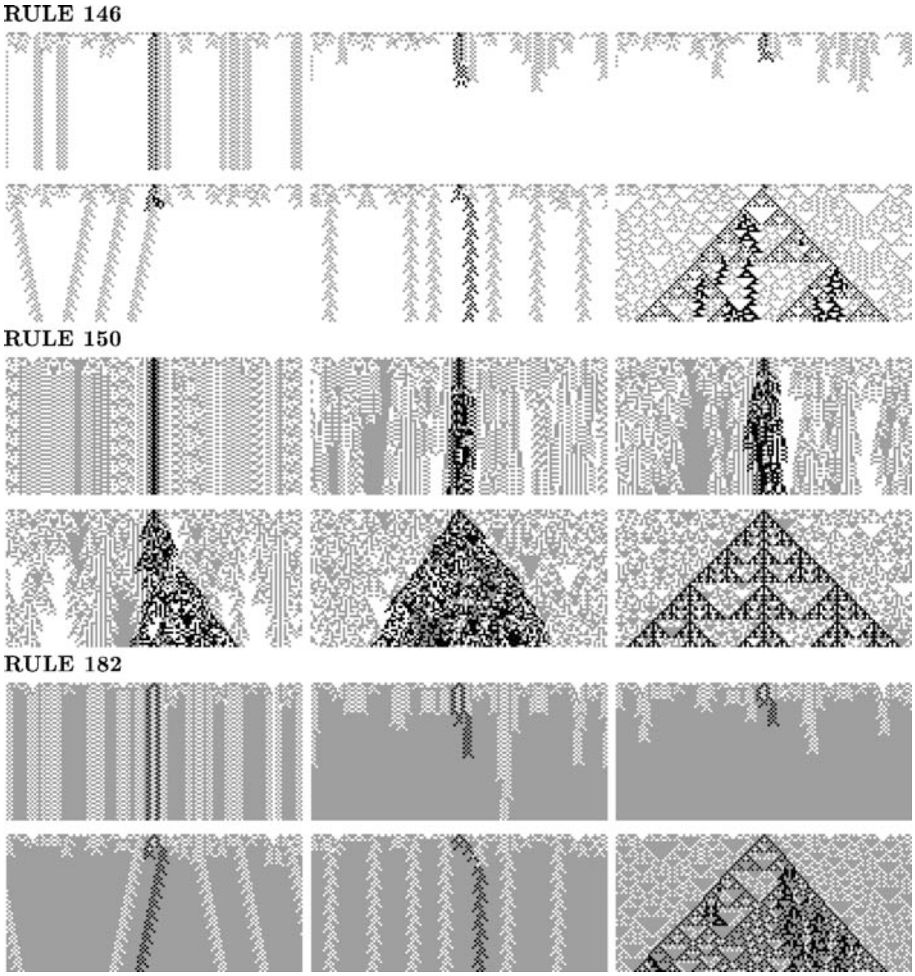
[0.8,0.9]. In the fully historic model, simple periodic patterns survive.

Rule 146 is affected by memory in much the same way as Rule 18 because their binary codes differ only in their β_1 value. The spatio-temporal of rule 182 and its equivalent Rule 146 are reminiscent, though those of Rule 182 look like a *negatives* photogram of those of Rule 146.

The effect of memory on rule 22 and the *complex* rule 54 is similar. Their spatio-temporal patterns in $\alpha = 0.6$ and $\alpha = 0.7$ keep the essential of the ahistoric, although the inverted triangles become enlarged and tend to be more sophisticated in their *basis*. A notable discontinuity is found for both rules ascending in the value of the memory factor: in $\alpha = 0.8$ and $\alpha = 0.9$ only a few simple structures survive. But unexpectedly, the patterns of the fully historic scenario differ markedly from the others, showing a high degree of synchronization.

The four remaining chaotic legal rules (90, 122, 126 and 150) show a much smoother evolution from the ahistoric to the historic scenario: no pattern evolves either to full extinction or to the preservation of only a few isolated persistent propagating structures (solitons). Rules 122 and 126, evolve in a similar form, showing a high degree of synchronization in the fully historic model.

As a rule, the effect of memory on the differences in patterns (DP) resulting from reversing the value of its initial center site is reminiscent of that on the spatio-temporal patterns, albeit this very much depends on the actual simulation run. In the case of rule 18 for example, damage is not present in the simulation of Fig. 5. The group of rules 90, 122, 126 and 150 shows a, let us say *canonical*, fairly gradual evolution from the ahistoric to the historic scenario, so that the DP appear more constrained as more historic memory is retained, with no extinction for any α



Cellular Automata with Memory, Figure 5
(*continued*)

value. Figure 6 shows the evolution of the fraction ρ_T of sites with value 1, starting at random ($\rho_0 = 0.5$). The simulation is implemented for the same rules as in Fig. 4, but with notably wider lattice: $N = 500$. A visual inspection of the plots in Fig. 6, ratifies the general features observed in the patterns in Fig. 5 regarding density. That also stands for damage spreading: as a rule, memory depletes the damaged region.

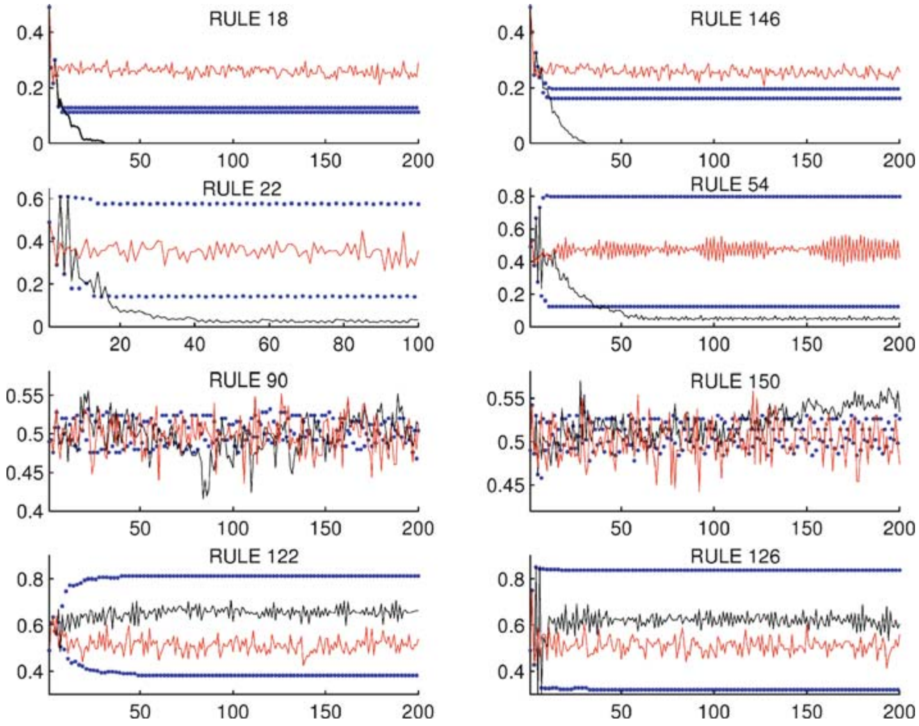
In one-dimensional $r = 2$ CA, the value of a given site depends on values of the nearest and next-nearest neighbors. *Totalistic* $r = 2$ rules with memory have the form: $\sigma_i^{(T+1)} = \phi\big(s_{i-2}^{(T)} + s_{i-1}^{(T)} + s_i^{(T)} + s_{i+1}^{(T)} + s_{i+2}^{(T)}\big)$. The effect of memory on these rules follows the way traced in the $r = 1$ context, albeit with a rich casuistic studied in [14].

Probabilistic CA So far the CA considered are deterministic. In order to study perturbations to deterministic CA as well as transitional changes from one deterministic CA to another, it is natural to generalize the deterministic CA framework to the probabilistic scenario. In the elementary scenario, the β are replaced by probabilities

$$p = P\big(\sigma_i^{(T+1)} = 1 \big/ \sigma_{i-1}^{(T)}, \sigma_i^{(T)}, \sigma_{i+1}^{(T)}\big) \ :$$

111	110	101	100	011	010	001	000	
β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	$\beta \in \{0,1\}$
p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	$p \in [0,1]$

As in the deterministic scenario, memory can be embedded in probabilistic CA (PCA) by featuring cells by a summary of past states s_i instead of by their last state σ_i: $p = P\big(\sigma_i^{(T+1)} = 1/s_{i-1}^{(T)}, s_i^{(T)}, s_{i+1}^{(T)}\big)$. Again, memory is embedded into the characterization of cells but not in the construction of the stochastic transition rules, as done in the canonical approach to PCA. We have



Cellular Automata with Memory, Figure 6

Evolution of the density starting at random in elementary legal rules. Color code: *blue* → full memory, *black* → $\alpha = 0.8$, *red* → ahistoric model

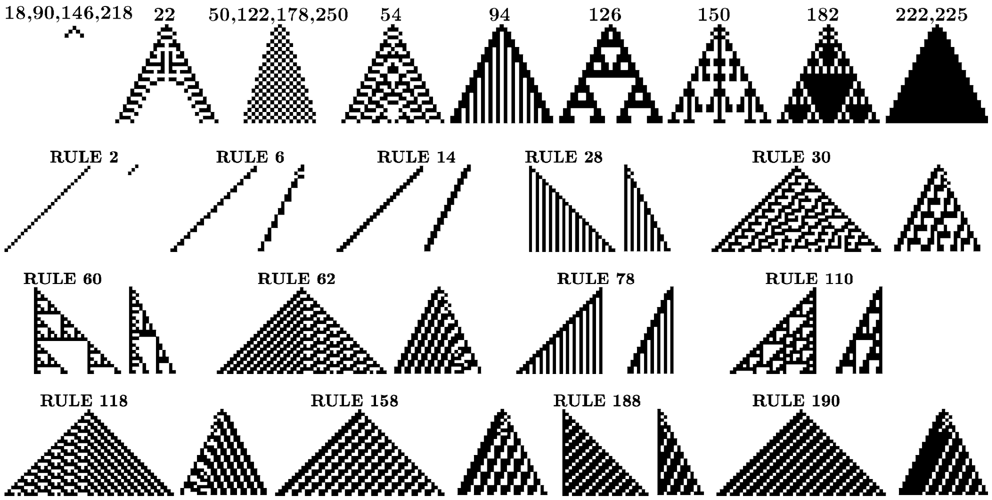
explored the effect of memory on three different subsets $(0, p_2, 0, p_4, p_2, 0, p_4, 0)$, $(0, 0, p_3, 1, 0, p_6, 1, 0)$, and $(p_1, p_2, p_1, p_2, p_2, 0, p_2, 0)$ in [9].

Other Memories

A number of average-like memory mechanisms can readily be proposed by using weights different to that implemented in the α-memory mechanism: $\delta(t) = \alpha^{T-t}$. Among the plausible choices of δ, we mention the weights $\delta(t) = t^c$ and $\delta(t) = c^t$, $c \in \mathbb{N}$, in which the larger the value of c, the more heavily is the recent past taken into account, and consequently the closer the scenario to the ahistoric model [19,21]. Both weights allow for integer-based arithmetics (*à la* CA) comparing $2\omega^{(T)}$ to $2\Omega(T)$ to get the featuring states s (a clear computational advantage over the α-based model), and are accumulative in respect to charge: $\omega_i^{(T)} = \omega_i^{(T-1)} + T^c \sigma_i^{(T)}$, $\omega_i^{(T)} = \omega_i^{(T-1)} + c^T \sigma_i^{(T)}$. Nevertheless, they share the same drawback: powers *explode*, at high values of t, even for $c = 2$.

Limited trailing memory would keep memory of only the last τ states. This is implemented in the context of average memory as: $\omega_i^{(T)} = \sum_{t=\top}^{T} \delta(t)\sigma_i^{(t)}$, with $\top = \max(1, T - \tau + 1)$. Limiting the trailing memory would approach the model to the ahistoric model ($\tau = 1$). In the geometrically discounted method, such an effect is more appreciable when the value of α is high, whereas at low α values (already close to the ahistoric model when memory is not limited) the effect of limiting the trailing memory is not so important. In the $k = 2$ context, if $\tau = 3$, provided that $\alpha > \alpha_3 = 0.61805$, the memory mechanism turns out to be that of selecting the mode of the last three states: $s_i^{(T)} = mode\left(\sigma_i^{(T-2)}, \sigma_i^{(T)}, \sigma_i^{(T-1)}\right)$, i. e. the elementary rule 232.

Figure 7 shows the effect of this kind of memory on legal rules. As is known, history has a dramatic effect on Rules 18, 90, 146 and 218 as their pattern dies out as early as at $T = 4$. The case of Rule 22 is particular: two *branches* are generated at $T = 17$ in the historic model; the patterns of the remaining rules in the historic model are much reminiscent of the ahistoric ones, but, let us say, *compressed*. Figure 7 shows also the effect of memory on some relevant quiescent asymmetric rules. Rule 2 *shifts* a single site live cell one space at every time-step in the ahistoric model; with the pattern dying at $T = 4$. This evolution is common



Cellular Automata with Memory, Figure 7

Legal (first row of patterns) and quiescent asymmetric elementary rules significantly affected by the *mode* of the three last states of memory

to all rules that just shift a single site cell without increasing the number of living cells at $T = 2$, this is the case of the important rules 184 and 226. The patterns generated by rules 6 and 14 are *rectified* (in the sense of having the lines in the spatio-temporal pattern slower slope) by memory in such a way that the total number of live cells in the historic and ahistoric spatio-temporal patterns is the same. Again, the historic patterns of the remaining rules in Fig. 7 seem, as a rule, like the ahistoric ones compressed [22].

Elementary rules (ER, noted f) can in turn act as memory rules:

$$s_i^{(T)} = f\left(\sigma_i^{(T-2)}, \sigma_i^{(T)}, \sigma_i^{(T-1)}\right)$$

Figure 8 shows the effect of ER memories up to $R = 125$ on rule 150 starting from a single site live cell up to $T = 13$. The effect of ER memories with $R > 125$ on rule 150 as well as on rule 90 is shown in [23]. In the latter case, complementary memory rules (rules whose rule number adds 255) have the same effect on rule 90 (regardless of the role played by the three last states in ϕ and the initial configuration). In the ahistoric scenario, Rules 90 and 150 are *linear* (or *additive*): i. e., any initial pattern can be decomposed into the superposition of patterns from a single site seed. Each of these configurations can be evolved independently and the results superposed (module two) to obtain the final complete pattern. The additivity of rules 90 and 150 remains in the historic model with linear memory rules.

Figure 9 shows the effect of elementary rules on the 2D parity rule with von Neumann neighborhood from a singe site live cell. This figure shows patterns from $T = 4$, being the three first patterns: ▪ ✚ ⁘. The consideration of CA rules as memory induces a fairly unexplored *explosion* of new patterns.

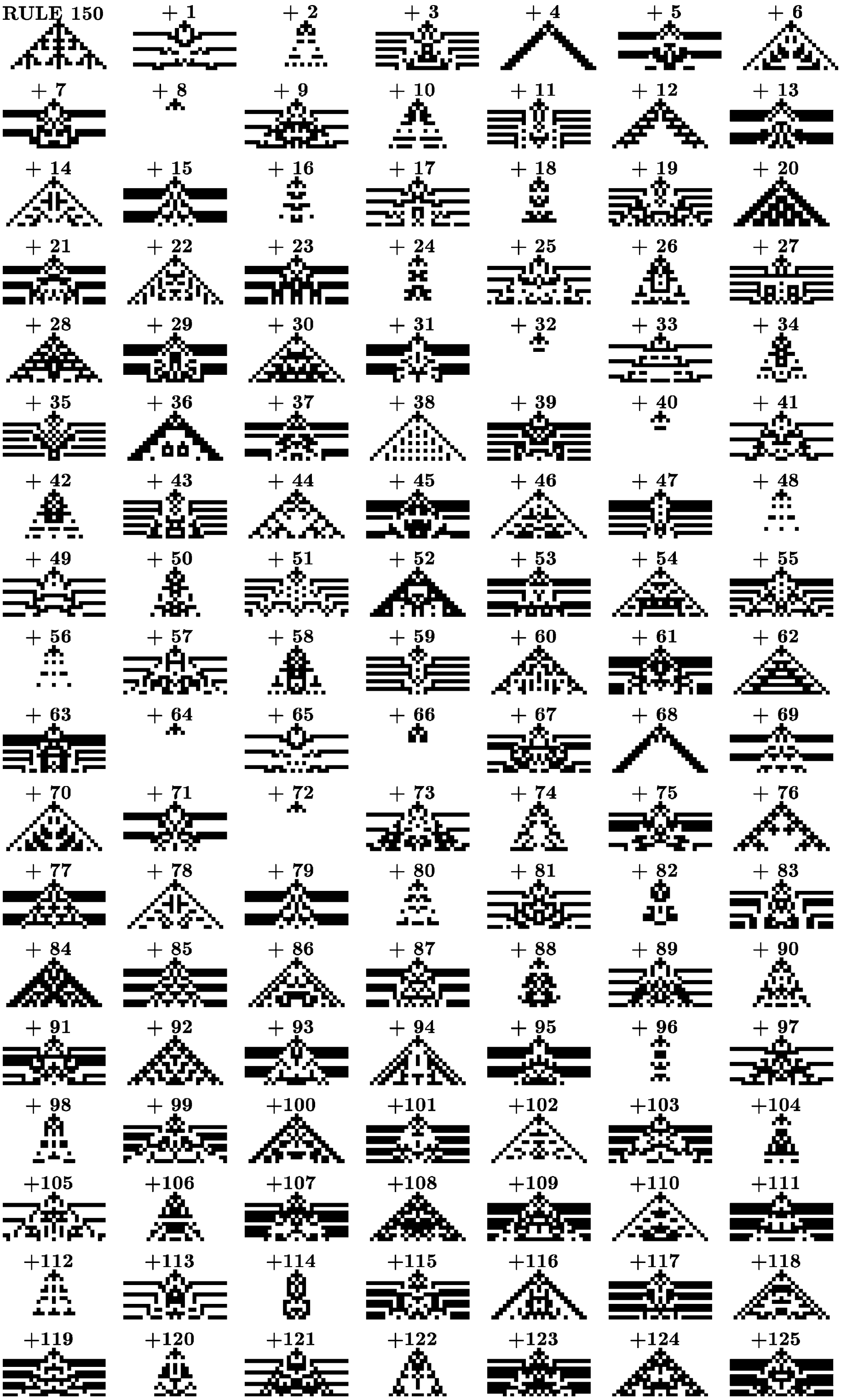
CA with Three Statesk

This section deals with CA with three possible values at each site ($k = 3$), noted $\{0, 1, 2\}$, so the rounding mechanism is implemented by comparing the unrounded weighted mean m to the hallmarks 0.5 and 1.5, assigning the last state in case on an equality to any of these values. Thus,

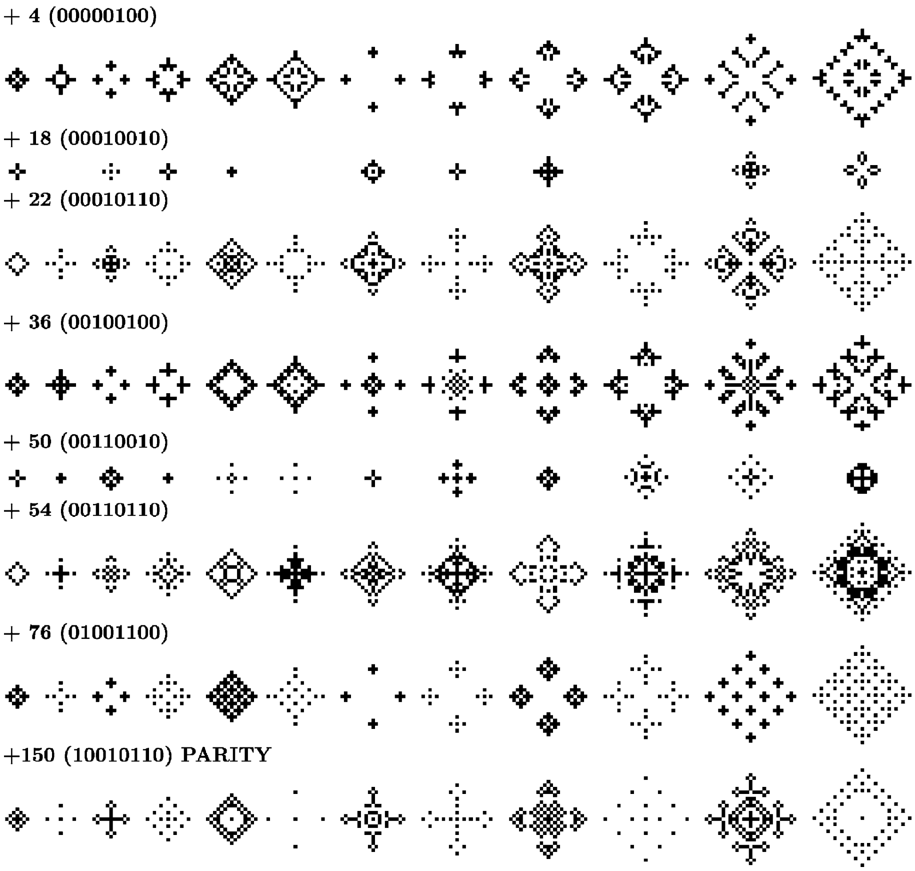
$$s^T = 0 \text{ if } m^T < 0.5, s^T = 1 \text{ if } 0.5 < m^T < 1.5, s^T = 2$$
$$\text{if } m^T > 1.5, \text{ and } s^T = \sigma^T \text{ if } m^T = 0.5 \text{ or } m^T = 1.5\,.$$

In the most unbalanced cell dynamics, historic memory takes effect after time step T only if $\alpha > \alpha_T$, with $3\alpha_T^T - 4\alpha_T + 1 = 0$, which in the temporal limit becomes $-4\alpha_* + 1 = 0 \Leftrightarrow \alpha_* = 0.25$. In general, in CA with k states (termed from 0 to $k-1$), the characteristic equation at T is $(2k-3)\alpha_T^T - (2k-1)\alpha_T + 1 = 0$, which becomes $-2(k-1)\alpha_* + 1 = 0$ in the temporal limit. It is then concluded that memory does not affect the scenario if $\alpha \le \alpha_*(k) = 1/(2(k-1))$.

We study first *totalistic* rules: $\sigma_i^{(T+1)} = \phi\left(\sigma_{i-1}^{(T)} + \sigma_i^{(T)} + \sigma_{i+1}^{(T)}\right)$, characterized by a sequence of ternary values (β_s) associated with each of the seven possible values of the sum (s) of the neighbors: $(\beta_6, \beta_5, \beta_4, \beta_3, \beta_2,$



Cellular Automata with Memory, Figure 8
The Rule 150 with elementary rules up to $R = 125$ as memory



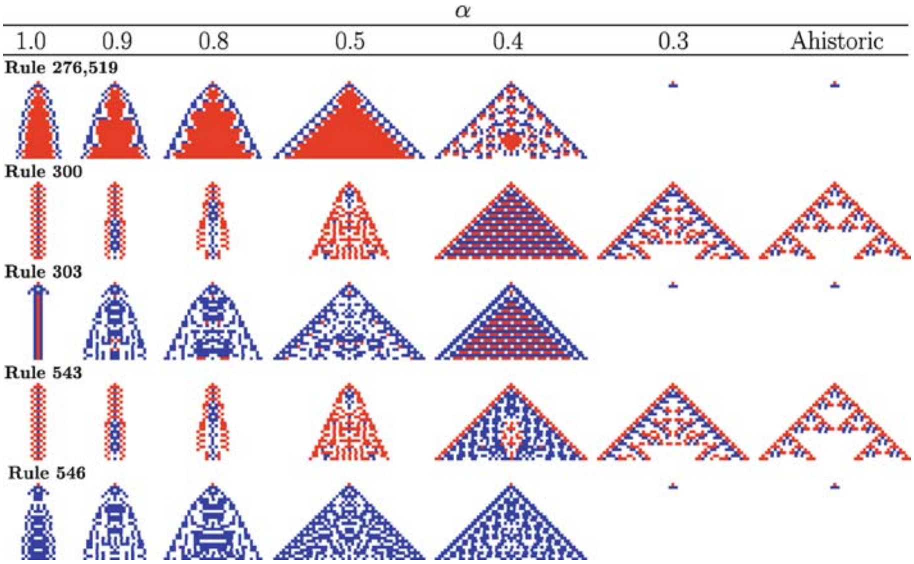
Cellular Automata with Memory, Figure 9
The parity rule with elementary rules as memory. Evolution from $T = 4 — 15$ in the Neumann neighborhood starting from a singe site live cell

β_1, β_0), with associated rule number $\mathcal{R} = \sum_{s=0}^{6} \beta_s 3^s \in [0, 2186]$.

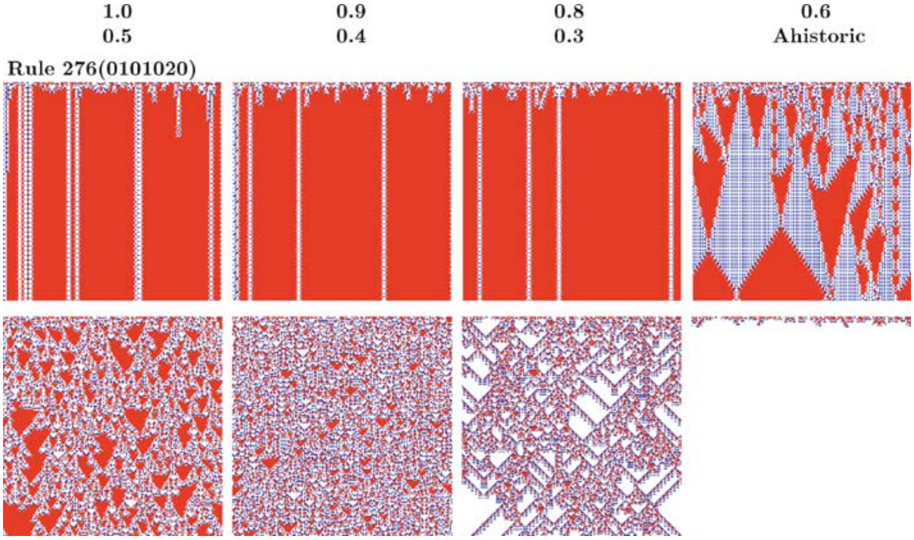
Figure 10 shows the effect of memory on quiescent ($\beta_0 = 0$) *parity* rules, i. e. rules with β_1, β_3 and β_5 non null, and $\beta_2 = \beta_4 = \beta_6 = 0$. Patterns are shown up to $T = 26$. The pattern for $\alpha = 0.3$ is shown to test its proximity to the ahistoric one (recall that if $\alpha \leq 0.25$ memory takes no effect). Starting with a single site seed it can be concluded, regarding proper three-state rules such as those in Fig. 10, that: *(i)* as an overall rule the patterns become more expanded as less historic memory is retained (smaller α). This characteristic *inhibition of growth* effect of memory is traced on rules 300 and 543 in Fig. 10, *(ii)* the transition from the fully historic to the ahistoric scenario tends to be gradual in regard to the amplitude of the spatio-temporal patterns, although their composition can differ notably, even at close α values, *(iii)* in contrast to the two-state scenario, memory fires the pattern of some three-state rules that die out in the ahistoric model, and no rule with memory dies out. Thus, the effect of memory on rules 276, 519, 303 and 546 is somewhat unexpected: they die out at $\alpha \leq 0.3$ but at $\alpha = 0.4$ the pattern expands, the expansion being inhibited (in Fig. 10) only at $\alpha \geq 0.8$. This activation under memory of rules that die at $T = 3$ in the ahistoric model is unfeasible in the $k = 2$ scenario.

The features in the evolving patterns starting from a single seed in Fig. 10 are qualitatively reflected starting at random as shown with rule 276 in Fig. 11, which is also *activated* (even at $\alpha = 0.3$) when starting at random. The effect of average memory (α and integer-based models, unlimited and limited trailing memory, even $\tau = 2$) and that of the mode of the last three states has been studied in [21].

When working with more than three states, it is an inherent consequence of averaging the tendency to *bias* the featuring state to the *mean* value: 1. That explains the *red-shift* in the previous figures. This led us to focus on a much more fair memory mechanism: the *mode*, in what follows.



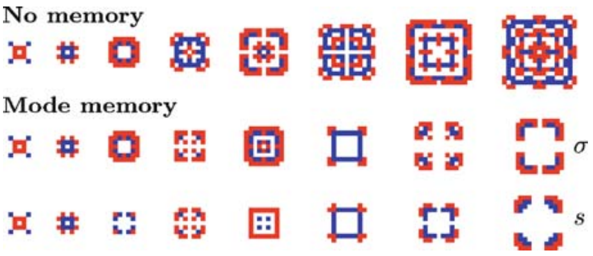
Cellular Automata with Memory, Figure 10
Parity $k = 3$ rules starting from a single $\sigma = 1$ seed. The red cells are at state 1, the blue ones at state



Cellular Automata with Memory, Figure 11
The $k = 3, \mathcal{R} = 276$ rule starting at random

Mode memory allows for manipulation of pure symbols, avoiding any computing/arithmetics.

In *excitable* CA, the three states are featured: resting *0*, excited *1* and refractory *2*. State transitions from excited to refractory and from refractory to resting are unconditional, they take place independently on a cell's neighborhood state: $\sigma_i^{(T)} = 1 \rightarrow \sigma_i^{(T+1)} = 2$, $\sigma_i^{(T)} = 2 \rightarrow \sigma_i^{(T+1)} = 0$. In [15] the excitation rule adopts a Pavlovian phenomenon of defensive inhibition: when strength of stimulus applied exceeds a certain limit the system 'shuts down', this can be naively interpreted as an inbuilt protection of energy loss and exhaustion. To simulate the phenomenon of defensive inhibition we adopt interval excitation rules [2], and a resting cell becomes excited only if one or two of its neighbors are excited: $\sigma_i^{(T)} = 0 \rightarrow \sigma_i^{(T)} = 1$ if $\sum_{j \in \mathcal{N}_i} \left(\sigma_j^{(T)} = 1\right) \in \{1, 2\}$ [3].



Cellular Automata with Memory, Figure 12
Effect of mode memory on the defensive inhibition CA rule

Figure 12 shows the effect of mode of the last three time steps memory on the defensive-inhibition CA rule with the Moore neighborhood, starting from a simple configuration. At $T = 3$ the outer excited cells in the actual pattern are not featured as excited but as resting cells (twice resting versus one excited), and the series of evolving patterns with memory diverges from the ahistoric evolution at $T = 4$, becoming less expanded. Again, memory tends to restrain the evolution.

The effect of memory on the *beehive* rule, a totalistic two-dimensional CA rule with three states implemented in the hexagonal tessellation [57] has been explored in [13].

Reversible CA

The second-order in time implementation based on the subtraction modulo of the number of states (noted $\ominus$): $\sigma_i^{(T+1)} = \phi(\sigma_j^{(T)} \in \mathcal{N}_i) \ominus \sigma_i^{(T-1)}$, readily reverses as: $\sigma_i^{(T-1)} = \phi(\sigma_j^{(T)} \in \mathcal{N}_i) \ominus \sigma_i^{(T+1)}$. To preserve the reversible feature, memory has to be endowed only in the pivotal component of the rule transition, so: $\sigma_i^{(T-1)} = \phi(s_j^{(T)} \in \mathcal{N}_i) \ominus \sigma_i^{(T+1)}$.

For reversing from T it is necessary to know not only $\sigma_i^{(T)}$ and $\sigma_i^{(T+1)}$ but also $\omega_i^{(T)}$ to be compared to $\Omega(T)$, to obtain:

$$s_i^{(T)} = \begin{cases} 0 & \text{if} \quad 2\omega_i^{(T)} < \Omega(T) \\ \sigma_i^{(T+1)} & \text{if} \quad 2\omega_i^{(T)} = \Omega(T) \\ 1 & \text{if} \quad 2\omega_i^{(T)} > \Omega(T) \, . \end{cases}$$

Then to progress in the reversing, to obtain $s_i^{(T-1)} = \text{round}\big(\omega_i^{(T-1)}/\Omega(T-1)\big)$, it is necessary to calculate $\omega_i^{(T-1)} = \big(\omega_i^{(T)} - \sigma_i^{(T)}\big)/\alpha$. But in order to avoid dividing by the memory factor (recall that operations with real numbers are not exact in computer arithmetic), it is preferable to work with $\gamma_i^{(T-1)} = \omega_i^{(T)} - \sigma_i^{(T)}$, and to compare these values to $\Gamma(T-1) = \sum_{t=1}^{T-1} \alpha^{T-t}$. This leads to:

$$s_i^{(T-1)} = \begin{cases} 0 & \text{if} \quad 2\gamma_i^{(T-1)} < \Gamma(T-1) \\ \sigma_i^{(T)} & \text{if} \quad 2\gamma_i^{(T-1)} = \Gamma(T-1) \\ 1 & \text{if} \quad 2\gamma_i^{(T-1)} > \Gamma(T-1) \, . \end{cases}$$

In general: $\gamma_i^{(T-\tau)} = \gamma_i^{(T-\tau+1)} - \alpha^{\tau-1}\sigma_i^{(T-\tau+1)}$, $\Gamma(T-\tau) = \Gamma(T-\tau+1) - \alpha^{\tau-1}$.

Figure 13 shows the effect of memory on the reversible parity rule starting from a single site live cell, so the scenario of Figs. 2 and 3, with the reversible qualification. As expected, the simulations corresponding to $\alpha = 0.6$ or below shows the ahistoric pattern at $T = 4$, whereas memory leads to a pattern different from $\alpha = 0.7$, and the pattern at $T = 5$ for $\alpha = 0.54$ and $\alpha = 0.55$ differ. Again, in the reversible formulation with memory, *(i)* the configuration of the patterns is notably altered, *(ii)* the speed of diffusion of the area affected are notably reduced, even by minimal memory ($\alpha = 0.501$), *(iii)* high levels of memory tend to freeze the dynamics since the early time-steps.

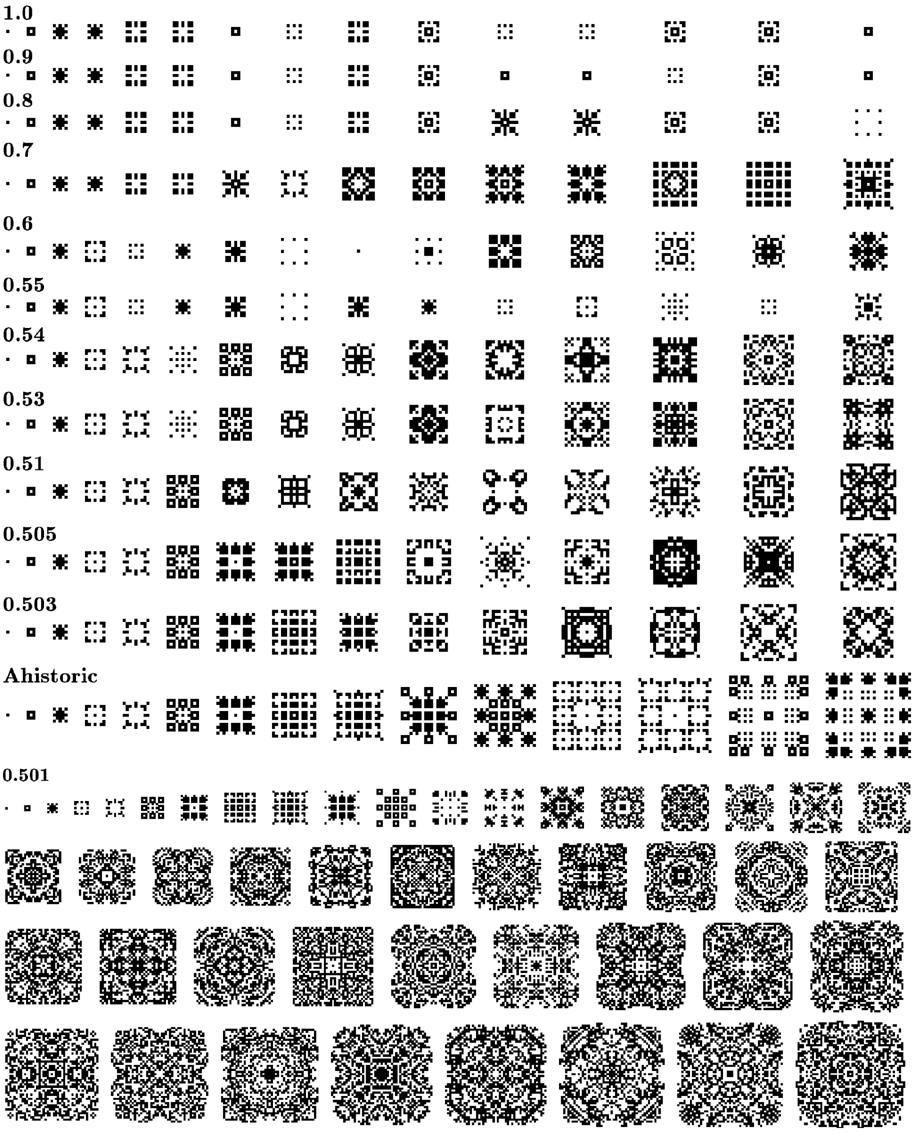
We have studied the effect of memory in the reversible formulation of CA in many scenarios, e. g., totalistic, $k = r = 2$ rules [7], or rules with three states [21].

Reversible systems are of interest since they preserve information and energy and allow unambiguous backtracking. They are studied in computer science in order to design computers which would consume less energy [51]. Reversibility is also an important issue in fundamental physics [31,41,52,53]. Geraldt 't Hooft, in a speculative paper [34], suggests that a suitably defined deterministic, local reversible CA might provide a viable formalism for constructing field theories on a Planck scale. Svozil [50] also asks for changes in the underlying assumptions of current field theories in order to make their discretization appear more CA-like. Applications of reversible CA with memory in cryptography are being scrutinized [30,42].

Heterogeneous CA

CA on networks have arbitrary connections, but, as proper CA, the transition rule is identical for all cells. This generalization of the CA paradigm addresses the intermediate class between CA and Boolean networks (BN, considered in the following section) in which, rules may be different at each site.

In networks two topological ends exist, random and regular networks, both display totally opposite geometric properties. Random networks have lower clustering coefficients and shorter average path length between nodes commonly known as *small world* property. On the other

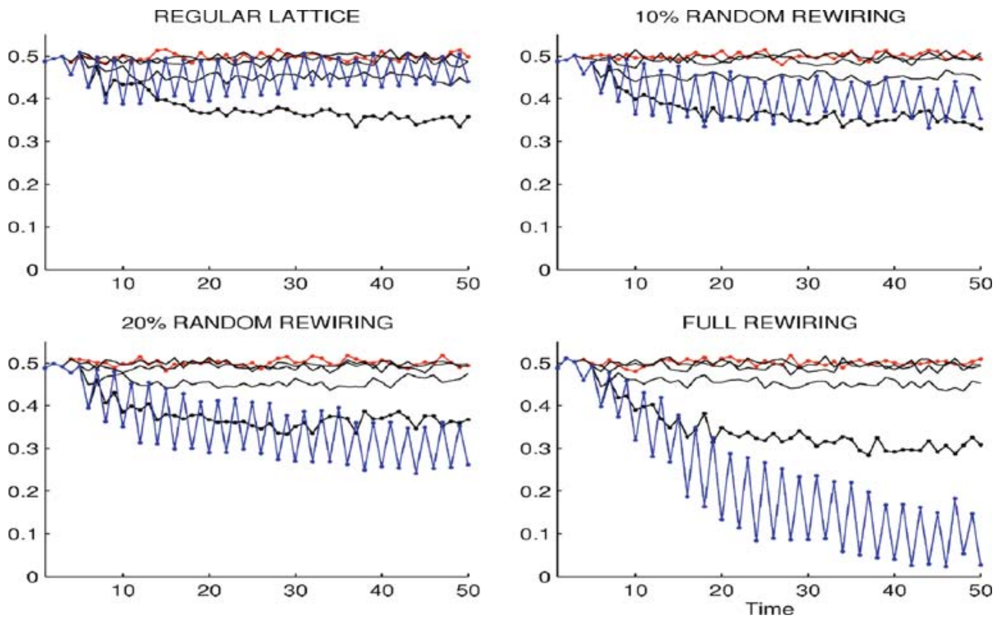


Cellular Automata with Memory, Figure 13
The reversible parity rule with memory

hand, regular graphs, have a large average path length between nodes and high clustering coefficients.

In an attempt to build a network with characteristics observed in real networks, a large clustering coefficient and a *small world* property, Watts and Strogatz (WS, [54]) proposed a model built by randomly rewiring a regular lattice. Thus, the WS model interpolates between regular and random networks, taking a single new parameter, the random rewiring degree, i. e.: the probability that any node redirects a connection, randomly, to any other. The WS model displays the high clustering coefficient common to regular lattices as well as the *small world* property (the *small world* property has been related to faster flow in the information transmission). The long-range links introduced by the randomization procedure dramatically reduce the diameter of the network, even when very few links are rewired.

Figure 14 shows the effect of memory and topology on the parity rule with four inputs in a lattice of size 65 × 65 with periodic boundary conditions, starting at random. As expected, memory depletes the Hamming distance between two consecutive patterns in relation to the *ahistoric*



Cellular Automata with Memory, Figure 14

The parity rule with four inputs: effect of memory and random rewiring. Distance between two consecutive patterns in the ahistoric model (*red*) and memory models of α levels: 0.6,0.7.0.8, 0.9 (*dotted*) and 1.0 (*blue*)

model, particularly when the degree of rewiring is high. With full memory, quasi-oscillators tend to appear. As a rule, the higher the curve the lower the memory factor α, but in the particular case of a regular lattice (and lattice with 10% of rewiring), the evolution of the distance in the full memory model turns out rather atypical, as it is maintained over some memory models with lower α parameters.

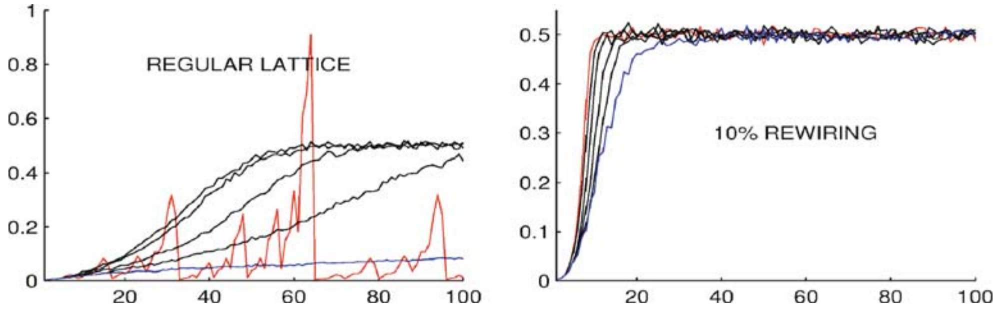
Figure 15 shows the evolution of the damage spread when reversing the initial state of the 3×3 central cells in the initial scenario of Fig. 14. The fraction of cells with the state reversed is plotted in the regular and 10% of rewiring scenarios. The plots corresponding to higher rates of rewiring are very similar to that of the 10% case in Fig. 15. Damage spreads fast very soon as rewiring is present, even in a short extent.

Boolean Networks

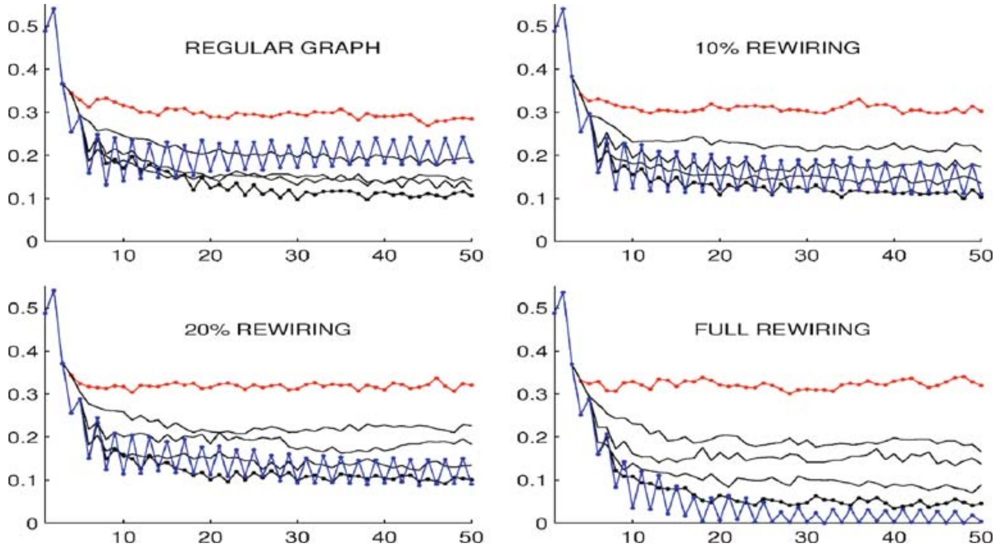
In Boolean Networks (BN,[38]), instead of what happens in canonical CA, cells may have arbitrary connections and rules may be different at each site. Working with totalistic rules: $\sigma_i^{(T+1)} = \phi_i(\sum_{j\in\mathcal{N}_i} s_j^{(T)})$.

The main features on the effect of memory in Fig. 14 are preserved in Fig. 16: (*i*) the ordering of the *historic* networks tends to be stronger with a high memory factor, (*ii*) with full memory, quasi-oscillators appear (it seems that full memory tends to induce *oscillation*), (*iii*) in the particular case of the regular graph (and a lesser extent in the networks with low rewiring), the evolution of the full memory model turns out rather atypical, as it is maintained over some of those memory models with lower α parameters. The relative Hamming distance between the *ahistoric* patterns and those of *historic* rewiring tends to be fairly constant around 0.3, after a very short initial transition period.

Figure 17 shows the evolution of the damage when reversing the initial state of the 3×3 central cells. As a rule in every frame, corresponding to increasing rates of random rewiring, the higher the curve the lower the memory factor α. The *damage vanishing* effect induced by memory does result apparently in the regular scenario of Fig. 17, but only full memory controls the damage spreading when the rewiring degree is not high, the dynamics with the remaining α levels tend to the damage propagation that characterizes the *ahistoric* model. Thus, with up to 10% of connections rewired, full memory notably controls the spreading, but this control capacity tends to disappear with a higher percentage of rewiring connections. In fact, with rewiring of 50% or higher, neither full memory seems to be very effective in altering the final rate of damage, which tends to reach a plateau around 30% regardless of



Cellular Automata with Memory, Figure 15
Damage up to $T = 100$ in the parity CA of Fig. 14



Cellular Automata with Memory, Figure 16
Relative Hamming distance between two consecutive patterns. Boolean network with totalistic, $K = 4$ rules in the scenario of Fig. 14

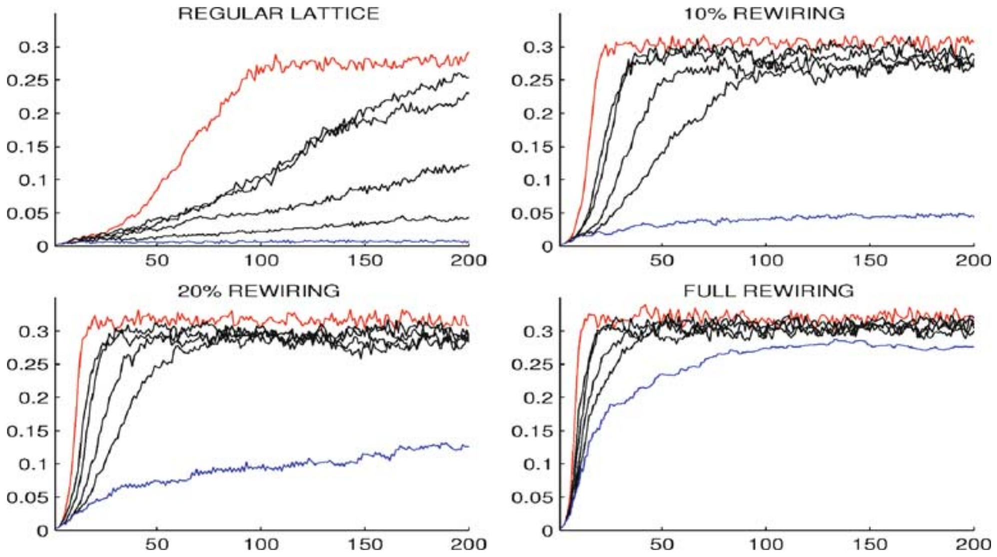
scenario. A level notably coincident with the percolation threshold in site percolation in the simple cubic lattice, and the critical point for the nearest neighbor Kaufmann model on the square lattice [49]: 0.31.

Structurally Dynamic CA

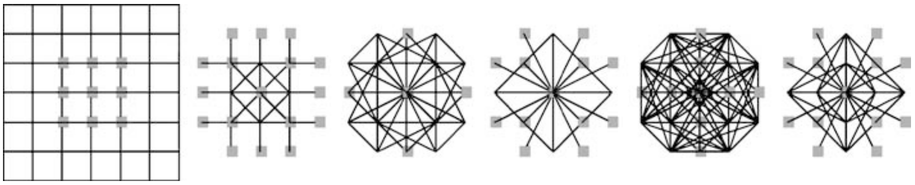
Structurally dynamic cellular automata (SDCA) were suggested by Ilachinski and Halpern [36]. The essential new feature of this model was that the connections between the cells are allowed to change according to rules similar in nature to the state transition rules associated with the conventional CA. This means that given certain conditions, specified by the *link transition rules*, links between rules may be created and destroyed; the neighborhood of each cell is dynamic, so, state and link configurations of an SDCA are both dynamic and continually interacting.

If cells are numbered 1 to N, their connectivity is specified by an $N \times N$ connectivity matrix in which $\lambda_i j = 1$ if cells i and j are connected; 0 otherwise. So, now: $\mathcal{N}_i^{(T)} = \{j/\lambda_{ij}^{(T)} = 1\}$ and $\sigma_i^{(T+1)} = \phi(\sigma_j^{(T)} \in \mathcal{N}_i^{(T)})$. The geodesic *distance* between two cells i and j, δ_{ij}, is defined as the number of links in the shortest path between i and j. Thus, i and j are *direct* neighbors if $\delta_{ij} = 1$, and are *next-nearest* neighbors if $\delta_i j = 2$, so $\mathcal{N}\mathcal{N}_i^{(T)} = \{j/\delta_{ij}^{(T)} = 2\}$. There are two types of link transition functions in an SDCA: *couplers* and *decouplers*, the former add new links, the latter remove links. The coupler and decoupler set determines the link transition rule: $\lambda_{ij}^{(T+1)} = \psi(l_{ij}^{(T)}, \sigma_i^{(T)}, \sigma_j^{(T)})$.

Instead of introducing the formalism of the SDCA, we deal here with just one example in which the decoupler rule removes all links connected to cells in which both values are zero ($\lambda_{ij}^{(T)} = 1 \to \lambda_{ij}^{(T+1)} = 0$ iff $\sigma_i^{(T)} + \sigma_j^{(T)} = 0$)



Cellular Automata with Memory, Figure 17
Evolution of the damage when reversing the initial state of the 3 × 3 central cells in the scenario of Fig. 16



Cellular Automata with Memory, Figure 18
The SDCA described in text up to $T = 6$

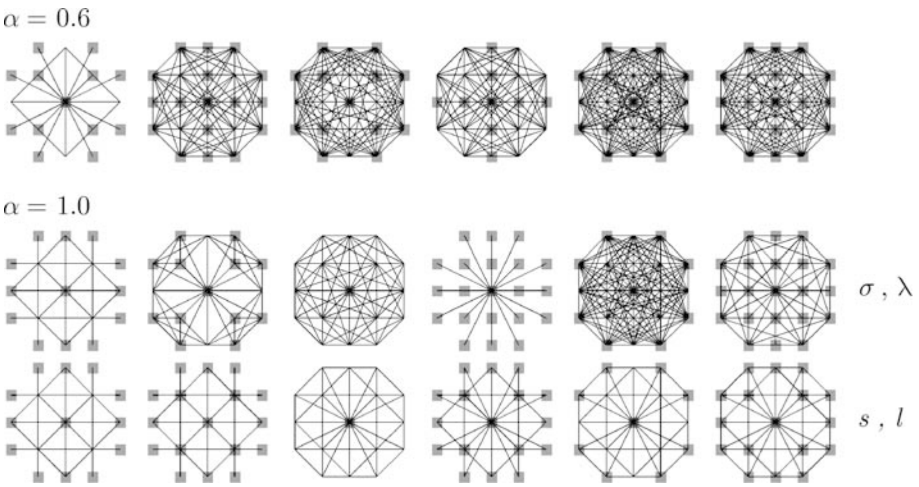
and the coupler rule adds links between all next-nearest neighbor sites in which both values are one ($\lambda_{ij}^{(T)} = 0 \to \lambda_i^{(T+1)} j = 1$ iff $\sigma_i^{(T)} + \sigma_j^{(T)} = 2$ and $j \in \mathcal{N}\mathcal{N}_i^{(T)}$). The SDCA with these transition rules for connections, together with the *parity* rule for mass states, is implemented in Fig. 18, in which the initial Euclidean lattice with four neighbors (so the generic cell □ has eight next-nearest neighbors: ⁘) is seeded with a 3 × 3 block of ones. After the first iteration, most of the lattice structure has decayed as an effect of the decoupler rule, so that the active value cells and links are confined to a small region. After $T = 6$, the link and value structures become periodic, with a periodicity of two.

Memory can be embedded in links in a similar manner as in state values, so the link between any two cells is featured by a mapping of its previous link values: $l_{ij}^{(T)} = l(\lambda_{ij}^{(1)}, \ldots, \lambda_{ij}^{(T)})$. The *distance* between two cells in the historic model (d_{ij}), is defined in terms of l instead of λ values, so that i and j are *direct* neighbors if $d_{ij} = 1$, and are *next-nearest* neighbors if $d_{ij} = 2$. Now: $N_i^{(T)} = \{j/d_{ij}^{(T)} = 1\}$, and $NN_i^{(T)} = \{j/d_{ij}^{(T)} = 2\}$. Generalizing the approach to embedded memory applied to states, the unchanged transition rules (ϕ and ψ) operate on the featured link and cell state values: $\sigma_i^{(T+1)} = \phi(s_j^{(T)} \in N_i)$, $\lambda_{ij}^{(T+1)} = \psi(l_i^{(T)} j, s_i^{(T)}, s_j^{(T)})$.

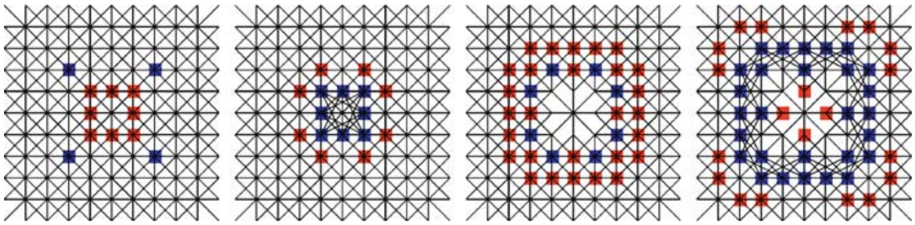
Figure 19 shows the effect of α-memory on the cellular automaton above introduced starting as in Fig. 18. The effect of memory on SDCA in the hexagonal and triangular tessellations is scrutinized in [11].

A plausible wiring dynamics when dealing with excitable CA is that in which the decoupler rule removes all links connected to cells in which both values are at refractory state ($\lambda_{ij}^{(T)} = 1 \to \lambda_{ij}^{(T+1)} = 0$ iff $\sigma_i^{(T)} = \sigma_j^{(T)}$=2) and the coupler rule adds links between all next-nearest neighbor sites in which both values are excited ($\lambda_{ij}^{(T)} = 0 \to \lambda_{ij}^{(T+1)} = 1$ iff $\sigma_i^{(T)} = \sigma_j^{(T)} = 1$ and $j \in \mathcal{N}\mathcal{N}_i^{(T)}$).

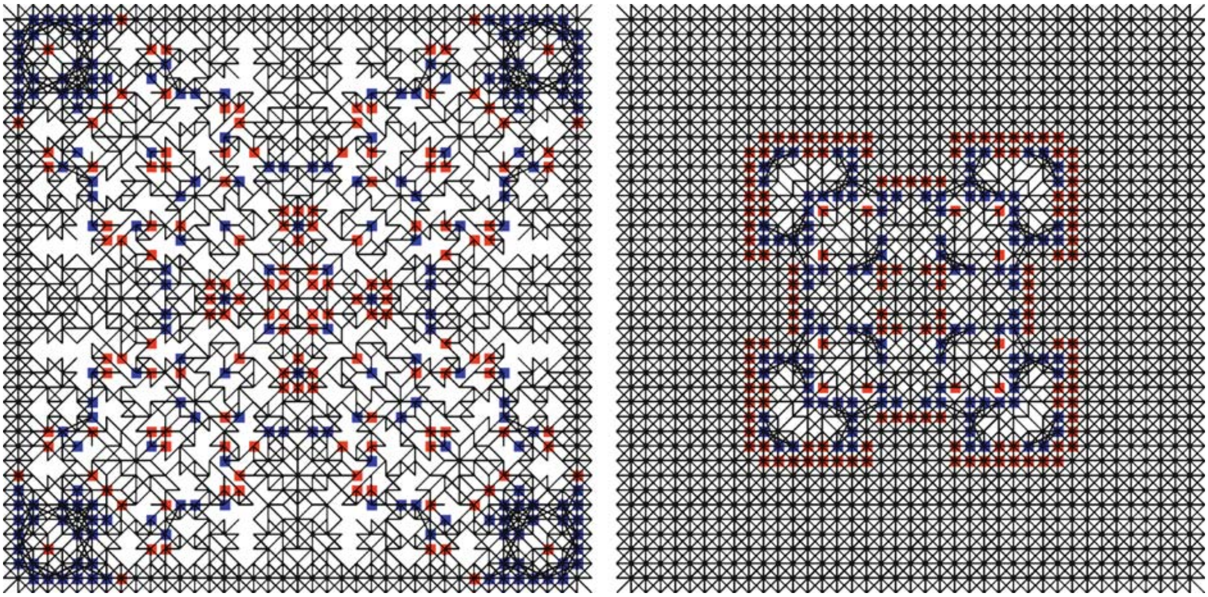
In the SDCA in Fig. 20, the transition rule for cell states is that of the generalized defensive inhibition rule: resting cell is excited if its ratio of excited and connected to the cell neighbors to total number of connected neighbors lies in



Cellular Automata with Memory, Figure 19
The SD cellular automaton introduced in text with weighted memory of factor α. Evolution from $T = 4$ up to $T = 9$ starting as in Fig. 18



Cellular Automata with Memory, Figure 20
The $k = 3$ SD cellular automaton described in text, up to $T = 4$



Cellular Automata with Memory, Figure 21
The SD cellular automaton starting as in Fig. 20 at $T = 20$, with no memory (*left*) and mode memory in both cell states and links

the interval [1/8,2/8]. The initial scenario of Fig. 20 is that of Fig. 12 with the wiring network revealed, that of an Euclidean lattice with eight neighbors, in which, the generic cell □ has 16 next-nearest neighbors: ▣. No decoupling is verified at the first iteration in Fig. 20, but the excited cells generate new connections, most of them lost, together with some of the initial ones, at $T = 3$. The excited cells at $T = 3$ generate a *crown* of new connections at $T = 4$. Figure 21 shows the ahistoric and mode memory patterns at $T = 20$. The figure makes apparent the preserving effect of memory.

The Fredkin's reversible construction is feasible in the SDCA scenario extending the $\ominus$ operation also to links: $\lambda_{ij}^{(T+1)} = \psi(\lambda_{ij}^{(T)}, \sigma_i^{(T)}, \sigma_j^{(T)}) \ominus \lambda_{ij}^{(T-1)}$. These automata may be endowed with memory as: $\sigma_i^{(T+1)} = \phi\big(s_j^{(T)} \in N_i^{(T)}\big) \ominus \sigma_i^{(T-1)}, \lambda_{ij}^{(T+1)} = \psi\big(l_{ij}^{(T)}, s_i^{(T)}, s_j^{(T)}\big) \ominus \lambda_{ij}^{(T-1)}$ [12].

The SDCA seems to be particularly appropriate for modeling the human brain function—updating links between cells imitates variation of synaptic connections between neurons represented by the cells—in which the relevant role of memory is apparent. Models similar to SDCA have been adopted to build a dynamical network approach to quantum space-time physics [45,46]. Reversibility is an important issue at such a fundamental physics level. Technical applications of SDCA may also be traced [47]. Anyway, besides their potential applications, SDCA with memory have an aesthetic and mathematical interest on their own [1,35]. Nevertheless, it seems plausible that further study on SDCA (and Lattice Gas Automata with dynamical geometry [40]) with memory should turn out to be profitable.

Memory in Other Discrete Time Contexts

Continuous-Valued CA

The mechanism of implementation of memory adopted here, keeping the transition rule unaltered but applying it to a function of previous states, can be adopted in any spatialized dynamical system. Thus, historic memory can be embedded in:

- *Continuous–valued* CA (or *Coupled Map Lattices* in which the state variable ranges in $\mathcal{R}$, and the transition rule φ is a continuous function [37]), just by considering m instead of σ in the application of the updating rule: $\sigma_i^{(T+1)} = \varphi\big(m_j^{(T)} \in \mathcal{N}_i^{(T)}\big)$. An elementary CA of this kind with memory would be [20]: $\sigma_i^{(T+1)} = \frac{1}{3}\big(m_{i-1}^{(T)} + m_i^{(T)} + m_{i+1}^{(T)}\big)$.
- *Fuzzy* CA, a sort of continuous CA with states ranging in the real [0,1] interval. An illustration of the effect of memory in fuzzy CA is given in [17]. The illustration operates on the elementary rule 90 : $\sigma_i^{(T+1)} = \big(\sigma_{i-1}^{(T)} \wedge (\neg\sigma_{i+1}^{(T)})\big) \vee \big((\neg\sigma_{i-1}^{(T)}) \wedge \sigma_{i+1}^{(T)}\big)$, which after fuzzification $(a \vee b \to \min(1, a+b), a \wedge b \to ab, \neg a \to 1-a)$ yields: $\sigma_i^{(T+1)} = \sigma_{i-1}^{(T)} + \sigma_{i+1}^{(T)} - 2\sigma_{i-1}^{(T)}\sigma_{i+1}^{(T)}$; thus incorporating memory: $\sigma_i^{(T+1)} = m_{i-1}^{(T)} + m_{i+1}^{(T)} - 2m_{i-1}^{(T)}m_{i+1}^{(T)}$.
- *Quantum* CA, such, for example, as the simple 1D quantum CA models introduced in [32]:

  $$\sigma_j^{(T+1)} = \frac{1}{N^{1/2}}\left(i\delta\sigma_{j-1}^{(T)} + \sigma_j^{(T)} + i\delta^*\sigma_{j+1}^{(T)}\right) ,$$

  which would become with memory [20]:

  $$\sigma_j^{(T+1)} = \frac{1}{N^{1/2}}\left(i\delta m_{j-1}^{(T)} + m_j^{(T)} + i\delta^* m_{j+1}^{(T)}\right) .$$

Spatial Prisoner's Dilemma

The Prisoner's Dilemma (PD) is a game played by two players (A and B), who may choose either to cooperate (C or 1) or to defect (D or 0). Mutual cooperators each score the *reward* R, mutual defectors score the *punishment* P ; D scores the *temptation* $\mathcal{T}$ against C, who scores S (*sucker's* payoff) in such an encounter. Provided that $\mathcal{T} > R > P > S$, mutual defection is the only equilibrium strategy pair. Thus, in a single round both players are to be *penalized* instead of both *rewarded*, but cooperation may be rewarded in an iterated (or spatial) formulation. The game is simplified (while preserving its essentials) if $P = S = 0$. Choosing $R = 1$, the model will have only one parameter: the *temptation* $\mathcal{T}$=b.

In the spatial version of the PD, each player occupies at a site (i,j) in a 2D lattice. In each generation the payoff of a given individual ($p_{i,j}^{(T)}$), is the sum over all interactions with the eight nearest neighbors and with its own site. In the next generation, an individual cell is assigned the decision ($d_{i,j}^{(T)}$) that received the highest payoff among all the cells of its Moore's neighborhood. In case of a tie, the cell retains its choice. The spatialized PD (SPD for short) has proved to be a promising tool to explain how cooperation can hold out against the ever-present threat of exploitation [43]. This is a task that presents problems in the classic *struggle for survival* Darwinian framework.

When dealing with the SPD, memory can be embedded not only in choices but also in rewards. Thus, in the *historic* model we dealt with, at T: (i) the payoffs coming from previous rounds are accumulated ($\pi_{i,j}^{(T)}$), and (ii) players are featured by a summary of past decisions ($\delta_{i,j}^{(T)}$). Again, in each round or generation, a given cell plays with each of the eight neighbors and itself, the decision δ in the cell of the neighborhood with the highest π being adopted. This approach to modeling memory

Cellular Automata with Memory, Table 1

Choices at $T = 1$ and $T = 2$; accumulated payoffs after $T = 1$ and $T = 2$ starting from a single defector in the SPD. $b > 9/8$

$\mathbf{d}^{(1)} = \delta^{(1)}$				
1	1	1	1	1
1	1	1	1	1
1	1	0	1	1
1	1	1	1	1
1	1	1	1	1

$\mathbf{p}^{(1)} = \pi^{(1)}$				
9	9	9	9	9
9	8	8	8	9
9	8	8b	8	9
9	8	8	8	9
9	9	9	9	9

$\mathbf{d}^{(2)} = \delta(2)$						
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	0	0	0	1	1
1	1	0	0	0	1	1
1	1	0	0	0	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

$\pi^{(2)} = \alpha\mathbf{p}^{(1)} + \mathbf{p}^{(2)}$						
$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 8$	$9\alpha + 7$	$9\alpha + 6$	$9\alpha + 7$	$9\alpha + 8$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$8\alpha + 5b$	$8\alpha + 3b$	$8\alpha + 5b$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$8\alpha + 3b$	8α	$8\alpha + 3b$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$8\alpha + 5b$	$8\alpha + 3b$	$8\alpha + 5b$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 8$	$9\alpha + 7$	$9\alpha + 6$	$9\alpha + 7$	$9\alpha + 8$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$

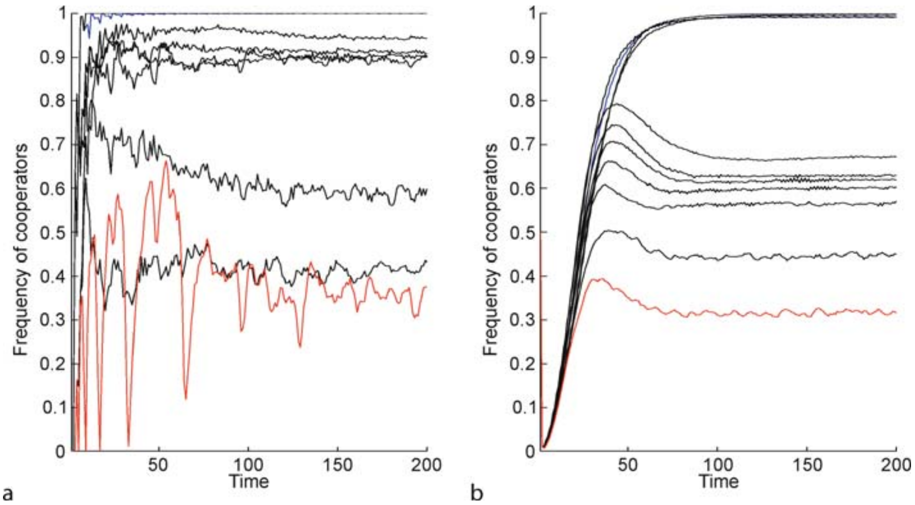
has been rather neglected, the usual being that of designing strategies that specify the choice for every possible outcome in the sequence of historic choices recalled [33,39].

Table 1 shows the initial scenario starting from a single defector if $8b > 9 \Leftrightarrow b > 1.125$, which means that neighbors of the initial defector become defectors at $T = 2$.

Nowak and May paid particular attention in their seminal papers to $b = 1.85$, a high but not excessive temptation value which leads to complex dynamics. After $T = 2$, defection can advance to a 5×5 square or be restrained as a 3×3 square, depending on the comparison of $8\alpha + 5 \times 1.85$ (the maximum π value of the recent defectors) with $9\alpha + 9$ (the π value of the non-affected players). As $8\alpha + 5 \times 1.85 = 9\alpha + 9 \rightarrow \alpha = 0.25$, i. e., if $\alpha > 0.25$, defection remains confined to a 3×3 square at $T = 3$. Here we see the paradigmatic effect of memory: it tends to avoid the spread of defection.

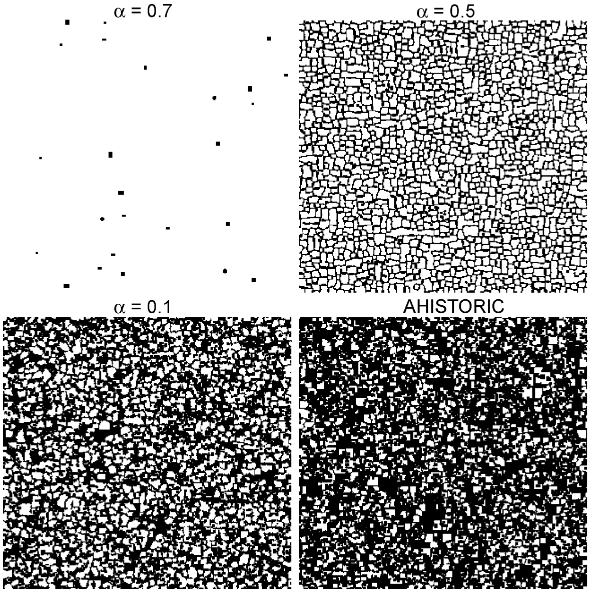
If memory is limited to the last three iterations: $\pi_{i,j}^{(T)} = \alpha^2 p_{i,j}^{(T-2)} + \alpha p_{i,j}^{(T-1)} + p_{i,j}^{(T)}$, $m_{i,j}^{(T)} = \left(\alpha^2 d_{i,j}^{(T-2)} + \alpha d_{i,j}^{(T-1)} + d_{i,j}^{(T)}\right)/(\alpha^2 + \alpha + 1)$, $\Rightarrow \delta_{i,j}^{(T)} = \text{round}(m_{i,j}^{(T)})$, with assignations at $T = 2$: $\pi_{i,j}^{(2)} = \alpha\pi_{i,j}^{(1)} + \pi_{i,j}^{(2)}$, $\delta_{i,j}^{(2)} = d_{i,j}^{(2)}$.

Memory has a dramatic restrictive effect on the advance of defection as shown in Fig. 22. This figure shows the frequency of cooperators (f) starting from a single defector and from a random configuration of defectors in a lattice of size 400×400 with periodic boundary conditions when $b = 1.85$. When starting from a single defector, f at time step T is computed as the frequency of cooperators within the square of size $(2(T - 1) + 1)^2$ centered on the initial D site. The ahistoric plot reveals the convergence of f to 0.318, (*which seems to be the same value regardless of the initial conditions* [43]). Starting from a single defector (*a*), the model with small memory ($\alpha = 0.1$) seems to reach a similar f value, but sooner and in a smoother way. The plot corresponding to $\alpha = 0.2$ still shows an early decay in f that leads it to about 0.6, but higher memory factor values lead f close to or over 0.9 very



Cellular Automata with Memory, Figure 22

Frequency of cooperators (f) with memory of the last three iterations. a starting from a single defector, b starting at random ($f(1) = 0.5$). The *red curves* correspond to the ahistoric model, the *blue ones* to the full memory model, the remaining curves to values of α from 0.1 to 0.9 by 0.1 intervals, in which, as a rule, the higher the α the higher the f for any given T



Cellular Automata with Memory, Figure 23
Patterns at $T = 200$ starting at random in the scenario of Fig. 22b

soon. Starting at random (b), the curves corresponding to $0.1 \leq \alpha \leq 0.6$ (thus with no memory of choices) do mimic the ahistoric curve but with higher f, as $\alpha \geq 0.7$ (also memory of choices) the frequency of cooperators grows monotonically to reach almost full cooperation: D persists as scattered unconnected small oscillators (*D-blinkers*), as shown in Fig. 23. Similar results are found for any temptation value in the parameter region $0.8 < b < 2.0$, in which spatial chaos is characteristic in the ahistoric model. It is then concluded that short-type memory supports cooperation.

As a natural extension of the described binary model, the 0-1 assumption underlying the model can be relaxed by allowing for *degrees* of cooperation in a continuous-valued scenario. Denoting by x the degree of cooperation of player A and by y the degree of cooperation of the player B, a consistent way to specify the pay-off for values of x and y other than zero or one is to simply interpolate between the extreme payoffs of the binary case. Thus, the payoff that the player A receives is:

$$G_A(x,y) = (x, 1-x)\begin{pmatrix} R & S \\ \mathcal{T} & P \end{pmatrix}\begin{pmatrix} y \\ 1-y \end{pmatrix} .$$

In the continuous-valued historic formulation it is $\delta \equiv m$, including $\delta^{(2)}_{i,j} = \left(\alpha d^{(1)}_{i,j} + d^{(2)}_{i,j}\right)/(\alpha+1)$. Table 2 illustrates the initial scenario starting from a single (full) defector. Unlike in the binary model, in which the initial defector never becomes a cooperator, the initial defector cooperates with degree $\alpha/(1+\alpha)$ at $T = 3$: its neighbors which received the highest accumulated payoff (those in the corners with $\pi^{(2)} = 8\alpha + 5b > 8b\alpha$), achieved this mean degree of cooperation after $T = 2$. Memory dramatically constrains the advance of defection in a smooth way, even for the low level $\alpha = 0.1$. The effect appears much more homogeneous compared to the binary model, with no special case for high values of α, as memory on decisions is always operative in the continuous-valued model [24]. The effect of unlimited trailing memory on the SPD has been studied in [5,6,7,8,9,10].

Discrete-Time Dynamical Systems

Memory can be embedded in any model in which time plays a dynamical role. Thus, Markov chains $\mathbf{p}'_{T+1} = \mathbf{p}'_T\mathbf{M}$ become with memory: $\mathbf{p}'_{T+1} = \boldsymbol{\imath}'_T\mathbf{M}$ with $\boldsymbol{\imath}_T$ being a weighted mean of the probability distributions up to T: $\boldsymbol{\imath}_T = \pi(\mathbf{p}_1, \ldots, \mathbf{p_T})$. In such scenery, even a minimal incorporation of memory notably alters the evolution of $\mathbf{p}$ [23]. Last but not least, conventional, non-spatialized, *discrete dynamical systems* become with memory: $x_{T+1} = f(m_T)$ with m_T being an average of past values. As an overall rule, memory leads the dynamics a fixed point of the map f [4].

We will introduce an example of this in the context of the PD game in which players follow the so-called Paulov strategy: a Paulov player cooperates if and only if both players opted for the same alternative in the previous move. The name Paulov stems from the fact that this strategy embodies an almost reflex-like response to the payoff: it repeats its former move if it was rewarded by $\mathcal{T}$ or R, but switches behavior if it was punished by receiving only P or S. By coding cooperation as 1 and defection as 0, this strategy can be formulated in terms of the choices x of Player A (Paulov) and y of Player B as: $x^{(T+1)} = 1- \mid x^{(T)} - y^{(T)} \mid$. The Paulov strategy has proved to be very successful in its contests with other strategies [44]. Let us give a simple example of this: suppose that Player B adopts an Anti-Paulov strategy (which cooperates to the extent Paulov defects) with $y^{(T+1)} = 1- \mid 1 - x^{(T)} - y^{(T)} \mid$. Thus, in an iterated Paulov-Anti-Paulov (PAP) contest, with $T(x,y) = \big(1 - |x-y|\,, 1 - |1-x-y|\big)$, it is $T(0,0) = T(1,1) = (1,0)$, $T(1,0) = (0,1)$, and $T(0,1) = (0,1)$, so that (0,1) turns out to be immutable. Therefore, in an iterated PAP contest, Paulov will always defect, and Anti-Paulov will always cooperate. Relaxing the 0-1 assumption in the standard formulation of the PAP contest, *degrees* of cooperation can be considered in a continuous-valued scenario. Now x and

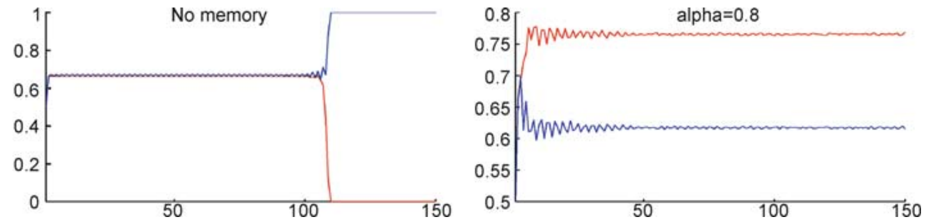
Cellular Automata with Memory, Table 2

Weighted mean degrees of cooperation after $T = 2$ and degree of cooperation at $T = 3$ starting with a single defector in the continuous-valued SPD with $b = 1.85$

$\delta^{(2)}$				
1	1	1	1	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	0	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	1	1	1	1

$\mathbf{d}^{(3)}(\alpha < 0.25)$						
1	1	1	1	1	1	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	1	1	1	1	1	1

$\mathbf{d}^{(3)}(\alpha < 0.25)$						
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		1
1	1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		1
1	1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		1
1	1	1	1	1	1	1
1	1	1	1	1	1	1



Cellular Automata with Memory, Figure 24

Dynamics of the mean values of *x* (*red*) and *y* (*blue*) starting from any of the points of the 1 × 1 square

y will denote the degrees of cooperation of players A and B respectively, with both x and y lying in [0,1].

In this scenario, not only (0,1) is a fixed point, but also $T(0.8, 0.6) = (0.8, 0.6)$. Computer implementation of the iterated PAP tournament turns out to be fully disrupting of the theoretical dynamics. The errors caused by the finite precision of the computer floating point arithmetics (a common problem in dynamical systems working modulo 1) make the final fate of every point to be (0,1). With no exceptions: even the *theoretically* fixed point (0.8,0.6) ends up as (0,1) in the computerized implementation.

A natural way to incorporate older choices in the strategies of decision is to feature players by a summary (m) of their own choices farther back in time. The PAP contest becomes in this way: $x^{(T+1)} = 1- \mid m_x^{(T)} - m_y^{(T)} \mid$, $y^{(T+1)} = 1- \mid 1 - m_x^{(T)} - m_y^{(T)} \mid$. The simplest historic extension results in considering only the two last choices: $m(z^{(T-1)}, z^{(T)}) = (\alpha z^{(T-1)} + z^{(T)})/(\alpha + 1)$ (z stands for both x and y) [10].

Figure 24 shows the dynamics of the mean values of x and y starting from any of the 101×101 lattice points of the 1×1 square with sides divided by 0.01 intervals. The dynamics in the ahistoric context are rather striking: immediately, at $T = 2$, both $\overline{x}$ and $\overline{y}$ increase from 0.5 up to app. $0.66(\simeq 2/3)$, a value which remains stable up to app. $T = 100$, but soon after Paulov cooperation plummets, with the corresponding firing of cooperation of Anti-Paulov: finite precision arithmetics leads every point to (0,1). With memory, Paulov not only keeps a permanent mean degree cooperation but it is higher than that of Anti-Paulov; memory tends to lead the overall dynamics to the ahistoric (theoretically) fixed point (0.8, 0.6).

Future Directions

Embedding memory in states (and in links if wiring is also dynamic) broadens the spectrum of CA as a tool for modeling, in a fairly natural way of easy computer implementation. It is likely that in some contexts, a transition rule with memory could match the *correct* behavior of the CA system of a given complex system (physical, biological, social and so on). A major impediment in modeling with CA stems from the difficulty of utilizing the CA complex behavior to exhibit a particular behavior or perform a particular function. Average memory in CA tends to inhibit *complexity*, inhibition that can be modulated by varying the depth of memory, but memory not of average type opens a notable new perspective in CA. This could mean a potential advantage of CA with memory over standard CA as a tool for modeling. Anyway, besides their potential applications, CA with memory (CAM) have an aesthetic and mathematical interest on their own.

Thus, it seems plausible that further study on CA with memory should turn out profitable, and, maybe that as a result of a further rigorous study of CAM it will be possible to paraphrase T. Toffoli in presenting CAM—as an alternative to (rather than an approximation of) integro-differential equations in modeling— phenomena with memory.

Bibliography

Primary Literature

1. Adamatzky A (1994) Identification of Cellular Automata. Taylor and Francis
2. Adamatzky A (2001) Computing in Nonlinear Media and Automata Collectives. IoP Publishing, London
3. Adamatzky A, Holland O (1998) Phenomenology of excitation in 2-D cellular automata and swarm systems. Chaos Solit Fract 9:1233–1265
4. Aicardi F, Invernizzi S (1982) Memory effects in Discrete Dynamical Systems. Int J Bifurc Chaos 2(4):815–830
5. Alonso-Sanz R (1999) The Historic Prisoner's Dilemma. Int J Bifurc Chaos 9(6):1197–1210
6. Alonso-Sanz R (2003) Reversible Cellular Automata with Memory. Phys D 175:1–30
7. Alonso-Sanz R (2004) One-dimensional, $r = 2$ cellular automata with memory. Int J Bifurc Chaos 14:3217–3248
8. Alonso-Sanz R (2004) One-dimensional, $r = 2$ cellular automata with memory. Int J BifurcChaos 14:3217–3248
9. Alonso-Sanz R (2005) Phase transitions in an elementary probabilistic cellular automaton with memory. Phys A 347:383–401 Alonso-Sanz R, Martin M (2004) Elementary Probabilistic Cellular Automata with Memory in Cells. Sloot PMA et al (eds) LNCS, vol 3305. Springer, Berlin, pp 11–20
10. Alonso-Sanz R (2005) The Paulov versus Anti-Paulov contest with memory. Int J Bifurc Chaos 15(10):3395–3407
11. Alonso-Sanz R (2006) A Structurally Dynamic Cellular Automaton with Memory in the Triangular Tessellation. Complex Syst 17(1):1–15. Alonso-Sanz R, Martin, M (2006) A Structurally Dynamic Cellular Automaton with Memory in the Hexagonal Tessellation. In: El Yacoubi S, Chopard B, Bandini S (eds) LNCS, vol 4774. Springer, Berlin, pp 30-40
12. Alonso-Sanz R (2007) Reversible Structurally Dynamic Cellular Automata with Memory: a simple example. J Cell Automata 2:197–201
13. Alonso-Sanz R (2006) The Beehive Cellular Automaton with Memory. J Cell Autom 1:195–211
14. Alonso-Sanz R (2007) A Structurally Dynamic Cellular Automaton with Memory. Chaos Solit Fract 32:1285–1295
15. Alonso-Sanz R, Adamatzky A (2008) On memory and structurally dynamism in excitable cellular automata with defensive inhibition. Int J Bifurc Chaos 18(2):527–539
16. Alonso-Sanz R, Cardenas JP (2007) On the effect of memory in Boolean networks with disordered dynamics: the $K = 4$ case. Int J Modrn Phys C 18:1313–1327
17. Alonso-Sanz R, Martin M (2002) One-dimensional cellular automata with memory: patterns starting with a single site seed. Int J Bifurc Chaos 12:205–226
18. Alonso-Sanz R, Martin M (2002) Two-dimensional cellular automata with memory: patterns starting with a single site seed. Int J Mod Phys C 13:49–65
19. Alonso-Sanz R, Martin M (2003) Cellular automata with accumulative memory: legal rules starting from a single site seed. Int J Mod Phys C 14:695–719
20. Alonso-Sanz R, Martin M (2004) Elementary cellular automata with memory. Complex Syst 14:99–126
21. Alonso-Sanz R, Martin M (2004) Three-state one-dimensional cellular automata with memory. Chaos, Solitons Fractals 21:809–834
22. Alonso-Sanz R, Martin M (2005) One-dimensional Cellular Automata with Memory in Cells of the Most Frequent Recent Value. Complex Syst 15:203–236
23. Alonso-Sanz R, Martin M (2006) Elementary Cellular Automata with Elementary Memory Rules in Cells: the case of linear rules. J Cell Autom 1:70–86
24. Alonso-Sanz R, Martin M (2006) Memory Boosts Cooperation. Int J Mod Phys C 17(6):841–852
25. Alonso-Sanz R, Martin MC, Martin M (2000) Discounting in the Historic Prisoner's Dilemma. Int J Bifurc Chaos 10(1):87–102
26. Alonso-Sanz R, Martin MC, Martin M (2001) Historic Life Int J Bifurc Chaos 11(6):1665–1682
27. Alonso-Sanz R, Martin MC, Martin M (2001) The Effect of Memory in the Spatial Continuous-valued Prisoner's Dilemma. Int J Bifurc Chaos 11(8):2061–2083
28. Alonso-Sanz R, Martin MC, Martin M (2001) The Historic Strategist. Int J Bifurc Chaos 11(4):943–966
29. Alonso-Sanz R, Martin MC, Martin M (2001) The Historic-Stochastic Strategist. Int J Bifurc Chaos 11(7):2037–2050
30. Alvarez G, Hernandez A, Hernandez L, Martin A (2005) A secure scheme to share secret color images. Comput Phys Commun 173:9–16
31. Fredkin E (1990) Digital mechanics. An informal process based on reversible universal cellular automata. Physica D 45:254–270
32. Grössing G, Zeilinger A (1988) Structures in Quantum Cellular Automata. Physica B 15:366
33. Hauert C, Schuster HG (1997) Effects of increasing the number of players and memory steps in the iterated Prisoner's Dilemma, a numerical approach. Proc R Soc Lond B 264:513–519
34. Hooft G (1988) Equivalence Relations Between Deterministic and Quantum Mechanical Systems. J Statistical Phys 53(1/2):323–344
35. Ilachinski A (2000) Cellular Automata. World Scientific, Singapore
36. Ilachinsky A, Halpern P (1987) Structurally dynamic cellular automata. Complex Syst 1:503–527
37. Kaneko K (1986) Phenomenology and Characterization of coupled map lattices, in Dynamical Systems and Sigular Phenomena. World Scientific, Singapore
38. Kauffman SA (1993) The origins of order: Self-Organization and Selection in Evolution. Oxford University Press, Oxford
39. Lindgren K, Nordahl MG (1994) Evolutionary dynamics of spatial games. Physica D 75:292–309
40. Love PJ, Boghosian BM, Meyer DA (2004) Lattice gas simulations of dynamical geometry in one dimension. Phil Trans R Soc Lond A 362:1667
41. Margolus N (1984) Physics-like Models of Computation. Physica D 10:81–95

42. Martin del Rey A, Pereira Mateus J, Rodriguez Sanchez G (2005) A secret sharing scheme based on cellular automata. Appl Math Comput 170(2):1356–1364
43. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. Nature 359:826
44. Nowak MA, Sigmund K (1993) A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature 364:56–58
45. Requardt M (1998) Cellular networks as models for Plank-scale physics. J Phys A 31:7797; (2006) The continuum limit to discrete geometries, arxiv.org/abs/math-ps/0507017
46. Requardt M (2006) Emergent Properties in Structurally Dynamic Disordered Cellular Networks. J Cell Aut 2:273
47. Ros H, Hempel H, Schimansky-Geier L (1994) Stochastic dynamics of catalytic CO oxidation on Pt(100). Pysica A 206:421–440
48. Sanchez JR, Alonso-Sanz R (2004) Multifractal Properties of R90 Cellular Automaton with Memory. Int J Mod Phys C 15:1461
49. Stauffer D, Aharony A (1994) Introduction to percolation Theory. CRC Press, London
50. Svozil K (1986) Are quantum fields cellular automata? Phys Lett A 119(41):153–156
51. Toffoli T, Margolus M (1987) Cellular Automata Machines. MIT Press, Massachusetts
52. Toffoli T, Margolus N (1990) Invertible cellular automata: a review. Physica D 45:229–253
53. Vichniac G (1984) Simulating physics with Cellular Automata. Physica D 10:96–115
54. Watts DJ, Strogatz SH (1998) Collective dynamics of Small-World networks. Nature 393:440–442
55. Wolf-Gladrow DA (2000) Lattice-Gas Cellular Automata and Lattice Boltzmann Models. Springer, Berlin
56. Wolfram S (1984) Universality and Complexity in Cellular Automata. Physica D 10:1–35
57. Wuensche A (2005) Glider dynamics in 3-value hexagonal cellular automata: the beehive rule. Int J Unconv Comput 1:375–398
58. Wuensche A, Lesser M (1992) The Global Dynamics of Cellular Automata. Addison-Wesley, Massachusetts

Books and Reviews

Alonso-Sanz R (2008) Cellular Automata with Memory. Old City Publising, Philadelphia (in press)

Cellular Automata Modeling of Complex Biochemical Systems

Lemont B. Kier[1], Paul G. Seybold[2]

[1] Center for the Study of Biological Complexity,
Virginia Commonwealth University, Richmond, USA

[2] Wright State University, Dayton, USA

Article Outline

Glossary

Agents Ingredients under study occupying the cells in the grid.

Asynchronous Each cell in turn, at random responds to a rule.

Cells Sections of the grid in which the agents exist.

Gravity The relative relationship of two agents within the frame of a grid with opposite bound edges.

Grid The frame containing the cells and the agents.

Movement The simulation of agent movement between cells accomplished by the disappearance of an agent in a cell and the appearance of that agent in an adjacent cell.

Neighborhood The cells immediately touching a given cell in the grid.

Rules Statements of actions to be taken by an agent under certain conditions. They may take the form of a probability of such an action.

Synchronous All cells in a grid exercise a simultaneous response to a rule.

Definition of the Subject

Cellular automata are discrete, agent-based models that can be used for the simulation of complex systems [1]. They are composed of:

- A grid of cells.
- A set of ingredients called agents that can occupy the cells.
- A set of local rules governing the behaviors of the agents.
- Specified initial conditions.

Once these components are defined, a simulation can be carried out. During the simulation the system evolves via a series of discrete time-steps, or iterations, in which the rules are applied to all of the ingredients of the system and the configuration of the system is regularly updated. A striking feature of the cellular automata (CA) models is that they treat not only the ingredients, or agents, of the model as discrete entities, as do the traditional models of physics and chemistry, but in the CA models time (iterations) and space (the cells) are also regarded as discrete, in contrast to the continuous forms assumed for these variables in the traditional, equation-based physical models. In practice, this distinction usually makes little difference, since the traditional continuous results appear naturally as limiting cases of the discrete CA analysis.

Introduction

Cellular automata (CA) were first proposed by the mathematical physicist John von Neumann [2] and the mathematician Stanislaw Ulam [3,4] more than a half century ago, and similar ideas were suggested at about the same time, in the 1940s, by the German engineer Konrad Zuse. Von Neumann's interest was in the construction of "self-reproducing automata". His original idea was to construct a series of mechanical devices or "automata" that would gather and assemble the parts necessary to reproduce themselves. A suggestion by Ulam led him to consider more abstract systems consisting of grids with moving agents, operating under sets of rules. The first such system proposed by von Neumann consisted of a two-dimensional grid of square cells, each having a set of possible states, along with a set of rules. With the development of modern digital computers, it has became increasingly clear that these very abstract ideas could in fact be usefully applied to the examination of real physical and biological systems, with interesting and informative results.

A number of research groups have developed different realizations of the CA paradigm for the study of a broad range of physical [5,6], chemical [7], biological [8,9], medical [10,11] and even sociological [12,13,14] phenomena. These models have contributed important new insights regarding the deeper, often hidden, factors underlying a host of complex phenomena. These diverse CA studies have been especially important in treating the often-surprising behaviors of systems where large numbers of diverse interactions between the system agents serve to hide the general patterns involved and, in addition, render the conventional, differential-equation-based methods difficult to implement or ineffective—i. e., complex systems.

Models and Simulations with Cellular Automata

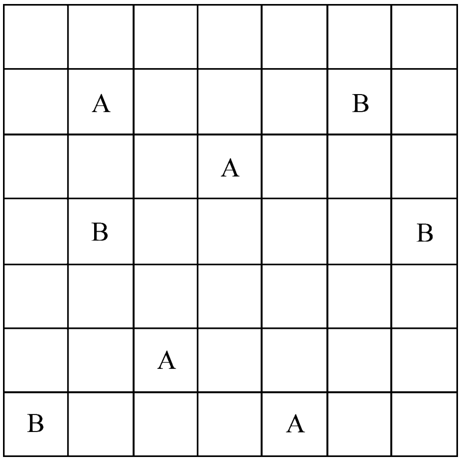
Cellular automata can be used to both model and to simulate real systems. A model is a substitute, usually greatly simplified, for what is the real thing. A model should capture from the real system and display in some revealing way its most important or interesting features. Where possible it should capture the essence of the system without being overly cumbersome or complicated. Many models in the physical sciences take the form of mathematical relationships, equations connecting some property with other parameters of the system. Some of these relationships are quite simple, while many mathematical relationships are more complicated, and rely on the techniques of calculus to describe the rates of change of the quantities involved. Overall, mathematical models have been exceedingly successful in depicting the broad outlines of an enormously diverse variety of phenomena in nature.

Simulations are active imitations of real things, and there are generally two different types of simulations, with different aims. In one approach a simulation is merely designed to match a certain behavior, often in a very limited context. Thus a mechanical noisemaker may simulate a desired sound, and does so through a very different mechanism than the real sound-maker. Such a simulation reveals little or nothing about the features of the original system, and is not intended to do so. Only the outcome, to some extent, matches reality. A hologram may look like a real object, but it is constructed from interfering light waves. A second type of simulation is more ambitious. It attempts to mimic at least some of the key features of the system under study, with the intent of gaining insight into how the system operates.

The Grid

The grid in a CA model may contain a single cell, or more commonly a larger collection. The grid might be one-, two-, or three-dimensional in form, although most studies have used two-dimensional grids. A moving agent may encounter an edge or boundary during its response to the rules. Three general types of two-dimensional grids are considered relating to the boundaries: (i) a box, (ii) a cylinder, and (iii) a torus. In the box grid moving agents encounter boundaries on all four sides; in the cylinder they encounter only top and bottom boundaries; and in the torus no boundaries restrict the agent movements. An illustration of a $7 \times 7 = 49$-cell grid of square cells occupied by two different types of agents, A and B, is shown in Fig. 1.

The nature of the grid type employed will normally depend on the boundary characteristics of the system of interest. For some systems, e. g., when the agents are either stationary or confined, a box grid is perfectly suitable. In other cases, one may need only a constraining top and bottom (or right and left sides), and a cylindrical grid will be used. The torus effectively simulates a small segment of a larger, unrestricted system by allowing cells to move off the top edge and appear at the bottom, or move off the bottom edge and appear at the top.



	A				B	
			A			
	B					B
		A				
B				A		

Cellular Automata Modeling of Complex Biochemical Systems, Figure 1
A two dimensional cellular automata grid with two different agents

The Cells

Structure

The cells can take a variety of shapes; they can be triangles, squares, hexagons or other shapes on the two-dimensional grid, with square cells being most common. Each cell in the grid can normally exist in a number of distinct states which define the occupancy of the cell. The cell can be empty or contain a specific agent, representing a particle, a type of molecule or isomer, a molecular electronic state, or some other entity pertinent to the study in question. The choice of the cell shape is based on the objective of the study. In the case of studies of water-related phenomena, for example, square cells are especially advantageous since water molecules, are quadravalent with respect to their participation in intermolecular hydrogen bonding. An individual water molecule can employ two hydrogen atoms and two lone pairs of electrons to form hydrogen bonds with its neighbors. This leads to the tetrahedral configuration found in ice, a structure that is retained to some extent in the liquid state. The four faces of a square cell thus correspond to the four hydrogen-bonding opportunities of a water molecule.

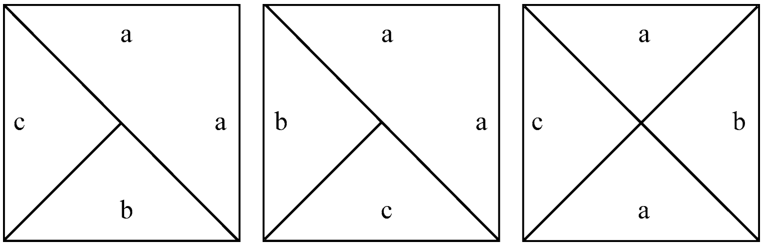
The interactions of an agent with other agents take place at the cell edges. Originally, cellular automata models routinely assumed that all of the edges of a given agent should obey the same rules. More recently, the idea of a variegated cell, in which each edge can have its own independent rules for interaction with other agents, has been introduced and shown to have considerable value in modeling. Examples of some types of variegated cells are shown in Fig. 2.

Cell Neighborhoods

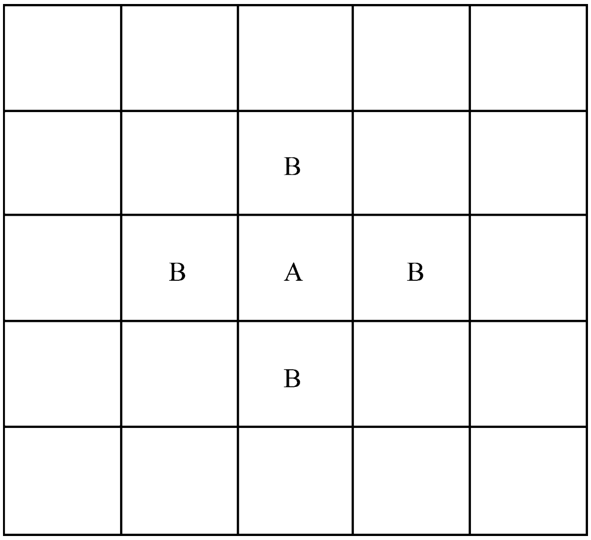
The movements and other actions of an agent on the grid are governed by rules, that depend only on the nature of the cells in close proximity to the agent. This proximate environment of a cell is called its neighborhood. The most common neighborhood used in two-dimensional cellular automata studies is called the von Neumann neighborhood, after the pioneer of the CA method. This neighborhood for a cell A refers to the four B cells adjoining its four faces (see Fig. 3). Another common neighborhood is the Moore neighborhood referring to the eight cells completely surrounding cell A, including those cells on the diagonals.

The Rules

Several different types of rules govern the behaviors of the agents on the grid and thereby the subsequent evolutions of the CA systems. The key features of all these rules are that they are local, involving only an agent itself and possibly those agents in its immediate neighborhood, and that they are uniformly applied throughout a CA study.



Cellular Automata Modeling of Complex Biochemical Systems, Figure 2
Cellular automata variegated cells



		B		
	B	A	B	
		B		

Cellular Automata Modeling of Complex Biochemical Systems, Figure 3
The von Neumann neighborhood

Movement Rules

Much of the dynamic character of cellular automata models is developed through the movements of the agents about the grid. During each time-step interval, or iteration, in the CA simulation an agent on the grid has the possibility of moving vertically or horizontally to an adjacent, unoccupied cell. In the absence of further restrictions, a free agent would therefore, over time, perform a random walk about the grid. Normally, however, there are other agents on the grid and their presence will influence the motion of the first agent. During each iteration the movement of every agent on the grid is computed based on rules that involve the status of its neighboring cells, i. e., whether these cells are empty or occupied, and, if occupied, by what types of agents. Deterministic cellular automata use a fixed set of rules, the values of which are immutable and uniformly applied to the agents. In probabilistic, or stochastic, cellular automata, the movements of the agent are based on probabilistic rules, embodied as probabilities of moving or not moving during each iteration.

Free Moving Probability P_m

The free moving probability $P_m(A)$ defines the probability that an agent A in a cell *i* will move to one of the four adjacent cells, *j*, in its von Neumann neighborhood if that space is unoccupied. An example would be the agent in cell A which might move to any of the unoccupied neighboring B cells. This probability is usually set at $P_m = 1.0$, which means that a movement in one of the allowed directions always happens. However, in some cases P_m can be set to lower values if certain species in the CA simulation are to be regarded as moving more slowly than others.

Joining Parameter *J*

The first of the two trajectory or interaction rules is the joining trajectory parameter, $J(AB)$, which defines the propensity of movement of an agent A toward or away from a second agent B when the two are separated by a vacant cell. It thus involves the extended von Neumann neighborhood of agent A, and has the effect of adding a short-range attraction or repulsion component to the interaction between agents A and B. J is a non-negative real number. When $J = 1$, species A has the same probability of movement toward or away from B as when the B cell is not present. When J is greater than 1, agent A has a greater probability of movement toward a B agent than when agent B is absent, simulating, in effect, a degree of short-range attraction. When J lies between 0 and 1, agent A has a lower probability of such movement, and this can be considered as a degree of mutual repulsion. When $J = 0$, agent A cannot move toward B at all.

Breaking Probability P_B

The second trajectory or interaction rule is the breaking probability, P_B. This parameter, in effect, assigns a persistence to the encounter between two agents that are in contact, i. e., adjacent to each other on the grid. The breaking rule assigns the probability $P_B(AB)$ that an agent A, adjacent to an agent B, will break apart from B. The value for P_B necessarily lies within the range 0 to 1. Low values of P_B imply a strong cohesion between A and B, whereas high values indicate little cohesion. Thus if $P_B = 0$, the agents will not separate from each other, and if $P_B = 1$ they have no tendency to adhere to one another. If P_B lies between these values there is an intermediate tendency to break apart. When molecule A is bordered two agents, B and C, the simultaneous probability of A breaking away is given by the product $P_B(AB) * P_B(AC)$. If agent A has three adjacent agents (B, C, and D), the simultaneous breaking probability of agent A, the probability that it will move to the remaining adjacent empty cell, is $P_B(AB) * P_B(AC) * P_B(AD)$. If agent A is surrounded by four agents, it cannot move.

Transition Rules

Transitions occur constantly in nature; molecules change from one tautomeric form to another, radioactive nuclei decay to form other nuclei, acids dissociate, proteins alter their shapes, molecules undergo transitions between electronic states, chemicals react to form new species, and so forth. Transition rules allow the simulation of these changes. They govern the probability that during each iteration of the simulation an agent will transform to a different type of agent. If $P_{\mathrm{T}}(\mathrm{AB}) = 1.0$ the transition A → B is certain to occur; if $P_{\mathrm{T}}(\mathrm{AB}) = 0.0$, it will never occur. But if, for example, $P_{\mathrm{T}}(\mathrm{AB}) = 0.5$, then during each iteration there will be a 50% chance that the transition A → B will occur. The first two cases can be considered deterministic, since they do not allow for different outcomes. The third case is stochastic, however, since it allows different outcomes, the agent might remain unchanged or it might transform to a different state. The transition probabilities may, in some cases, depend on the conditions prevailing in neighboring cells. For example, the transformation probability $P_{\mathrm{T}}(\mathrm{AB})$ might depend on the occupancies of neighboring cells. In reaction simulations two agents A and B that come in contact on the grid will have a probability $P_{\mathrm{R}}(\mathrm{AB})$ of reacting, or transforming, to other species, say, C and D, during such an encounter. In this case the reaction probability $P_{\mathrm{R}}(\mathrm{AB})$ defines the probability that the reaction $\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + \mathrm{D}$ will occur when A and B encounter one another in the course of their motions. If $P_{\mathrm{R}}(\mathrm{AB}) = 1$ the reaction will take place on every encounter, but if $P_{\mathrm{R}}(\mathrm{AB}) = 0.1$, for example, only 10% of such encounters will lead to reaction.

Relative Gravity Rules

The simulation of a gravity effect can be introduced into a cellular automaton model in two different ways. Separation phenomena like the de-mixing of immiscible liquids can be simulated using a relative gravity rule. For this, a boundary condition is first imposed at the upper and lower edges of the grid to apply vertical limits on the motions of the agents (a cylindrical grid). The differential effect of gravity on different agents A and B is simulated by introducing reciprocal rules governing their tendencies to exchange positions when they come together. When one agent moves to a position on top of the other the rules are applied. The first rule, $G_{\mathrm{R}}(\mathrm{AB})$, applies when A is above B and is the probability that agent A will exchange places with agent B, so that A will appear below, and B above. The complementary rule is $G_{\mathrm{R}}(\mathrm{BA})$, which expresses the probability that molecule B, originally above A, will exchange positions with A and end up below.

When $G_{\mathrm{R}}(\mathrm{AB})$ is greater than $G_{\mathrm{R}}(\mathrm{BA})$ there will be an overall tendency for the A agents to congregate below the B agents, and when $G_{\mathrm{R}}(\mathrm{AB})$ is less than $G_{\mathrm{R}}(\mathrm{BA})$ the A agents will tend toward the upper part of the grid. In the first case the A's can be thought of as forming a more dense liquid than the B's, and in the latter case, a less dense liquid. The G_{R} rules are probabilities that the events will occur.

Absolute Gravity Rule

In other simulations an absolute gravity rule, denoted by $G_{\mathrm{A}}(\mathrm{A})$, is more appropriate. This rule favors motion in a preferred direction. For example, one might wish to simulate the motions of different gas molecules, some heavier than others, in a gravitational field. The value $G_{\mathrm{A}}(\mathrm{A}) = 0$ is the neutral value, so that the movement probabilities are equal in all four directions. Values greater than $G_{\mathrm{A}}(\mathrm{A}) = 0$ increase the likelihood of downward movements. Thus a value of $G_{\mathrm{A}}(\mathrm{A}) = 0.2$ would impose a slight tendency of the agents of species A to move downward on the grid, and $G_{\mathrm{A}}(\mathrm{A}) = 0.5$ would impose a much stronger tendency.

Cell Rotation Rules

In those cases where variegated agents are used it is necessary to insure that there exists a balanced representation of the possible rotational states of these agents on the grid. To accomplish this, the variegated cells are rotated randomly, by 90°, −90°, +180°, or −180°, during every iteration of the run. Only free cells rotate; when a variegated agent has a neighboring agent in its von Neumann neighborhood it does not rotate.

Application of the Rules

A complete time-step (iteration) in a CA model involves the application of all of the applicable rules to all of the agents on the grid. During an iteration the movement rules can be applied either simultaneously (synchronously) or sequentially (asynchronously). Synchronous application of the governing movement rules for a CA simulation, as outlined above, can lead in some instances to conflicts, e. g., the assigning of two agents to move to the same empty cell. As a result, synchronous rule application may not be practical for cellular automata studies of some kinds. In asynchronous application of the rules, in contrast, agents are selected in random order for application of the movement rules, and such potential conflicts are avoided.

Examples

Structure of Bulk Water

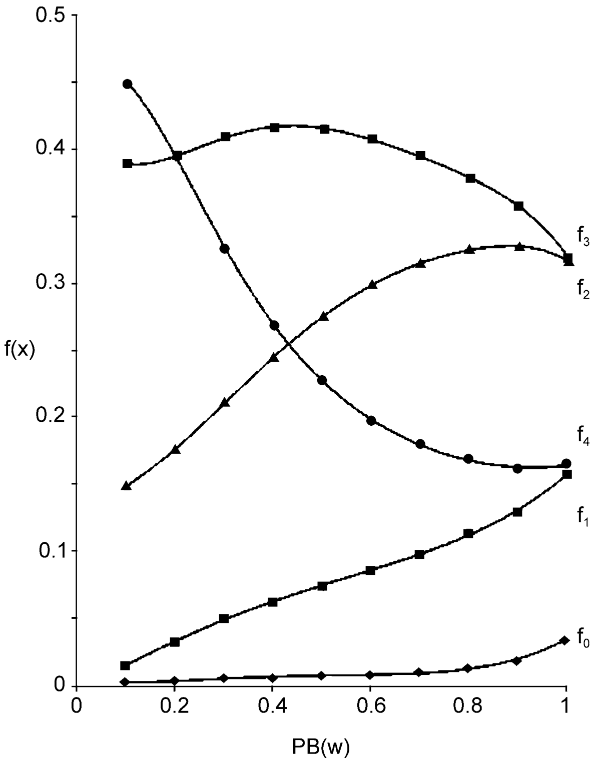
The physical properties of water, such as its viscosity, heat capacity, heat of vaporization, surface tension, dielectric constant, etc., are commonly related to temperature under normal pressure. It is accepted that underlying changes in the structure of water are responsible for the changes in the properties, although the precise physical basis for this relationship has not been established. Cellular automata models of bulk water in the liquid state at different temperatures produce a series of structural patterns, differing in the extent of binding between the individual water molecules. The fractions of water molecules unbound, bound to one, two, three or four neighbors, designated, f_0, f_1, f_2, f_3, and f_4, can be employed as structural descriptors of bulk liquid water at various temperatures. The significance of these descriptors can be evaluated by examining the correlations of these descriptors with these physical properties. A linkage between the J and P_B rules for water models has been postulated and tested leading to the relationship [15]:

$$\mathrm{Log}\, J = -1.5 P_B + 0.6\,. \tag{1}$$

The temperature of the water has been found to relate with the P_B value as

$$\mathrm{T}\left({}^\circ\mathrm{C}\right) = 100 P_B\,. \tag{2}$$

Using the f_x parameters as structure parameters in regression analysis, correlations arise between them and a variety of physical properties. Several relationships are given in Table 1, where r^2 is the coefficient of determination and s is the standard error of the fit. The observed relationships suggest that the bulk water f_x descriptors encode information allowing them to be used in correlations with a variety of physical properties. This reveals possibilities of using these descriptors in analyzing properties and behavior entirely on the basis of structure rather than using an inter-connected web of physical properties. The relationship between the f_x descriptors and the temperature (P_B value) is shown in Fig. 4.



Cellular Automata Modeling of Complex Biochemical Systems, Figure 4
The calculated distribution of water bonding states, *f(x)*, and the modeled temperature using *PB(W)* values

Solution Phenomena

Solute Rules The solubility of a compound in a particular solvent reflects the intermolecular interactions between the molecular species involved. Thus the breaking and joining rules pertaining to solute-solute and solute-solvent interactions may be assigned on the basis of some anticipation of the magnitude of these intermolecular terms. For the solute (S), a large value of J(SS) encodes a high tendency for its molecules to be attracted to one another.

Cellular Automata Modeling of Complex Biochemical Systems, Table 1
Equations relating f_x values to properties of water

Property	Equation	r^2	s
Temperature	$T\,(^\circ\mathrm{C}) = -490.28 f_0 + 622.60 f_1 + 4.46$	0.996	1.90
Heat capacity	$C_p\,[\mathrm{Cal/g/^\circ C}] = 1.0478 - 0.0488 f_2 - 0.0620 f_3 - 0.0446 f_4$	0.995	0.0002
Surface tension	$\gamma\,(\mathrm{dynes/cm}) = -93.72 f_1 + 75.33$	0.996	1.90
Vapor pressure	$\mathrm{Log}\,P_v\,(\mathrm{mm\,Hg}) = -24.30 f_0 + 15.64 f_1 + 0.90$	0.997	0.035
Compressibility	$\kappa\,(10^6/\mathrm{Bar}) = 79.60 - 43.61 f_2 - 39,57 f_3 - 30.32 f_4$	0.991	0.190
Dielectric constant	$\varepsilon = -178.88 f_1 + 55.84$	0.994	0.735
Viscosity	$\eta\,(\mathrm{centipoises}) = 1.439 f_4 + 0.202$	0.965	0.045

The companion rule, $P_B(SS)$, encodes the ease or difficulty of a pair of joined solute molecules to break apart. To illustrate the cooperativity of these two rules for a solute, a molecule with rules $J(SS) = 2.0$ and $P_B(SS) = 0.2$ would be expected to have a strong tendency to remain as an insoluble aggregate or crystal. Such a solute would likely have a high melting point. In contrast, a solute with $J(SS) = 0.5$ and $P_B(SS) = 0.8$ would be expected to have a lower melting point and be more soluble. The choice of these rules govern the general behavior of solute molecules toward themselves [16].

The Hydrophobic Effect The hydrophobic effect refers to the action of relatively non-polar (hydrophobic) substances on the organization of the water molecules in their vicinity. A common expression is that water becomes "more structured" or organized in the neighborhood of a hydrophobic solute. This phenomenon has been simulated using a cellular automata model [17]. In this model the breaking probability of water-solute molecule pairs, $P_B(WS)$, was systematically increased, thus encoding an increasing hydrophobicity of the solute and a decreasing probability that the solute molecules would associate with the water molecules.

It was observed that low $P_B(WS)$ values, representing strong water-solute attractions, produced solution configurations in which the solute molecules were heavily surrounded by water molecules. Conversely, at high $P_B(WS)$ values most of the solute molecules were found outside of the water clusters and within the solution cavities. These configurations leave the water clusters relatively free of solute and hence more internally structured or organized. These models thus reflect the molecular system level conditions present in the hydrophobic effect. These results agree with molecular dynamics simulations and the interpretation presented agrees with recent experimental evidence and models proposed for the hydrophobic effect.

Solute Dissolution Dissolution refers to the breaking apart of a solute aggregate within a solvent to form a solution. Cellular automata simulations of the dissolution process have been carried out in which the solute aggregates started as solid blocks surrounded by water [18]. The solute species were endowed with specific $P_B(SS)$, $J(SS)$, $P_B(WS)$ and $J(WS)$ rules. The system attributes recorded during the ensuing system evolutions were the fraction $f_0(S)$ of solutes unbound to other solute molecules, the average number of solute-solute joined faces $T(S)$, and the average distance $D(S)$ that solute molecules traveled from the center of the block at a specific iteration. The $f_0(S)$ values were interpreted as representing the extent of dissolution of the solute, the decrease in the $T(S)$ values were used to characterize the extent of disruption of the solute block, and the $D(S)$ values quantified the extent of diffusion of the solutes into the surrounding water.

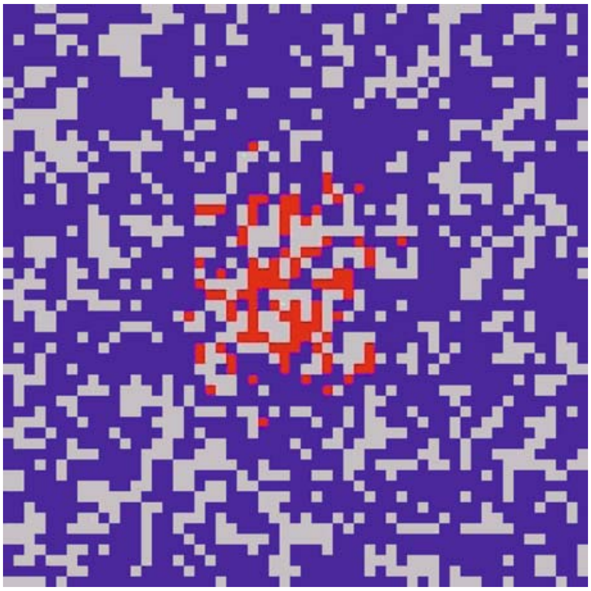
The extent and rate of the solute block disruption, $T(S)$, is primarily a function of the $P_B(S)$ rule, with secondary influence from the $P_B(WS)$ rule. A high value of $P_B(S)$ implies a high probability of solute molecules separating from each other, and hence a strong tendency toward crystal disruption. The f_0 fraction serves as a measure of the extent of solute solubility. The extent of dissolution depends upon both the $P_B(S)$ and the $P_B(WS)$ rules. High values of $P_B(S)$ and low values of $P_B(WS)$ promote an extensive degree of dissolution. A low value of $P_B(WS)$ characterizes a solute that is relatively hydrophilic. The simulations also implied that solutes with high $P_B(WS)$ values should diffuse more rapidly than those with low values of equal size. It was also observed that simulations of higher temperatures led to faster disruption of the block of solute, and more extensive diffusion of the solute molecules through the bulk water, in accord with the normal expectations.

An unexpected observation arises from the graphical display of the disintegration of the solute block of cells. The early stages of the disruption occurred as a series of intrusions of cavities, rather than water molecules, into the block. The cavities roamed throughout the block, behaving as "particles". The entrance of water into the block structure appeared much later, after significant disruption and loss of solute molecules from the block. This behavior is shown in Fig. 5.

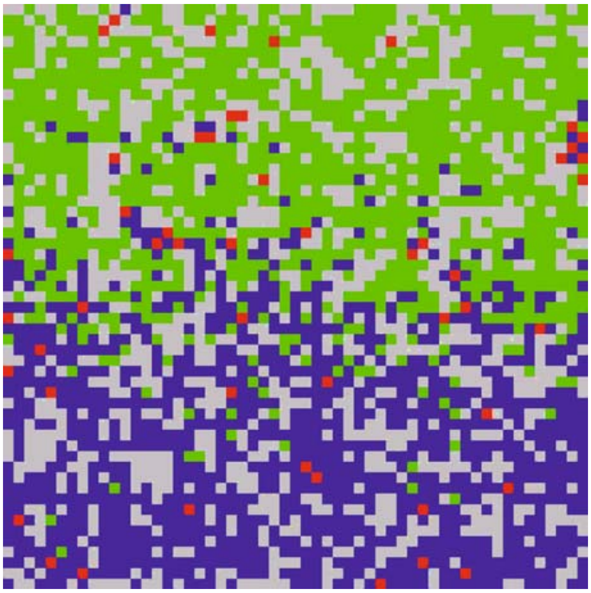
Diffusion in Water

Models of the diffusion process in water have been reported. One study revealed that hydrophobic solutes diffuse faster than hydrophilic ones, comparing molecules of similar size [19]. A model of diffusion as a function of water temperature produced the expected result of greater diffusion with higher temperature. A series of dilute solutions were modeled where the relative polarity of a solute S_1, was varied. A second solute, S_2, was introduced at the center of the grid. The dynamics showed that the solute, S_2, diffused faster when there is no co-solute and when the polarity of S_2 is low. The presence of the co-solute revealed that the diffusion of S_2 was fastest when the co-solvent, S_1 was hydrophobic.

In another study, the grid was divided into two halves, the upper half, containing a solution of a non-polar solute while the lower half contained a solution of a polar solute [20]. Between these two halves, a thin layer of cells



Cellular Automata Modeling of Complex Biochemical Systems, Figure 5
A model of the dissolution of a crystal in water



Cellular Automata Modeling of Complex Biochemical Systems, Figure 6
A model of the de-mixing of immiscible liquids

simulated a solution made up of a solute of intermediate polarity. The dynamics created a model in which the solute in the middle layer preferentially diffused into the upper layer containing the non-polar solute. A second study simulated an aqueous solution of intermediate polarity in the middle of the grid. Surrounding this solution were regions with highly polar, intermediately polar and non-polar solutes in solution. A fourth region around the center was pure water. The solute in the central section of the grid was allowed to freely diffuse. It diffused more rapidly into the pure water, however it secondarily preferred the more non-polar solution quadrant. The diffusion of a solute is modeled to be faster when it is hydrophobic and when co-solutes are also hydrophobic. The rate of diffusion of a solute in water was studied as a function of the water temperature and the hydropathic state of the solute.

Oil-Water Demixing

Models of the separation of two immiscible liquids and the partitioning of a solute between them have been reported [21]. The emerging configuration is an immiscible, two phase system. An interface formed with a greater concentration of water in the lower half while the second liquid dominated the upper half. The interface was unorganized with large "fingers" of cells from each half projecting into the other half of the grid. Figure 6 shows the interface in this model. This interface structure and behavior has been observed using other modeling paradigms. The simulated events of the demixing process present an intriguing model of a phenomenon that might be difficult to examine experimentally.

Another study tested the ability of this model to simulate the partitioning of solute molecules between the two phases, governed by their relative hydrophobicity. The addition of a small number of solute molecules was made to the initial, random mixture. As the dynamics proceeded, it was observed that the solute was associating with the patches of solvent to which it had the closest parameter-governed affinity. As a relatively stable configuration developed, the ratio of the solute molecules among the two phases became relatively constant. This ratio is the partition coefficient of the solute between the two phases. The dominant rules influencing the partition coefficient were the $P_B(WS_2)$ and $P_B(S_1S_2)$, the affinity of the solute for the two liquids, where S_2 is the solute molecule.

Observing the course of the dynamics there is a constantly changing pattern from the random configuration at the outset to the eventual formation of a disturbed interface and separated compartments of the two solvents. The solute molecules move rapidly to the patches in which the rules have ordained an affinity. The solute molecules have essentially partitioned themselves among the patches long before the two phases and the interface have formed.

Chemical Kinetics

Acid Dissociation The dissociation of an acid and the influence of the environment on this process was the subject of a recent study [22]. A cell modeling an organic acid molecule was divided into two parts, one face representing the carboxyl group, Y, and the other three faces, X, representing the non-dissociating, non-polar parts of the acid molecule. The strength of the acid, i. e. its propensity to dissociate, was governed by the probability rule, P_D. The hydronium ion was endowed with greater mobility since it is known to move very rapidly from one oxygen to another within the hydrogen bonded system of water. This was accomplished by allowing any of the four possible neighboring water molecules, relative to H to exchange positions with the hydronium cell, H.

An initial test of the model was to vary the P_D value and monitor the concentration of products. As expected, an increase in the P_D rule produced an increase in the calculated acid dissociation constant, K_a. A second study examined the influence of acid concentration on the observed properties. As expected, the K_a was approximately constant over a modest concentration range. A study on solution environment influences modeled the presence of another molecule in the solution. This co-solute was endowed with an attribute of non-dissociation. It's hydrophobicity was varied in a series of studies. The dissociation of the acid decreased when the hydrophobicity of the co-solute decreased. The interaction of two acids of different strength was also simulated using the same basic model. The observed dissociations revealed a strong and unequal influence of the two acids on each other. Both acids exhibit a suppression in their dissociations relative to their behavior in pure solution. The weaker acid is significantly more suppressed than the stronger acid. The decrease in dissociation of the two acids in a mixture cannot be readily calculated from the acid concentrations and their individual dissociation constants because of the complicating influences of ionic solvation effects on the water structure plus temperature factors.

First-Order Kinetics Many important natural processes ranging from nuclear decay to unimolecular chemical reactions, are first-order, or can be approximated as first-order [23]. This means that these processes depend only on the concentration to the first power of the transforming species itself. A cellular automaton model for such a system takes on an especially simple form, since rules for the movements of the ingredients are unnecessary and only transition rules for the inter-converting species need to be specified. Recent work described such a general cellular automaton model for first-order kinetics and tested its ability to simulate a number of classic first-order phenomena.

The prototype first-order transition is radioactive decay A → B, in which the concentration [A] of a species A decreases according to the rule that each A ingredient has a probability $P_t(\mathrm{AB})$ per unit time (here, per iteration) of converting to some other form B. For small numbers of A ingredients the actual decay curve observed for [A] is rather jagged and only roughly exponential, as a result of the irregular decays expected in this very finite, stochastic system. However, as the number of decaying ingredients is increased the decay curve approaches the smooth exponential fall-off expected for a deterministic system obeying the rate equation

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -k[A] \,. \tag{3}$$

When a reverse transition probability $P_t(\mathrm{BA})$ for the transition B → A is included the model simulates the first-order equilibrium:

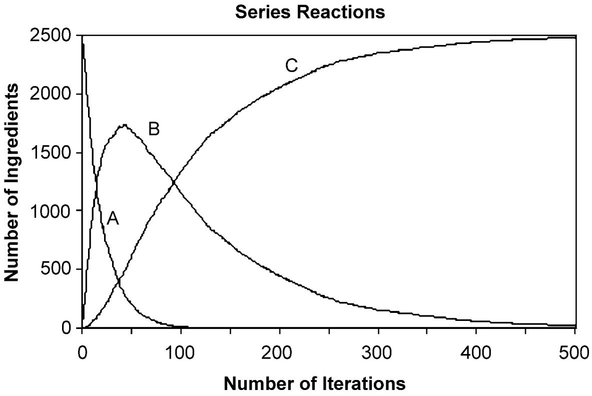
$$\mathrm{A} \leftrightarrow \mathrm{B} \,.$$

Here too, the finite size of the system causes notable fluctuations, in this case in the value of the equilibrium constant K, which fluctuates with time about the deterministic value

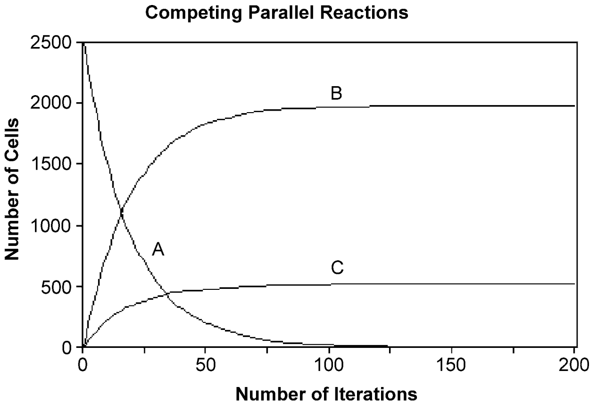
$$\mathrm{K} = \frac{P_t(\mathrm{AB})}{P_t(\mathrm{BA})} \,. \tag{4}$$

As an example, for 10 trials with 400 ingredients taking $P_t(\mathrm{AB}) = 0.05$ and $P_t(\mathrm{BA}) = 0.04$, we found K= 1.27 with a standard deviation of 0.13, compared to the deterministic value of 1.25. As a further test of the model one can ask whether it is ergodic in the sense that the average of K over time, i. e., for a single system observed for a long time after reaching equilibrium, is equal to the average K for identical systems taken at a particular time in a large number of trials. When this was tested for 1000 time steps (separated by 100 iterations) vs. 1000 trials the results were statistically identical, indicating that the first-order cellular automaton model is sensibly ergodic [2].

The first-order model can also be used to examine sequences of transformations of the form A → B → C . . .. For the simple example A → B → C the concentration of the initial reactant A falls exponentially, that of the intermediate species B rises then falls, and that of C builds up as it is fed from B. These time-dependent changes are illustrated in Fig. 7 using specific transformation probabilities.



Cellular Automata Modeling of Complex Biochemical Systems, Figure 7
Typical time-dependent variations of the ingredient populations for a two-step reaction A → B → C



Cellular Automata Modeling of Complex Biochemical Systems, Figure 8
Typical variations in the ingredient populations for competing reaction A → B → C

Kinetic and Thermodynamic Reaction Control Parallel competing reactions

$$\begin{array}{l} A \rightarrow B \\ \downarrow \\ C \end{array}$$

can also be simulated. An especially interesting example occurs when the reactions are reversible [24]:

$$A \leftrightarrow B \quad \text{and} \quad A \leftrightarrow C\,.$$

With properly chosen transformation probabilities, this model can be utilized to examine the conditions governing thermodynamic and kinetic control of reactions. An illustration is shown in Fig. 8 for the conditions $P_t(AB) = 0.01$, $P_t(AC) = 0.001$, $P_t(BA) = 0.02$, and $P_t(CA) = 0.0005$, using a set of 10 000 ingredients. Starting with all species A, in the initial stages of the reaction the kinetically-favored product B is produced in excess, whereas at later times the thermodynamically-favored product C gains dominance. The cellular automata model is in good agreement with those found in a deterministic, numerical solution for the same conditions. For example, the cellular automata model yields final equilibrium concentrations for species B and C of $[B] = 0.1439 \pm 0.0038$ and $[C] = 0.5695 \pm 0.0048$ compared to reported deterministic values of 0.14 and 0.571, respectively.

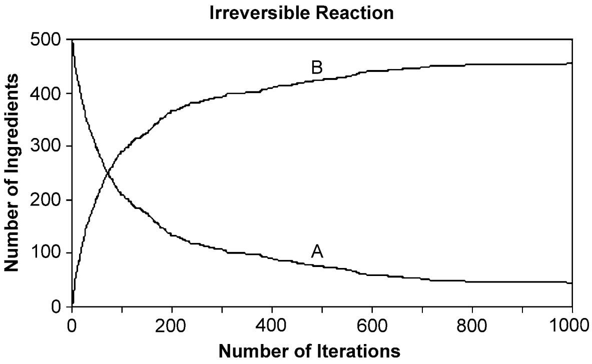
Excited-State Kinetics Another important application of the first-order model is to the examination of the ground and excited state kinetics of atoms and molecules [25]. One illustration of the excited-state cellular automata model is the dynamics of the excited-state transitions of oxygen atoms [26]. The oxygen atom has a ^3P ground state and ^1S and ^1D excited states. Emissions from the latter two excited states play an important role in the dramatic light displays—the *Aurora Borealis*, or "Northern lights"—seen under certain conditions in the northern polar skies, and similar emissions have been detected in the atmospheres of Mars and Venus. The excited states are believed to be mainly produced by dissociative recombinations of ionized oxygen molecules and electrons generated in the atmosphere by ultraviolet bombardment during the daylight hours:

$$O_2^+ + e^- \rightarrow O^* + O^{**}\,.$$

In this, the species O^* and O^{**} are unspecified atomic oxygen states, which may be any of the species ^3P, ^1S, or ^1D. The most prominent feature in the atmospheric displays is normally the green spin-allowed $^1S \rightarrow {}^1D$ transition appearing at 5577 angstroms.

Using transition probabilities taken from the compilation of Okabe [27] we have simulated the dynamics associated with these atomic transitions under both pulse and steady-state conditions. For the pulse simulations two starting conditions were examined: the first in which all ingredients started in the upper ^1S excited state, and the second in which the ingredients started in a distribution believed characteristic of that produced by the dissociative recombination process shown above. The simulations yield excited state lifetimes and luminescence quantum yields consistent with the experimental observations for these properties.

Second-Order Kinetics Several groups have developed cellular automata models for particular reaction-diffusion



Cellular Automata Modeling of Complex Biochemical Systems, Figure 9
Illustration of the variation with time of the ingredient populations for an irreversible reaction A → B

systems. In particular, the Belousov–Zhabotinsky oscillating reaction has been examined in a number of studies. Attention has also been directed to the simpler $A + B \rightarrow C$ reaction, using both lattice-gas models and a generalized Margolus diffusion approach. We have recently developed a simple, direct cellular automaton model [28] for hard-sphere bimolecular chemical reactions of the form

$$A + B \rightarrow C + D\,.$$

As before, the different species are assigned different colors in the visualization. In this model the reactant and product species diffuse about the grid in random walks. When the species A and B encounter each other (come to adjacent cells) on the grid the probability that these species transform to C and D is determined by an assigned reaction probability $P_r(AB)$. The simulations take place on a toroidal space such that ingredients leaving the grid on one side appear at the opposite edge. Initially the ingredients are placed randomly on the grid.

The production of species C over time for starting counts of 100 A and 200 B ingredients on a $100 \times 100 = 10\,000$ cell grid provides an example shown in Fig. 9. The expected second-order rate law

$$\frac{d[C]}{dt} = k[A][B] \tag{5}$$

is found to be obeyed in the simulations, subject to fluctuations expected for a system containing a finite number of reacting ingredients. When the results from a number of trials are combined to mimic, in effect, the results for a much larger system, the fluctuations become relatively small and the results approach the deterministic forms.

When a back reaction $C + D \rightarrow A + B$ with probability $P_r(CD)$ is included in the automaton rules the equilibrium

$$A + B \equiv C + D$$

can be simulated. Once the system has stabilized from its initial non-equilibrium concentrations, fluctuations about the equilibrium concentrations occur over time, and the relative size of these fluctuations decreases as the number of ingredients increases. A variation of this theme occurs in a pseudo first-order reaction. This type of reaction involves an actual second-order reaction $A + B \rightarrow C + D$ in which one of the reactants, say B, is present in sufficient excess that its variation becomes effectively unnoticeable. Simulations with 50 A ingredients and 1000 B ingredients bear out the expectation that the reaction kinetics are such that the rate of production of C and D appears to depend only on the concentration of A.

Enzyme Reactions

A recent study on the kinetics of an enzyme reaction considered the Michaelis–Menten model [29]. The rules selected for this reaction included a probability of the conversion, P_c, of an enzyme-substrate pair, ES, to an enzyme-product pair, EP. The Michaelis–Menten model was observed and characteristic Lineweaver–Burk plots were found from the model. The systematic variation of the hydrophobicity of substrates and products showed that a lower affinity between a substrate and water leads to a greater extent of the reaction at a common point along the reaction progress curve. This influence is greater than the influence of the affinity between the substrate and the enzyme. The water-substrate affinity appears to primarily influence the concentration of the ES complex at the observed point along the reaction progress curve. A low affinity between water and substrate favors a high ES concentration at this point. A hydrophilic substrate appears to be more entrapped in the water continuum, hence to be less available to the enzyme. It was also observed that an accumulation of product molecules around the enzymes coincides with a decline in the reaction rate. A biochemical network was reported to be modeled using cellular automata [30]. A protein kinase was modeled and studied using various concentrations of substrates and changes in enzyme competence.

An Anticipatory Model

An anticipatory enzyme system has been modeled using the dynamic characteristics of cellular automata [31]. A concentration of an intermediate product influences the competence of an enzyme down stream. This anticipation of the future event creates a condition in which the concentration of a later substrate is suppressed, a property characteristic of the system. The dynamics revealed concentrations over time, influenced by the presence or absence of a feed-forward or pre-adaptation state in the system. The concentration of A steadily diminishes as successive concentrations of B, C and D rise and fall at the same levels. The concentration of E rises at the end of the run, eventually becoming the only ingredient in the system. The concentration of D is approximately 0.25 in a non-anticipatory model. In contrast, with an anticipatory or feed-forward step in the system there is created an additional amount of enzyme specific for substrate D. This enzyme, $e_{2,4}$, is available at a future time to catalyze the conversion of D to E. This creates a property of the system in which the concentration of ingredient, D, is not allowed to accumulate to its normal level. The concentration of D in an anticipatory model is approximately 0.13, about one half of the D concentrations for the non-anticipatory models. The concentration of B therefore serves as a predictor of the concentration of D at a later time.

Micelle Formation

A micelle is a structure formed from the close interaction of hydrophobic fragments of amphiphiles plus the electrostatic encounters with the surrounding water. Typically they often assume a spherical structure with the non-polar fragments in the interior and the polar fragments on the periphery, in aqueous solution. The formation of these structures is a dynamic process which has been modeled using cellular automata. The model of an amphiphile was created by treating each face of a square automaton cell as an independent structure [32]. Each face of this variegated cell can have its own set of $P_{\mathrm{B}}(\mathrm{X})$ and $J(\mathrm{X})$ values. For the micelle study three of the faces were considered as equivalent and were endowed with rules modeling a hydrophobic or non-polar part of the amphiphile. The other face was treated as a polar fragment of the molecule and assigned characteristic rules. The outcome of the dynamics was the creation of structures in which the non-polar fragments were in the interior of an aggregation of cells while the polar fragment lay on the periphery. The interpretation of these organized clusters is that they model a micelle. The dominant influence on the formation of these structures is the extent of non-polar character of the designated three sides of the cell. Of secondary influence is the polarity of the remaining face of the cell. If this is too polar, the micelle formation is retarded. Both of these influences produce models that agree with experiment. Other studies on these variegated cells depicting an amphiphile revealed a temperature effect on the critical micelle concentration (cmc) which was minimal at about $P_{\mathrm{B}}(\mathrm{W}) = 0.25$, corresponding to experiment. The onset of the cmc was modeled and shown to be dependent upon a modestly polar fragment of the amphiphile.

Membrane Permeability

An extension of the micelle and diffusion models was a simulation of the diffusion of a solute through a layer of hydrophobic cells simulating a membrane separating two water compartments [33]. A membrane layer five cells wide membrane was positioned on a grid between two water compartments. The membrane cells were endowed with a $P_{\mathrm{B}}(\mathrm{WS})$ rule making them hydrophobic. The membrane cells could only move about within the layer according to their rule response. The two water cell compartments on either side were assigned identical rules but were colored differently in order to monitor their origins after some movement into and through the membrane layer. The dynamics revealed that water molecules from both compartments pass into and through the membrane as expected. To model the behavior of a solute in this environment, a few cells simulating solute molecules were positioned randomly near the lower edge of the membrane surface. These cells were endowed with rules making them hydrophilic. As the dynamics proceeded, it was observed that more water molecules from the upper compartment passed into and through the membrane than water from the lower compartment. Since the solute molecules were hydrophilic, the membrane was relatively impervious to their passage. The behavior of this model is in agreement with experimental observations collectively referred to as an osmotic effect. This model of diffusion was solute concentration dependent.

As the hydrophobicity of the solute increased, it was observed that an increasing number of solute particles passed through the membrane from the lower aqueous compartment. There was no accumulation of solute molecules within the membrane. At a level of hydrophobicity midway on the scale, i. e. about $P_{\mathrm{B}}(\mathrm{WS}) = 0.5$, there was a very abrupt change in this behavior. At this critical hydrophobicity the number of solute molecules passing through the membrane dropped sharply. At higher $P_{\mathrm{B}}(\mathrm{WS})$ values the number of cells passing through the membrane fell to nearly zero. At this critical hydrophobicity, the accumulation of solute molecules in the membrane increased sharply.

Chromatographic Separation

Models of chromatographic separation have been reported, derived from cellular automata [34]. The solvent cells were randomly distributed over the grid at the initiation of each run. The stationary phase, designated B is simulated by the presence of cells randomly distributed over the grid, replacing W cells. These B cells are immobile and are positioned at least 3 cells from another B cell. The solute cells, usually simulating two different compounds are represented by a few cells each. The movements of the solutes and mobile phase in the grid were governed by rules denoting the joining and breaking of like or unlike cells. The position of each solute cell was recorded at a row in the column after a certain number of iterations. The position of the peak maximum was determined by summing the number of cells found in groups of ten rows on the column then plotting these averages against the iteration time.

The gravity parameter for each ingredient in the simulation defines the flow rate. The polarity of the solvent, W, is encoded in the relative self-association experienced. This is governed by the rules, P_{B}(WW) and J(WW). The migration of the solutes was found to be faster when the solvent was non-polar. Another study modeled the influence of the relative affinities of solutes for the stationary phase, B. This affinity is encoded in the parameters, P_{B}(SB) and J(SB). High values of P_{B}(SB) and low values of J(SB) denote a weak affinity. A single solute was used in this study with five different sets of parameters. These studies revealed that solutes with a greater affinity for the stationary phase migrated at a slower rate. These parameters characterize the structural differences among solutes that give rise to different migratory rates and separations in chromatography.

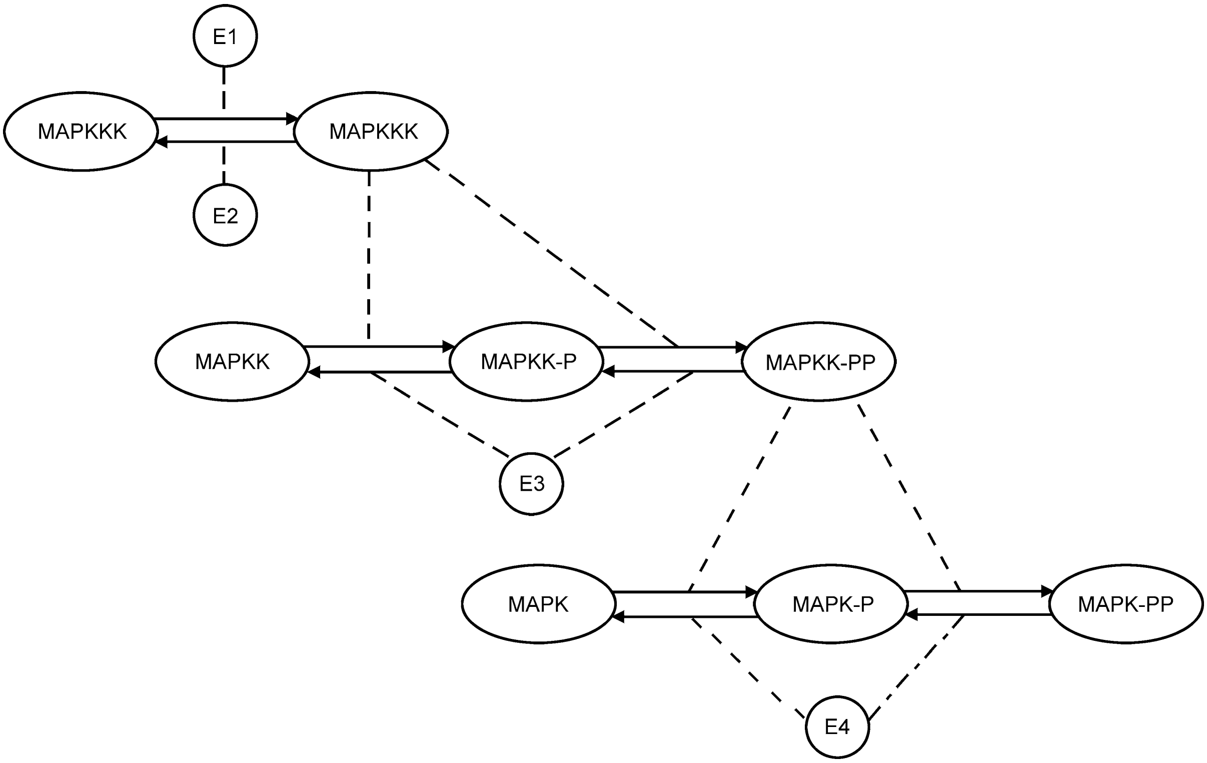
Modeling Biochemical Networks

The Network Dynamic evolutionary networks have recently been recognized as a universal approach to complex systems including biochemical systems. Network topology is generally used in characterizing these, focusing on their connectivity, neighborhood and distance relationships. Network complexity has also been recently quantitatively characterized [35]. The large size of the metabolic, protein, and gene regulatory networks makes impractical many of the traditional methods for dynamic modeling. A study on the MAPK signaling cascade has recently introduced the potential of cellular automata as a basic method for the dynamic modeling of networks for biological and medical applications [30].

The CA Modeling Design Each molecule involved in the MAPK pathway, Fig. 10, was represented by a number of cells in the CA grid. The numbers chosen reflect the relative concentration of that protein. Each of the cells representing all other proteins move about freely in the grid. They may encounter each other but this has no consequence. The only encounters that have a consequence are those between a specific protein (substrate) and a specific enzyme, as shown in the network. When such an encounter occurs, there is modeled a complex (enzyme-substrate). This complex has an assigned probability of converging to a new complex (enzyme-product). Following this there is a probability assigned for the separation of these two species.

The studies of the MAPK cascade were performed using a CA grid of 100 by 100 cells. Each model was obtained as the average of 50 runs, each of which included 5000 iterations. A network to be studied was represented by groups of CA cells, each group representing one of the network species. The number of cells in each group reflects the relative concentrations of each network ingredient. We have systematically altered the initial concentrations of several proteins (MAPKKK, MAPKK, and MAPK) and the competencies of several enzymes (MAPKK- and MAPK-proteases, and the hypothetical enzymes E1 and E2 that affect the forward and reverse reactions of activation and deactivation of MAPKKK). The basic variable was the concentration of MAPKKK, which was varied within a 25-fold range from 20 to 500 cells. The concentrations of MAPKK and MAPK were kept constant (500 or 250 cells) in most of the models. The four enzymes, denoted by E1, E2, E3, and E4, were represented in the CA grid by 50 cells each. In one series of models, we kept the transition probabilities of three of the enzymes the same, ($P = 0.1$), and varied the probability of the fourth enzyme within the 0 to 1 range. In another series, *all* enzyme probabilities were kept constant, whereas the concentrations of substrates were varied. The last series varied both substrate concentrations and enzyme propensities. Recorded were the variations in the concentrations of the three substrates MAPKKK, MAPKK, and MAPK, and the products MAPKKK*, MAPKK-P, MAPKK-PP, MAPK-P, and MAPK-PP.

Modeling Enzymes Network Activity Upgrading or downgrading enzymes activity is one of the typical ways the cell reacts to stress and interactions with pathogens. A systematic study was made of the variations of one of the four enzymes E1 to E4 at constant concentrations of the substrates MAPKKK, MAPKK and MAPK, and constant propensity of the other three enzymes. This is illustrated in Fig. 11 with the variation of the MAPKK-protease (E3 enzyme in Fig. 10), which reverses the two-step reaction of MAPKK phosphorilation. It is shown that the concentration of the MAPKK- and MAPK-diphos-



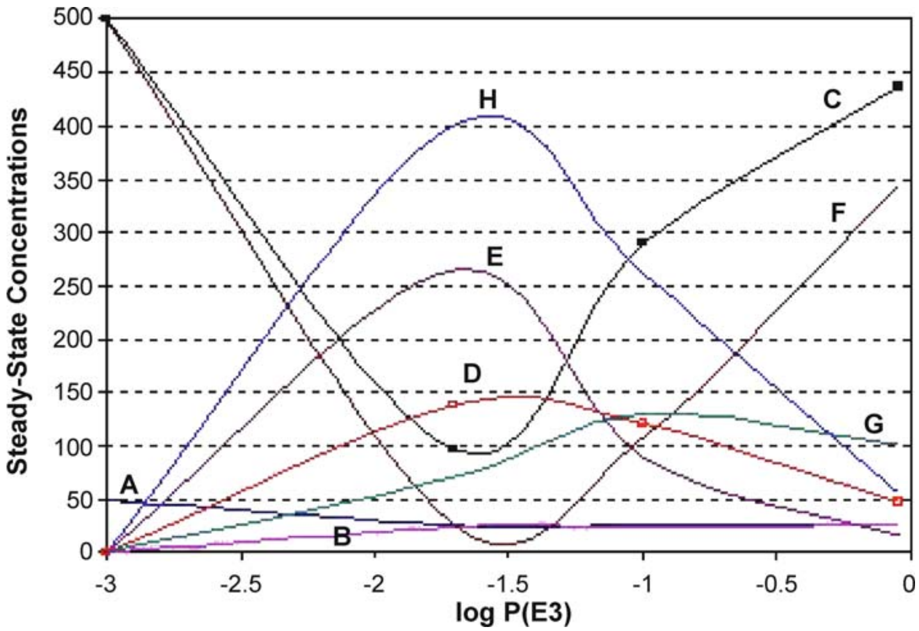
Cellular Automata Modeling of Complex Biochemical Systems, Figure 10
The MAPK signaling cascade. The substrates and products are represented by *oval contours*, reactions by *arrows*, and the catalysts action by *dashed lines*. E3 and E4 are MAPKK-protease and MAPK-protease, respectively. P stands for phosphate, PP for diphosphate

phates (marked as E and H respectively) passes through a maximum near relatively low enzyme transition probability ($P \approx 0.02$). At the point of its maximum, the concentration of MAPK-PP reaches over 80% of its maximum, whereas that of MAPKK-PP is slightly over 50%. This shows the potential for a strong influence on the concentrations of the two diphosphates in the MAPK cascade by inhibiting the MAPKK-protease. In contrast, the level of steady-state concentrations of the two monophosphates (marked by D and G in Fig. 11) is not sensitive to the activity of the enzyme modeled, except for the extreme case of very strong inhibition ($P \to 0.001$).

Future Directions

Applications of cellular automata as a modeling paradigm in chemistry is in its infancy. Literature searches show little attention to the use of this method to study chemical phenomena. Chemistry is about dynamic systems, thus a rich harvest of information is possible using this paradigm. This potential is illustrated here where primarily solution systems are modeled. A broader use is clearly on the horizon.

The study of bulk water structure is one example of this potential. It points to a way of encoding the structure of an evanescent substance with some degree of validity. A recent example of this potential is a series of studies of the influence of protein surface amino acid side chains on the structure of nearby water [36]. Because of the varying hydropathic states of the surface side chains, a series of passageways (chreodes) was postulated to exist through the water. These chreodes were invoked in the facilitated 2-D diffusion of ligands and substrates to receptors and enzyme active sites. The interference of these chreodes was proposed in a theory of the action of general anesthetics. As an extension of this theory, the induction of sleep was proposed to arise from the interference of these chreodes by elemental nitrogen accumulated through respiration over several hours. The remarkable correspondence of kinetic and enzyme CA models with reality is excellent and foretells the use of these in many applications. The treatment of networks, foretells of another area of great



Cellular Automata Modeling of Complex Biochemical Systems, Figure 11
Influence of the MAPKK-protease propensity *P*(E3) on the steady-state concentrations of the MAPK cascade species (A = MAPKKK, B = MAPKKK*, C = MAPKK, D = MAPKK-P, E = MAPKK-PP, F = MAPK, F = MAPK-P, H = MAPK-PP). Enzyme propensities *P*(E1) = *P*(E2) = *P*(E4) = 0.1, substrate initial concentrations $[A_o] = 50$, $[C_o] = [F_o] = 500$

possibilities of models with CA. The beginning of this approach is shown above with the MAPK mode. There is success here that is not achievable with classical differential equations.

A noteable advantage and a golden opportunity exists for any CA chemical model. That is the didactic value of any CA model. You can see the dissolution of a crystal, the changes of concentration in a CA enzyme model, the changes in structure as the temperature changes, and many more. A student will never forget a CA model of these while an equation is quietly forgotten.

New aspects of CA models are on the horizon. Multidimensional models, nested hierarchies, multi-grid assemblies, all are being explored and will soon surface as viable methods rich in information.

Bibliography

Primary Literature

1. Kier LB, Seybold PG, Cheng C-K (2005) Modeling chemical systems using cellular automata. Springer, Dordrecht
2. Von Neumann J (1966) Burks A (ed) Theory of self-reproducing automata. University of Illinois Press, Champaign
3. Ulam SM (1952) Proc Int Congr Math 2:264, held in 1950
4. Ulam SM (1976) Adventures of a mathematician. Charles Scribner's Sons, New York
5. Tofolli T, Margolas N (1987) Cellular automata machines: A new environment for modeling. MIT Press, Cambridge
6. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign
7. Kapral R, Showalter K (1995) Chemical waves and patterns. Kluwer, Boston
8. Ermentrout GB, Edelstein-Keshet L (1993) J Theoret Biol 160: 97–133
9. Grimm V, Revilla E, Berger U et al (2005) Science 310:987–991
10. Wu-Pong S, Cheng C-K (1999) Pharmacokinetic simulations using cellular automata in a pharmacokinetics course. Am J Pharm Educ 63:52–55
11. Moreira N (2006) In pixels and in health. Sci News Jan 21:40–44
12. Batty M (2005) Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals. MIT Press, Cambridge
13. Kohler TA, Gumerman GJ (2000) Dynamics in human and primate societies: Agent-based modelling of social and spatial processes. Oxford University Press, New York
14. White R (2005) Modelling multi-scale processes in a cellular automata framework. In: Portugali J (ed) Complex artificial environments. Springer, New York, pp 165–178
15. Kier LB, Cheng C-K (1994) A cellular automata model of water. J Chem Inf Comp Sci 34:647–654
16. Kier LB, Cheng C-K (1994) A cellular automata model of an aqueous solution. J Chem Inf Comp Sci 34:1334–1341
17. Kier LB, Cheng C-K, Testa B (1995) A cellular automata model of the hydrophobic effect. Pharm Res 12:615–622
18. Kier LB, Cheng C-K (1995) A cellular automata model of dissolution. Pharm Res 12:1521–1528

19. Kier LB, Cheng C-K, Testa B (1997) A cellular automata model of diffusion in aqueous systems. J Pharm Sci 86:774–781
20. Kier LB, Cheng C-K, Seybold PG (2001) A cellular automata model of aqueous systems. Revs Comput Chem 17:205–238
21. Cheng C-K, Kier LB (1995) A cellular automata model of oil-water partitioning. J Chem Inf Comput Sci 35:1054–1061
22. Kier LB, Cheng C-K, Tute M, Seybold PG (1998) A cellular automata model of acid dissociation. J Chem Inf Comp Sci 38:271–278
23. Seybold PG, Kier LB, Cheng C-K (1997) J Chem Inf Comput Sci 37:386–391
24. Neuforth A, Seybold PG, Kier LB, Cheng C-K (2000) Cellular automata models of kinetically and thermodynamically controlled reactions, vol A. Int J Chem Kinetic 32:529–534
25. Seybold PG, Kier LB, Cheng C-K (1998) Stochastic cellular automata models of molecular excited state dynamics. J Phys Chem A 102:886–891
26. Seybold PG, Kier LB, Cheng C-K (1999) Aurora Borealis: Stochastic cellular automata simulation of the excited state dynamics of Oxygen atoms. Int J Quantum Chem 75:751–756
27. Okabe H (1978) Photochemistry of small molecules. Wiley, New York, p 370
28. Moore J, Seybold PG; To be published personal correspondence
29. Kier LB, Cheng C-K, Testa B (1996) A cellular automata model of enzyme kinetics. J Molec Graph 14:227–234
30. Kier LB, Bonchev D, Buck G (2005) Modeling biochemical networks: A cellular automata approach. Chem Biodiv 2:233–43
31. Kier LB, Cheng C-K (2000) A cellular automata model of an anticipatory system. J Molec Graph Model 18:29–35
32. Kier LB, Cheng C-K, Testa B (1996) Cellular automata model of micelle formation. Pharm Res 13:1419–1426
33. Kier LB Cheng C-K (1997) A cellular automata model of membrane permeability. J Theor Biol 186:75–85
34. Kier LB, Cheng C-K, Karnes HT (2000) A cellular automata model of chromatography. Biomed Chrom 14:530–539
35. Bonchev D (2003) Complexity of protein–protein interaction networks, complexes and pathways. In: Conn M (ed) Handbook of proteomics methods. Humana, New York, pp 451–462
36. Kier LB (2007) Water as a complex system: Its role in ligand diffusion, general anesthesia, and sleep. Chem Biodiv 4:2473–2479

Selected Bibliography of Works on Cellular Automata

At this time thousands of scientific articles have been published describing cellular automata studies of topics ranging from applications dealing with physical and biological systems to investigations of traffic control and topics in the social sciences. It would be impossible to describe all of these studies within a limited space, but it may be useful to provide a short list of representative investigations on a limited variety of topics, permitting starting points for readers who might wish to further examine applications in these more narrow subjects. Below we give a short selection of publications, some of which, although not explicitly referring to CA, cover the same approach or a related approach.

Artificial Life

Adami C (1998) An introduction to artificial life. Springer, New York

Langton CG, Farmer JD, Rasmussen S, Taylor C (1992) Artificial life, vol II. Addison-Wesley, Reading

Biological Applications (General)

Maini PK, Deutsch A, Dormann S (2003) Cellular automaton modeling of biological pattern formation. Birkhäuser, Boston

Sigmund K (1993) Games of life: Explorations in ecology, evolution, and behaviour. Oxford University Press, New York

Solé R, Goodman B (2000) Signs of life: How complexity pervades biology. Basic Books, New York; A tour-de-force general introduction to biological complexity, with many examples

Books

Chopard B, Droz M (1998) Cellular automata modeling of physical systems. Cambridge University Press, Cambridge

Gaylord RJ, Nishidate K (1996) Modeling nature: Cellular automata simulations with Mathematica®. Telos, Santa Clara

Griffeath D, Moore C (2003) New constructions in cellular automata. In: Santa Fe Institute Studies in the Sciences of Complexity Proceedings. Oxford University Press, New York

Ilachinski A (2001) Cellular automata: A discrete universe. World Scientific, Singapore

Kier LB, Seybold PG, Cheng C-K (2005) Modeling chemical systems using cellular automata. Springer, Dodrecht

Manneville P, Boccara N, Vishniac GY, Bidaux R (1990) Cellular automata and modeling of complex physical systems. Springer, New York, pp 57–70

Schroeder M (1991) Fractals, chaos, power laws. WH Freeman, New York

Toffoli T, Margolus N (1987) Cellular automata machines: A new environment for modeling. MIT Press, Cambridge

Wolfram S (1994) Cellular automata and complexity: Collected papers. Westview Press, Boulder

Wolfram S (2002) A new kind of science. Wolfram Media, Champaign

Emergent Properties

Gruner D, Kapral R, Lawniczak AT (1993) Nucleation, domain growth, and fluctuations in a bistable chemical system. J Chem Phys 96:2762–2776

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Nat Acad Sci USA 79:2554–2558

Kauffman S (1984) Emergent properties in random complex automata. Physica D 10:145–156

Ottino JM (2004) Engineering complex systems. Nature 427:399

Evolution

Farmer JD, Kauffman SA, Packard NH (1986) Autocatalytic replication of polymers. Physica D 22:50–67

Solé RV, Bascompté J, Manrubia SC (1996) Extinctions: Bad genes or weak chaos? Proc Roy Soc Lond B 263:1407–1413

Solé RV, Manrubia SC (1997) Criticality and unpredictability in macroevolution. Phys Rev E 55:4500–4508

Solé RV, Manrubia SC, Benton M, Bak P (1997) Self-similarity of extinction statistics in the fossil record. Nature 388:764–767

Solé RV, Montoya JM, Erwin DH (2002) Recovery from mass extinction: Evolutionary assembly in large-scale biosphere dynamics. Phil Trans Royal Soc 357:697–707

Excited State Phenomena

Seybold PG, Kier LB, Cheng C-K (1998) Stochastic cellular automata models of molecular excited-state dynamics. J Phys Chem A 102:886–891; Describes general cellular automata models of molecular excited states

Seybold PG, Kier LB, Cheng C-K (1999) Aurora Borealis: Stochastic cellular automata simulations of the excited-state dynamics of Oxygen atoms. Int J Quantum Chem 75:751–756; This paper examines the emissions and excited-state transitions of atomic Oxygen responsible for some of the displays of the Aurora Borealis

First-Order Chemical Kinetics

Hollingsworth CA, Seybold PG, Kier LB, Cheng C-K (2004) First-order stochastic cellular automata simulations of the Lindemann mechanism. Int J Chem Kinetic 36:230–237

Neuforth A, Seybold PG, Kier LB, Cheng C-K (2000) Cellular automata models of kinetically and thermodynamically controlled reactions. Int J Chem Kinetic 32:529–534; A study of kinetic and thermodynamic reaction control

Seybold PG, Kier LB, Cheng C-K (1997) Simulation of first-order chemical kinetics using cellular automata. J Chem Inf Comput Sci 37:386–391; This paper illustrates a number of first-order cellular automata models

Fluid Flow

Malevanets A, Kapral R (1998) Continuous-velocity lattice-gas model for fluid flow. Europhys Lett 44:552

Game of Life

Alpert M (1999) Not just fun and games. Sci Am 40:42; A profile of John Horton Conway

Gardner M (1970) The fantastic combinations of John Conway's new solitaire game "life". Sci Amer 223:120–123

Gardner M (1971) On cellular automata, self-reproduction, the Garden of Eden and the game of "life". Sci Amer 224:112–117"

Note: There are many examples on the web of applets that allow you to play the Game of Life. Since these come and go, you are urged to locate them using a search engine.

Geology

Barton CC, La Pointe PR (1995) Fractals in petroleum geology and earth processes. Plenum, New York

Turcotte DL (1997) Fractals and chaos in geology and geophysics, 2nd edn. Cambridge University Press, New York

Historical Notes

Ulam SM (1952) Random processes and transformations. Proc Int Congr Math 2:264, held in 1950

Ulam SM (1976) Adventures of a mathematician. Charles Scribner's Sons, New York

Von Neumann J (1966) Burks A (ed) Theory of self-replicating automata. University of Illinois Press, Urbana

Zuse K (1982) The computing universe. Int J Theoret Phys 21:589

Liquid Phase Interactions

Cheng C-K, Kier LB (1995) A cellular automata model of oil-water partitioning. J Chem Info Comput Sci 35:1054–1059

Kier LB, Cheng C-K, Testa B (1996) A cellular automata model of micelle formation. Pharm Res 13:1419–1422

Malevanets A, Kapral R (1999) Mesoscopic model for solvent dynamics. J Chem Phys 110:8605–8613

Oscillations

Chavez F, Kapral R (2002) Oscillatory and chaotic dynamics in compartmentalized geometries. Phys Rev E 65:056203

Chavez F, Kapral R, Rousseau G, Glass L (2001) Scroll waves in spherical shell geometries. Chaos 11:757

Goryachev A, Strizhak P, Kapral R (1997) Slow manifold structure and the emergence of mixed-mode oscillations. J Chem Phys 107:2881

Hemming C, Kapral R (2002) Phase front dynamics in inhomogeneously forced oscillatory systems. Physica A 306:199

Pattern Formation

Kapral R, Showalter K (1994) Chemical waves and patterns. Kluwer, Dordrecht

Parrish JK, Edelstein-Keshet L (1999) Complexity, pattern, and evolutionary trade-offs in animal aggregation. Science 284:99–101

Veroney JP, Lawniczak AT, Kapral R (1996) Pattern formation in heterogeneous media. Physica D 99:303–317

Physics Applications

Signorini J (1990) Complex computing with cellular automata. In: Manneville P, Boccara N, Vishniac GY, Bidaux R (eds) Cellular automata and modeling of complex physical systems. Springer, New York, pp 57–70

Toffoli T (1984) Cellular automata as an alternative (rather than an approximation of) differential equations in modeling physics. Physica D 10:117–127

Vichniac GY (1984) Simulating physics with cellular automata. Physica D 10:96–116

Population Biology and Ecology

Bascompté J, Solé RV (1994) Spatially-induced bifurcations in single species population dynamics. J Anim Ecol 63:256–264

Bascompté J, Solé RV (1995) Rethinking complexity: Modelling spatiotemporal dynamics in ecology. Trends Ecol Evol 10:361–366

Bascompté J, Solé RV (1996) Habitat fragmentation, extinction thresholds in spatialy explicit models. J Anim Ecol 65:465

Deutsch A, Lawniczak AT (1999) Probabilistic lattice models of collective motion, aggregation: From individual to collective dynamics. J Math Biosci 156:255–269

Fuks H, Lawniczak AT (2001) Individual-based lattice models for the spatial spread of epidemics. Discret Dyn Nat Soc 6(3):1–18

Gamarra JGP, Solé RV (2000) Bifurcations, chaos in ecology: Lynx returns revisited. Ecol Lett 3:114–121

Levin SA, Grenfell B, Hastings A, Perelson AS (1997) Mathematical, computational challenges in population biology, ecosystems science. Science 275:334–343

Montoya JM, Solé RV (2002) Small world patterns in food webs. J Theor Biol 214:405–412

Nowak MA, Sigmund K (2004) Population dynamics in evolutionary ecology. In: Keinan E, Schechter I, Sela M (eds) Life sciences for the 21st century. Wiley-VCH, Cambridge, pp 327–334

Solé RV, Alonso D, McKane A (2000) Connectivity, scaling in S-species model ecosystems. Physica A 286:337–344

Solé RV, Manrubia SC, Kauffman S, Benton M, Bak P (1999) Criticality, scaling in evolutionary ecology. Trends Ecol Evol 14:156–160

Solé RV, Montoya JM (2001) Complexity, fragility in ecological networks. Proc Roy Soc 268:2039–2045

Random Walks

Berg HC (1983) Random walks in biology. Princeton University Press, Princeton

Hayes B (1988) How to avoid yourself. Am Sci 86:314–319

Lavenda BH (1985) Brownian motion. Sci Am 252(2):70–85

Shlesinger MF, Klafter J (1989) Random walks in liquids. J Phys Chem 93:7023–7026

Slade G (1996) Random walks. Am Sci 84:146–153

Weiss GH (1983) Random walks, their applications. Am Sci 71:65–71

Reviews

Kapral R, Fraser SJ (2001) Chaos, complexity in chemical systems. In: Moore JH, Spencer ND (eds) Encyclopedia of chemical physics, physical chemistry, vol III. Institute of Physics Publishing, Philadelphia, p 2737

Kier LB, Cheng C-K, Testa (1999) Cellular automata models of biochemical phenomena. Future Generation Compute Sci 16:273–289

Kier LB, Cheng C-K, Seybold PG (2000) Cellular automata models of chemical systems. SAR QSAR Environ Res 11:79–102

Kier LB, Cheng C-K, Seybold PG (2001) Cellular automata models of aqueous solution systems. In: Lipkowitz KM, Boyd DB (eds) Reviews in computational chemistry, vol 17. Wiley-VCH, New York, pp 205–225

Turcotte DL (1999) Self-organized criticality. Rep Prog Phys 62:1377–1429

Wolfram S (1983) Cellular automata. Los Alamos Sci 9:2–21

Second-Order Chemical Kinetics

Boon JP, Dab D, Kapral R, Lawniczak AT (1996) Lattice-gas automata for reactive systems. Phys Rep 273:55–148

Chen S, Dawson SP, Doolen G, Jenecky D, Lawiczak AT (1995) Lattice methods for chemically reacting systems. Comput Chem Engr 19:617–646

Self-Organized Criticality

Bak P (1996) How nature works. Springer, New York

Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation for 1/f noise. Phys Rev Lett 59:381–384; A classic paper introducing the "sandpile" cellular automaton

Turcotte DL (1999) Self-organized criticality. Rep Prog Phys 62:1377–1429

Social Insect Behavior

Cole BJ (1991) Short-term activity cycles in ants: Generation of periodicity by worker inaction. Amer Nat 137:144–259

Cole BJ (1996) Mobile cellular automata models of ant behavior: Movement activity of Leptothorax Allardycei. Amer Nat 148:1–15

Deneubourg J-L, Goss S, Franks NR, Pasteels JM (1989) The blind leading the blind: Modeling chemically mediated Army ant raid patterns. J Insect Behav 2:719–772

Goss S, Deneubourg J-L (1988) Autocatalysis as a source of synchronized rhythmical activity in social insects. Insectes Sociaux 35:310–315

Solé RV, Miramontes O, Goodwin BC (1993) Oscillations, chaos in ant societies. J Theoret Biol (1993) 161:343–357

Social Sciences

Gaylord RJ (1998) D'Andria LJ (1998) Simulating society: A Mathematica toolkit for modeling socioeconomic behavior. Springer/Telos, New York

Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco

Traffic Rules

Huang P-H, Kong L-J, Liu M-R (2002) A study of a main-road cellular automata traffic flow model. Chin Phys 11:678–683

Nagel K, Wolf DE, Wagner P, Simon P (1998) Two-lane rules for cellular automata: A systematic approach. Phys Rev E 58:1425–1437

Water

Kier LB, Cheng C-K (1994) A cellular automata model of water. J Chem Info Comput Sci 34:647

Cellular Automata Modeling of Physical Systems

BASTIEN CHOPARD
Computer Science Department, University of Geneva,
Geneva, Switzerland

Article Outline

Cellular automata offer a powerful modeling framework to describe and study physical systems composed of interacting components. The potential of this approach is demonstrated in the case of applications taken from various fields of physics, such as reaction-diffusion systems, pattern formation phenomena, fluid flows and road traffic models.

Glossary

BGK models Lattice Boltzmann models where the collision term Ω is expressed as a deviation from a given local equilibrium distribution $f^{(0)}$, namely $\Omega = (f^{(0)} - f)/\tau$, where f is the unknown particle distribution and τ a relaxation time (which is a parameter of the model). BGK stands for Bhatnager, Gross and Krook who first considered such a collision term, but not specifically in the context of lattice systems.

CA Abbreviation for cellular automata or cellular automaton.

Cell The elementary spatial component of a CA. The cell is characterized by a state whose value evolves in time according to the CA rule.

Cellular automaton System composed of adjacent cells or sites (usually organized as a regular lattice) which evolves in discrete time steps. Each cell is characterized by an internal state whose value belongs to a finite set. The updating of these states is made in parallel according to a local rule involving only a neighborhood of each cell.

Conservation law A property of a physical system in which some quantity (such as mass, momentum or energy) is locally conserved during the time evolution. These conservation laws should be included in the microdynamics of a CA model because they are essential ingredients governing the macroscopic behavior of any physical system.

Collision The process by which the particles of a LGA change their direction of motion.

Continuity equation An equation of the form $\partial_t \rho + \operatorname{div} \rho \mathbf{u} = 0$ expressing the mass (or particle number) conservation law. The quantity ρ is the local density of particles and $\mathbf{u}$ the local velocity field.

Critical phenomena The phenomena which occur in the vicinity of a continuous phase transition, and are characterized by very long correlation length.

Diffusion A physical process described by the equation $\partial_t \rho = D\nabla^2 \rho$, where ρ is the density of a diffusing substance. Microscopically, diffusion can be viewed as a random motion of particles.

DLA Abbreviation of Diffusion Limited Aggregation. Model of a physical growth process in which diffusing particles stick on an existing cluster when they hit it. Initially, the cluster is reduced to a single seed particle and grows as more and more particles arrive. A DLA cluster is a fractal object whose dimension is typically 1.72 if the experiment is conducted in a two-dimensional space.

Dynamical system A system of equations (differential equations or discretized equations) modeling the dynamical behavior of a physical system.

Equilibrium states States characterizing a closed system or a system in thermal equilibrium with a heat bath.

Ergodicity Property of a system or process for which the time-averages of the observables converge, in a probabilistic sense, to their ensemble averages.

Exclusion principle A restriction which is imposed on LGA or CA models to limit the number of particles per site and/or lattice directions. This ensures that the dynamics can be described with a cellular automata rule with a given maximum number of bits. The consequence of this exclusion principle is that the equilibrium distribution of the particle numbers follows a Fermi–Dirac-like distribution in LGA dynamics.

FHP model Abbreviation for the Frisch, Hasslacher and Pomeau lattice gas model which was the first serious candidate to simulate two-dimensional hydrodynamics on a hexagonal lattice.

Fractal Mathematical object usually having a geometrical representation and whose spatial dimension is not an integer. The relation between the size of the object and its "mass" does not obey that of usual geometrical objects. A DLA cluster is an example of a fractal.

Front The region where some physical process occurs. Usually the front includes the locations in space that are first affected by the phenomena. For instance, in a reaction process between two spatially separated reactants, the front describes the region where the reaction takes place.

HPP model Abbreviation for the Hardy, de Pazzis and Pomeau model. The first two-dimensional LGA aimed at modeling the behavior of particles colliding on a square lattice with mass and momentum conservation. The HPP model has several physical drawbacks that have been overcome with the FHP model.

Invariant A quantity which is conserved during the evolution of a dynamical system. Some invariants are imposed by the physical laws (mass, momentum, energy) and others result from the model used to describe physical situations (spurious, staggered invariants). Collisional invariants are constant vectors in the space where the Chapman–Enskog expansion is performed, associated to each quantity conserved by the collision term.

Ising model Hamiltonian model describing the ferromagnetic paramagnetic transition. Each local classical spin variables $s_i = \pm 1$ interacts with its neighbors.

Isotropy The property of continuous systems to be invariant under any rotations of the spatial coordinate system. Physical quantities defined on a lattice and obtained by an averaging procedure may or may not be isotropic, in the continuous limit. It depends on the type of lattice and the nature of the quantity. Second-order tensors are isotropic on a 2D square lattice but fourth-order tensors need a hexagonal lattice.

Lattice The set of cells (or sites) making up the spatial area covered by a CA.

Lattice Boltzmann model A physical model defined on a lattice where the variables associated to each site represent an average number of particles or the probability of the presence of a particle with a given velocity. Lattice Boltzmann models can be derived from cellular automata dynamics by an averaging and factorization procedure, or be defined per se, independently of a specific realization.

Lattice gas A system defined on a lattice where particles are present and follow given dynamics. Lattice gas automata (LGA) are a particular class of such a system where the dynamics are performed in parallel over all the sites and can be decomposed in two stages: (i) propagation: the particles jump to a nearest-neighbor site, according to their direction of motion and (ii) collision: the particles entering the same site at the same iteration interact so as to produce a new particle distribution. HPP and FHP are well-known LGA.

Lattice spacing The separation between two adjacent sites of a regular lattice. Throughout this book, it is denoted by the symbol Δ_r.

LB Abbreviation for Lattice Boltzmann.

LGA Abbreviation for Lattice Gas Automaton. See lattice gas model for a definition.

Local equilibrium Situation in which a large system can be decomposed into subsystems, very small on a macroscopic scale but large on a microscopic scale such that each sub-system can be assumed to be in thermal equilibrium. The local equilibrium distribution is the function which makes the collision term of a Boltzmann equation vanish.

Lookup table A table in which all possible outcomes of a cellular automata rule are pre-computed. The use of a lookup table yields a fast implementation of a cellular automata dynamics since however complicated a rule is, the evolution of any configuration of a site and its neighbors is directly obtained through a memory access. The size of a lookup table grows exponentially with the number of bits involved in the rule.

Margolus neighborhood A neighborhood made of two-by-two blocks of cells, typically in a two-dimensional square lattice. Each cell is updated according to the values of the other cells in the same block. A different rule may possibly be assigned dependent on whether the cell is at the upper left, upper right, lower left or lower right location. After each iteration, the lattice partition defining the Margolus blocs is shifted one cell right and one cell down so that at every other step, information can be exchanged across the lattice. Can be generalized to higher dimensions.

Microdynamics The Boolean equation governing the time evolution of a LGA model or a cellular automata system.

Moore neighborhood A neighborhood composed of the central cell and all eight nearest and next-nearest neighbors in a two-dimensional square lattice. Can be generalized to higher dimensions.

Multiparticle models Discrete dynamics modeling a physical system in which an arbitrary number of particles is allowed at each site. This is an extension of an LGA where no exclusion principle is imposed.

Navier–Stokes equation The equation describing the velocity field **u** in a fluid flow. For an incompressible fluid ($\partial_t \rho = 0$), it reads

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \nu \nabla^2 \mathbf{u}$$

where ρ is the density and P the pressure. The Navier–Stokes equation expresses the local momentum conservation in the fluid and, as opposed to the Euler equation, includes the dissipative effects with a viscosity term $\nu \nabla^2 \mathbf{u}$. Together with the continuity equation, this is the fundamental equation of fluid dynamics.

Neighborhood The set of all cells necessary to compute a cellular automaton rule. A neighborhood is usually composed of several adjacent cells organized in a simple geometrical structure. Moore, von Neumann and Margolus neighborhoods are typical examples.

Occupation numbers Boolean quantities indicating the presence or absence of a particle in a given physical state.

Open system A system communicating with the environment by exchange of energy or matter.

Parallel Refers to an action which is performed simultaneously at several places. A parallel updating rule corresponds to the updating of all cells at the same time as if resulting from the computations of several independent processors.

Partitioning A technique consisting of dividing space in adjacent domains (through a partition) so that the evolution of each block is uniquely determined by the states of the elements within the block.

Phase transition Change of state obtained when varying a control parameter such as the one occurring in the boiling or freezing of a liquid, or in the change between ferromagnetic and paramagnetic states of a magnetic solid.

Propagation This is the process by which the particles of a LGA are moved to a nearest neighbor, according to the direction of their velocity vector $\mathbf{v_i}$. In one time step Δ_t the particle travel from cell $\mathbf{r}$ to cell $\mathbf{r} + \mathbf{v_i}\Delta_t$ where $\mathbf{r} + \mathbf{v_i}\Delta_t$ is the nearest neighbor in lattice direction i.

Random walk A series of uncorrelated steps of length unity describing a random path with zero average displacement but characteristic size proportional to the square root of the number of steps.

Reaction-diffusion systems Systems made of one or several species of particles which diffuse and react among themselves to produce some new species.

Scaling hypothesis A hypothesis concerning the analytical properties of the thermodynamic potentials and the correlation functions in a problem invariant under a change of scale.

Scaling law Relations among the critical exponents describing the power law behaviors of physical quantities in systems invariant under a change of scale.

Self-organized criticality Concept aimed at describing a class of dynamical systems which naturally drive themselves to a state where interesting physics occurs at all scales.

Site Same as a cell, but preferred terminology in LGA and LB models.

Spatially extended systems Physical systems involving many spatial degrees of freedom and which, usually, have rich dynamics and show complex behaviors. Coupled map lattices and cellular automata provides a way to model spatially extended systems.

Spin Internal degree of freedom associated to particles in order to describe their magnetic state. A widely used case is the one of classical Ising spins. To each particle, one associates an "arrow" which is allowed to take only two different orientations, up or down.

Time step Interval of time separating two consecutive iterations in the evolution of a discrete time process, like a CA or a LB model. Throughout this work the time step is denoted by the symbol Δ_t.

Universality The phenomenon whereby many microscopically different systems exhibit a critical behavior with quantitatively identical properties such as the critical exponents.

Updating operation consisting of assigning a new value to a set of variables, for instance those describing the states of a cellular automata system. The updating can be done in parallel and synchronously as is the case in CA dynamics or sequentially, one variable after another, as is usually the case for Monte–Carlo dynamics. Parallel, asynchronous updating is less common but can be envisaged too. Sequential and parallel updating schemes may yield different results since the interdependencies between variables are treated differently.

Viscosity A property of a fluid indicating how much momentum "diffuses" through the fluid in a inhomogeneous flow pattern. Equivalently, it describes the stress occurring between two fluid layers moving with different velocities. A high viscosity means that the resulting drag force is important and low viscosity means that this force is weak. Kinematic viscosity is usually denoted by ν and dynamic viscosity is denoted by $\eta = \nu\rho$ where ρ is the fluid density.

von Neumann neighborhood On a two-dimensional square lattice, the neighborhood including a central cell and its nearest neighbors north, south, east and west.

Ziff model A simple model describing adsorption–dissociation–desorption on a catalytic surface. This model is based upon some of the known steps of the reaction A–B_2 on a catalyst surface (for example CO–O_2).

Definition of the Subject

The computational science community has always been faced with the challenge of bringing efficient numerical tools to solve problems of increasing difficulty. Nowadays, the investigation and understanding of the so-called complex systems, and the simulation of all kinds of phenomena originating from the interaction of many components are of central importance in many area of science.

Cellular automata turn out to be a very fruitful approach to address many scientific problems by providing an efficient way to model and simulate specific phenomena for which more traditional computational techniques are hardly applicable.

The goal of this article is to provide the reader with the foundation of this approach, as well as a selection of simple applications of the cellular automata approach to the modeling of physical systems. We invite the reader to consult the web site http://cui.unige.ch/~chopard/CA/Animations/img-root.html in order to view short movies about several of the models discussed in this article.

Introduction

Cellular automata (hereafter termed CA) are an idealization of the physical world in which space and time are of discrete nature. In addition to space and time, the physical quantities also take only a finite set of values. Since it has been proposed by von Neumann in the late 1940s, the cellular automata approach has been applied to a large range of scientific problems (see for instance [4,10,16,35,42,51]). International conferences (e. g. ACRI) and dedicated journals (J. of Cellular Automata) also describe current developments.

When von Neumann developed the concept of CA, its motivation was to extract the abstract (or algorithmic) mechanisms leading to self-reproduction of biological organisms [6].

Following the suggestions of S. Ulam, von Neumann addressed this question in the framework of a fully discrete universe made up of simple cells. Each cell was characterized by an internal state, which typically consists of a finite number of information bits. Von Neumann suggested that this system of cells evolves, in discrete time steps, like simple automata which only know of a simple recipe to compute their new internal state. The rule determining the evolution of this system is the same for all cells and is a function of the states of the cell itself and its neighbors.

Similarly to what happens in any biological system, the activity of the cells takes place simultaneously. The same clock is assumed to drive the evolution of every cell and the updating of their internal state occurs synchronously.

Such a fully discrete dynamical system (cellular space), as invented by von Neumann, is now referred to as a cellular automaton.

Among the early applications of CA, the *game of life* [19] is famous. In 1970, the mathematician John Conway proposed a simple model leading to complex behaviors. He imagined a two-dimensional square lattice, like a checkerboard, in which each cell can be either alive (state one) or dead (state zero). The updating rule of the game of life is as follows: a dead cell surrounded by exactly three living cells gets back to life; a living cell surrounded by less than two or more than three neighbors dies of isolation or overcrowdness. Here, the surrounding cells correspond to the neighborhood composed of the four nearest cells (North, South, East and West), plus the four second nearest neighbors, along the diagonals. It turns out that the game of life automaton has an unexpectedly rich behavior. Complex structures emerge out of a primitive "soup" and evolve so as to develop some skills that are absent of the elementary cells (see Fig. 1).

The game of life is a cellular automata capable of universal computations: it is always possible to find an initial configuration of the cellular space reproducing the behavior of any electronic gate and, thus, to mimic any computation process. Although this observation has little practical interest, it is very important from a theoretical point of view since it assesses the ability of CAs to be a nonrestrictive computational technique.

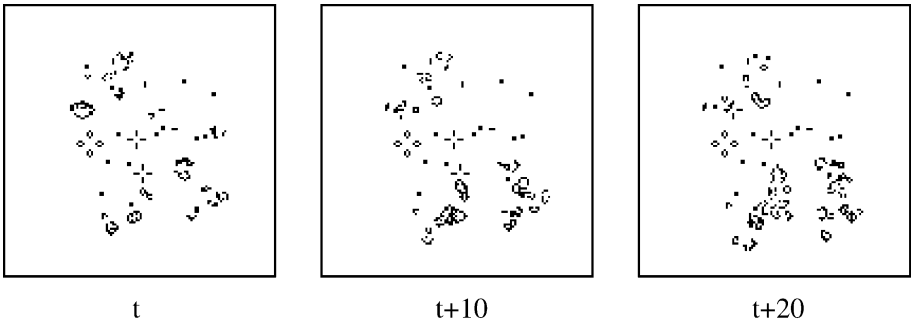
A very important feature of CAs is that they provide simple models of complex systems. They exemplify the fact that a collective behavior can emerge out of the sum of many, simply interacting, components. Even if the basic and local interactions are perfectly known, it is possible that the global behavior obeys new laws that are not obviously extrapolated from the individual properties, as if the whole were more than the sum of all the parts. These properties make cellular automata a very interesting approach to model physical systems and in particular to simulate complex and nonequilibrium phenomena.

The studies undertaken by S. Wolfram in the 1980s [50,51] clearly establishes that a CA may exhibit many of the behaviors encountered in continuous systems, yet in a much simpler mathematical framework. A further step is to recognize that CAs are not only behaving similarly to some dynamical processes, they can also represent an actual model of a given physical system, leading to macroscopic predictions that could be checked experimentally.

This fact follows from statistical mechanics which tells us that the macroscopic behavior of many systems is often only weakly related to the details of its microscopic reality. Only symmetries and conservation laws survive to the change of observation level: it is well known that the flows of a fluid, a gas or even a granular media are very similar at a macroscopic scale, in spite of their different microscopic nature.

When one is interested in the global or macroscopic properties of a system, it is therefore a clear advantage to invent a much simpler microscopic reality, which is more appropriate to the available numerical means of investigation.

An interesting example is the FHP fluid model proposed by Frisch, Hasslacher and Pomeau in 1986 [18] which can be viewed as a fully discrete molecular dynamic and yet behaves as predicted by the Navier–Stokes equation when the observation time and length scales are much larger than the lattice and automaton time step.



Cellular Automata Modeling of Physical Systems, Figure 1

The game of life automaton. *Black dots* represent living cells whereas dead cells are *white*. The figure shows the evolution of a random initial configuration and the formation of spatial structures, with possibly some emerging functionalities. The figure shows the evolution of some random initial configurations

A cellular automata model can then be seen as an idealized universe with its own microscopic reality but, nevertheless, with the same macroscopic behavior as given in the real system.

The cellular automata paradigm presents nevertheless some weaknesses inherent to its discrete nature. In the early 1990s, Lattice Boltzmann (LB) models were proposed to remedy some of these problems, using real-valued states instead of Boolean variables. It turns out that LB models are indeed a very powerful approach which combines numerical efficiency with the advantage of having a model whose microscopic components are intuitive. LB-fluids are more and more used to solve complex flows such as multi-component fluids or complicated geometries problems. See for instance [10,40,41,49] for an introduction to LB models.

Definition of a Cellular Automata

In order to give a definition of a cellular automaton, we first present a simple example, the so-called parity rule. Although it is very basic, the rule we discuss here exhibits a surprisingly rich behavior. It was proposed initially by Edward Fredkin in the 1970s [3] and is defined on a two-dimensional square lattice.

Each site of the lattice is a cell which is labeled by its position $\mathbf{r} = (i, j)$ where i and j are the row and column indices. A function $\psi(\mathbf{r}, t)$ is associated with the lattice to describe the state of each cell $\mathbf{r}$ at iteration t. This quantity can be either 0 or 1.

The cellular automata rule specifies how the states $\psi(\mathbf{r}, t+1)$ are to be computed from the states at iteration t. We start from an initial condition at time $t = 0$ with a given configuration of the values $\psi(\mathbf{r}, t = 0)$ on the lattice. The state at time $t = 1$ will be obtained as follows

(1) Each site $\mathbf{r}$ computes the sum of the values $\psi(\mathbf{r}', 0)$ on the four nearest neighbor sites $\mathbf{r}'$ at north, west, south, and east. The system is supposed to be periodic in both i and j directions (like on a torus) so that this calculation is well defined for all sites.

(2) If this sum is even, the new state $\psi(\mathbf{r}, t = 1)$ is 0 (white) and, else, it is 1 (black).

The same rule (steps 1 and 2) is repeated over to find the states at time $t = 2, 3, 4, \ldots$.

From a mathematical point of view, this cellular automata parity rule can be expressed by the following relation

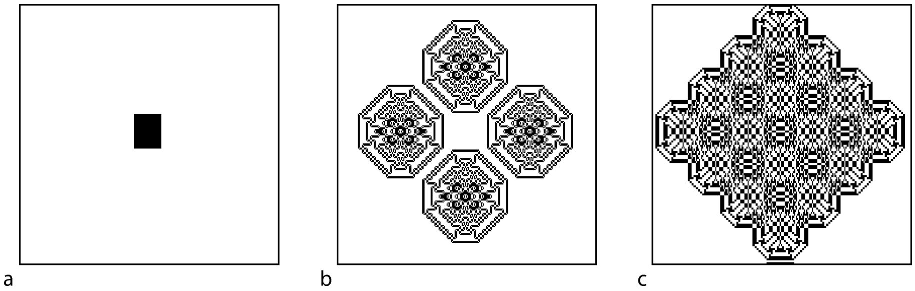
$$\begin{aligned}\psi(i, j, t+1) = &\psi(i+1, j, t) \oplus \psi(i-1, j, t)\\ &\oplus \psi(i, j+1, t) \oplus \psi(i, j-1, t)\end{aligned} \quad (1)$$

where the symbol $\oplus$ stands for the exclusive OR logical operation. It is also the sum modulo 2: $1 \oplus 1 = 0 \oplus 0 = 0$ and $1 \oplus 0 = 0 \oplus 1 = 1$.

When this rule is iterated, very nice geometric patterns are observed, as shown in Fig. 2. This property of generating complex patterns starting from a simple rule is generic of many cellular automata rules. Here, complexity results from some spatial organization which builds up as the rule is iterated. The various contributions of successive iterations combine together in a specific way. The spatial patterns that are observed reflect how the terms are combined algebraically.

A computer implementation of this CA can be given. In Fig. 3 we propose as, an illustration, a Matlab program (the reader can also consider Octave which is a free version of Matlab).

On the basis of this example we now give a definition of a cellular automata. Formally, a cellular automata is a tuple $(A, \Psi, R, \mathcal{N})$ where



Cellular Automata Modeling of Physical Systems, Figure 2
The ⊕ rule (or parity rule) on a 256 × 256 periodic lattice. a Initial configuration; b and c configurations after $t_b = 93$ and $t_c = 110$ iterations, respectively

```
nx=128; ny=128;  % size of the domain: 128x128
a=zeros(nx,ny);  % the states are first initialized to 0

north=[ 2:nx,1];    % vectors to access the neighbors
south=[nx, 1:nx-1]; % corresponding to a cyclic permutation
east=[2:ny,1];      % of 1:nx  or 1:ny
west=[ny,1:ny-1];

% a central patch  is initialized with 1's
a(nx/2-3:nx/2+2, ny/2-4:ny/2+3)=1;

for t=1:65   % let us do 65 iterations
     pcolor(a)       % build a graphical representation
     axis off
     axis square
     shading flat
     drawnow         % display it
     somme=a(north,:) + a(south,:) + a(:,west) + a(:,east);
     a=mod(somme,2);
end
```

Cellular Automata Modeling of Physical Systems, Figure 3
A example of a Matlab program for the parity rule

(i) A is a regular lattice of cells covering a portion of a d-dimensional space.

(ii) A set $\Psi(\mathbf{r},t) = \{\Psi^{(1)}(\mathbf{r},t), \Psi^{(2)}(\mathbf{r},t), \ldots, \Psi^{(m)}(\mathbf{r},t)\}$ of m Boolean variables attached to each site $\mathbf{r}$ of the lattice and giving the local state of the cells at time t.

(iii) A set R of rules, $R = \{R^{(1)}, R^{(2)}, \ldots, R^{(m)}\}$, which specifies the time evolution of the states $\Psi(\mathbf{r},t)$ in the following way

$$\Psi^{(j)}(\mathbf{r}, t+\Delta_t) = R^{(j)}(\Psi(\mathbf{r},t), \Psi(\mathbf{r}+\mathbf{v_1},t), \\ \Psi(\mathbf{r}+\mathbf{v_2},t), \ldots, \Psi(\mathbf{r}+\mathbf{v_q},t)) \tag{2}$$

where $\mathbf{r}+\mathbf{v_k}$ designate the cells belonging to the neighborhood $\mathcal{N}$ of cell $\mathbf{r}$.

In the above definition, the rule R is identical for all sites and is applied simultaneously to each of them, leading to synchronous dynamics. As the number of configurations of the neighborhood is finite, it is common to precompute all the values of R in a lookup table. Otherwise, an algebraic expression can be used and evaluated at each iteration, for each cell, as in Eq. (1).

It is important to notice that the rule is *homogeneous*, that is it cannot not depend explicitly on the cell position $\mathbf{r}$. However, spatial (or even temporal) inhomogeneities can be introduced anyway by having some $\Psi_j(\mathbf{r})$ systematically

1 in some given locations of the lattice to mark particular cells on which a different rule applies. Boundary cells are a typical example of spatial inhomogeneities. Similarly, it is easy to alternate between two rules by having a bit which is 1 at even time steps and 0 at odd time steps.

The neighborhood $\mathcal{N}$ of each cell (i. e. the spatial region around each cell used to compute the next state) is usually made of its adjacent cells. It is often restricted to the nearest or next to nearest neighbors, otherwise the complexity of the rule is too large. For a two-dimensional cellular automaton, two neighborhoods are often considered: the von Neumann neighborhood which consists of a central cell (the one which is to be updated) and its four geographical neighbors North, West, South, and East. The Moore neighborhood contains, in addition, the second nearest neighbor North-East, North-West, South-West, and South-East, that is a total of nine cells. Another interesting neighborhood is the Margolus neighborhood briefly described in the glossary.

According to the above definition, a cellular automaton is deterministic. The rule **R** is some well-defined function and a given initial configuration will always evolve identically. However, it may be very convenient for some applications to have a certain degree of randomness in the rule. For instance, it may be desirable that a rule selects one outcome among several possible states, with a probability p. Cellular automata whose updating rule is driven by some external probabilities are called *probabilistic* cellular automata. On the other hand, those which strictly comply with the definition given above, are referred to as *deterministic* cellular automata.

Probabilistic cellular automata are a very useful generalization because they offer a way to adjust the parameters of a rule in a continuous range of values, despite the discrete nature of the cellular automata world. This is very convenient when modeling physical systems in which, for instance, particles are annihilated or created at some given rate.

Limitations, Advantages, Drawbacks, and Extensions

The interpretation of the cellular automata dynamics in terms of simple "microscopic" rules offers a very intuitive and powerful approach to model phenomena that are very difficult to include in more traditional approaches (such as differential equations). For instance, boundary conditions are often naturally implemented in a cellular automata model because they have a natural interpretation at this level of description (e. g. particles bouncing back on an obstacle).

Numerically, an advantage of the CA approach is its simplicity and its adequation to computer architectures and parallel machines. In addition, working with Boolean quantities prevent numerical instabilities since an exact computation is made. There is no truncation or approximation in the dynamics itself. Finally, a CA model is an implementation of an N-body system where all correlations are taken into account, as well as all spontaneous fluctuations arising in a system made up of many particles.

On the other hand, cellular automata models have several drawbacks related to their fully discrete nature. An important one is the statistical noise requiring systematic averaging processes. Another one is the little flexibility to adjust parameters of a rule in order to describe a wider range of physical situations.

The Lattice Boltzmann approach solves several of the above problems. On the other hand, it may be numerically unstable and, also, requires some hypotheses of molecular chaos which reduces the some of the richness of the original CA dynamics [10].

Finally, we should remark that the CA approach is not a rigid framework but should allow for many extensions according to the problem at hand. The CA methodology is a philosophy of modeling where one seeks a description in terms of simple but essential mechanisms. Its richness and interest comes from the microscopic contents of its rule for which there is, in general, a clear physical or intuitive interpretation of the dynamics directly at the level of the cell.

Einstein's quote "Everything should be made as simple as possible, but not simpler" is a good illustration of the CA methodology to the modeling of physical systems.

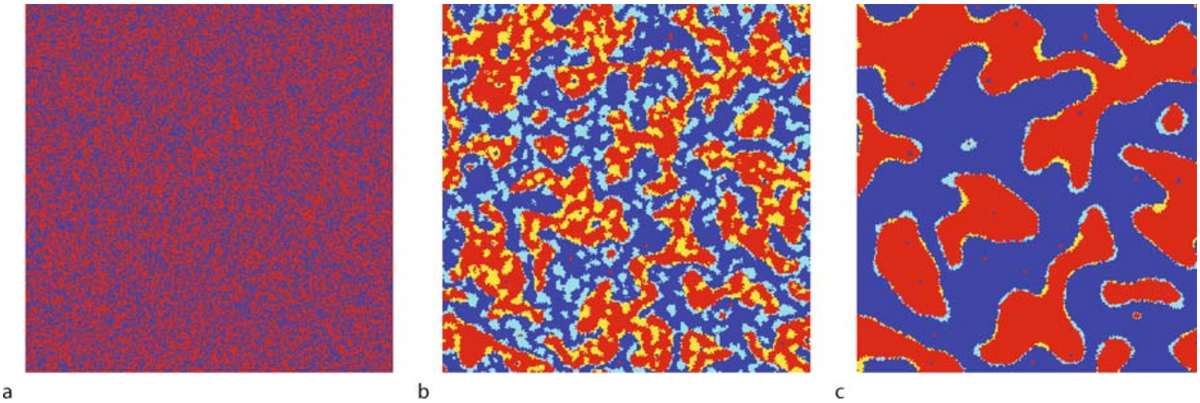
Applications

Many physical situations, like fluid flows, pattern formation, reaction-diffusion processes, nucleation-aggregation growth phenomena, phase transition, population dynamics, or traffic processes are very well suited to the cellular automata approach because both space and time play a central role in the behavior of these systems.

Below we describe several applications which illustrate the potential of the approach and that can be extended in order to address a wide class of scientific problems.

A Growth Model

A simple class of cellular automata rules consists of the so-called *majority rules*. The updating selects the new state of each cell so as to conform to the value currently held by the majority of the neighbors. Typically, in these majority rules, the state is either 0 or 1.



Cellular Automata Modeling of Physical Systems, Figure 4
Evolution of the twisted majority. The inherent "surface tension" present in the rule tends to separate the *red phases* $s = 1$ from the blue phase $s = 0$. The snapshots a, b and c correspond to $t = 0$, $t = 72$, and $t = 270$ iterations, respectively. The other colors indicate how "capes" have been eroded and "bays" filled: *light blue* shows the blue regions that have been eroded during the last few iterations and *yellow* marks the red regions that have been filled

A very interesting behavior is observed with the twisted majority rule proposed by G. Vichniac [44]: in two-dimensions, each cell considers its Moore neighborhood (i. e. itself plus its eight nearest neighbors) and computes the sum of the cells having a value 1. This sum can be any value between 0 and 9. The new state $s(t+1)$ of each cell is then determined from this local sum, according to the following table

$$\begin{array}{ccccccccccc} \text{sum}(t) & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ s(t+1) & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \end{array} \tag{3}$$

As opposed to the plain majority rule, here, the two middle entries of the table have been swapped. Therefore, when there is a slight majority of 1 around a cell, it turns to 0. Conversely, if there is a slight majority of 0, the cell becomes 1.

Surprisingly enough this rule describes the interface motion between two phases, as illustrated in Fig. 4. Vichniac has observed that the normal velocity of the interface is proportional to its local curvature, as required by the Allen–Cahn [21] equation. Of course, due to its local nature, the rule cannot detect the curvature of the interface directly. However, as the rule is iterated, local information is propagated to the nearest neighbors and the radius of curvature emerges as a collective effect.

This rule is particularly interesting when the initial configuration is a random mixture of the two phases, with equal concentration. Otherwise, some pathological behaviors may occur. For instance, an initial square of 1's surrounded by zero's will not evolve: 90-degree angles are not eroded and remain stable structures.

Ising-Like Dynamics

The Ising model is extensively used in physics. Its basic constituents are spins s_i which can be in one of two states: $s_i \in \{-1, 1\}$. These spins are organized on a regular lattice in d-dimensions and coupled in the sense that each pair (s_i, s_j) of neighbor spins contributes an amount $-Js_is_j$ to the energy of the system. Intuitively, the dynamics of such a system is that a spin flips ($s_i \to -s_i$) if this is favorable in view of the energy of the local configuration.

Vichniac [44], in the 1980s, proposed a CA rule, called Q2R, simulating the behavior of Ising spin dynamics. The model is as follows:

We consider a two-dimensional square lattice such that each site holds a spin s_i which is either up ($s_i = 1$) or down ($s_i = 0$) (instead of ±1). The coupling between spins is assumed to come from the von Neumann neighborhood (i. e. north, west, south, and east neighbors).

In this simple model, the spins will flip (or not flip) during their discrete time evolution according to a local energy conservation principle. This means we are considering a system which cannot exchange energy with its surroundings. The model will be a microcanonical cellular automata simulation of Ising spin dynamics, without a temperature but with a critical energy.

A spin s_i can flip at time t to become $1 - s_i$ at time $t+1$ if and only if this move does not cause any energy change. Accordingly, spin s_i will flip if the number of its neighbors with spin up is the same as the number of its neighbors with spin down. However, one has to remember that the motion of all spins is simultaneous in a cellular automata. The decision to flip is based on the assumption that the neighbors are not changing. If they are allowed to flip too, (because they obey the same rule), then energy may not be conserved.

A way to cure this problem is to split the updating in two phases and consider a partition of the lattice in odd and even sites (e. g. the white and black squares of a chessboard in 2D): first, one flips the spins located at odd positions, according to the configuration of the even spins. In the second phase, the even sublattice is updated according to the odd one. The spatial structure (defining the two sublattices) is obtained by adding an extra bit b to each lattice site, whose value is 0 for the odd sublattice and 1 for the even sublattice. The flipping rule described earlier is then regulated by the value of b. It takes place only for those sites for which $b = 1$. Of course, the value of b is also updated at each iteration according to $b(t + 1) = 1 - b(t)$, so that at the next iteration, the other sublattice is considered. In two-dimensions, the Q2R rule can be then expressed by the following expressions

$$s_{ij}(t{+}1) = \begin{cases} 1 - s_{ij}(t) & \text{if } b_{ij} = 1 \text{ and } s_{i-1,j} \\ & +s_{i+1,j} + s_{i,j-1} + s_{i,j+1} = 2 \\ s_{ij}(t) & \text{otherwise} \end{cases} \tag{4}$$

and

$$b_{ij}(t + 1) = 1 - b_{ij}(t) \tag{5}$$

where the indices (i, j) label the Cartesian coordinates and $s_{ij}(t = 0)$ is either one or zero.

The question is now how well does this cellular automata rule perform to describe an Ising model? Figure 5 shows a computer simulation of the Q2R rule, starting from an initial configuration with approximately 11% of spins $s_{ij} = 1$ (Fig. 5a). After a transient phase (figures b and c), the system reaches a stationary state where domains with "up" magnetization (white regions) are surrounded by domains of "down" magnetization (black regions).

In this dynamics, energy is exactly conserved because that is the way the rule is built. However, the number of spins down and up may vary. In the present experiment, the fraction of spins up increases from 11% in the initial state to about 40% in the stationary state. Since there is an excess of spins down in this system, there is a resulting macroscopic magnetization.

It is interesting to study this model with various initial fractions ρ_s of spins up. When starting with a random initial condition, similar to that of Fig. 5a, it is observed that, for many values of ρ_s, the system evolves to a state where there is, in the average, the same amount of spin down and up, that is no macroscopic magnetization. However, if the initial configuration presents a sufficiently large excess of one kind of spins, then a macroscopic magnetization builds up as time goes on. This means there is a phase transition between a situation of zero magnetization and a situation of positive or negative magnetization.

It turns out that this transition occurs when the total energy E of the system is low enough (a low energy means that most of the spins are aligned and that there is an excess of one species over the other), or more precisely when E is smaller than a critical energy E_c. In that sense, the Q2R rule captures an important aspect of a real magnetic system, namely a non-zero magnetization at low energy (which can be related to a low temperature situation) and a transition to a nonmagnetic phase at high energy.

However, Q2R also exhibits unexpected behavior that is difficult to detect from a simple observation. There is a breaking of ergodicity: a given initial configuration of energy E_0 evolves without visiting completely the region of the phase space characterized by $E = E_0$.

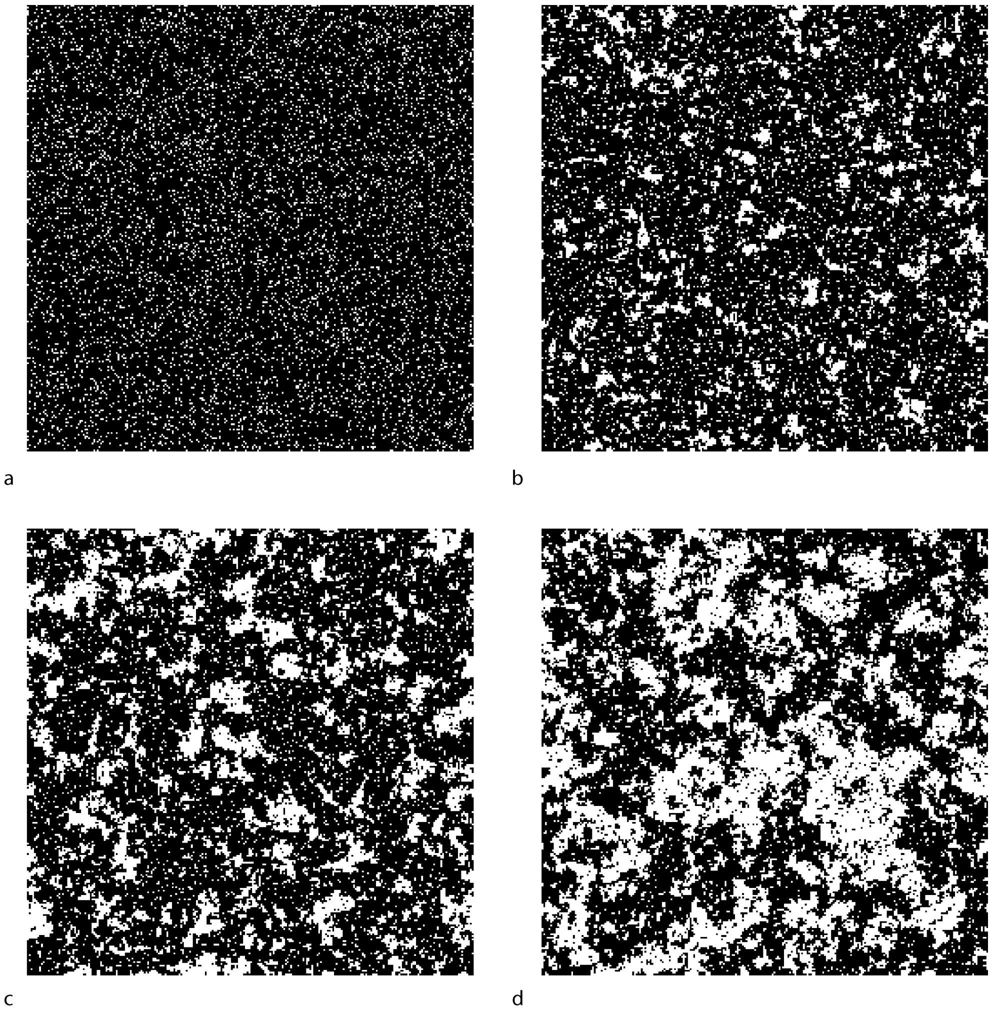
This is illustrated by the following simple 1D example, where a ring of four spins with periodic boundary condition are considered.

$$\begin{aligned} t&: \ 1001 \\ t+1&: \ 1100 \\ t+2&: \ 0110 \\ t+3&: \ 0011 \\ t+4&: \ 1001 \end{aligned} \tag{6}$$

After four iterations, the system cycles back to its original state. The configuration of this example has $E_0 = 0$. As we observed, it never evolves to 0111, which is also a configuration of zero energy. This nonergodicity means that not only energy is conserved during the evolution of the automaton, but also another quantity which partitions the energy surface in independent regions.

Competition Models and Cell Differentiation

In Sect. "A Growth Model" we have discussed a majority rule in which the cells imitate their neighbors. In some sense, this corresponds to a cooperative behavior between the cells. A quite different situation can be obtained if the cells obey competitive dynamics. For instance, we may imagine that the cells compete for some resources at the expense of their nearest neighbors. A winner is a cell of state 1 and a loser a cell of state 0. No two winner cells can be neighbors and any loser cell must have at least one winner neighbor (otherwise nothing would have prevented it to also win).

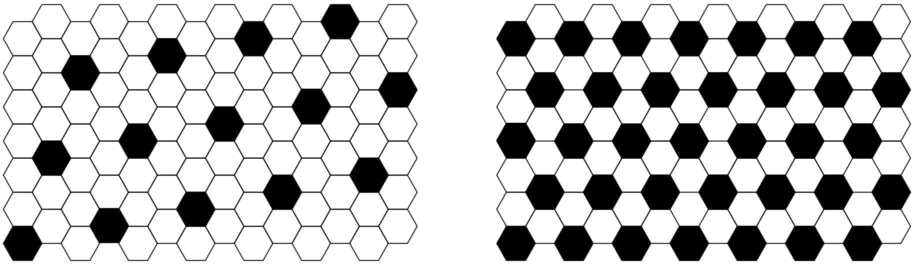


Cellular Automata Modeling of Physical Systems, Figure 5
Evolution of a system of spins with the Q2R rule. *Black* represents the spins down $s_{ij} = 0$ and *white* the spins up $s_{ij} = 1$. The four images a, b, c, and d show the system at four different times $t_a = 0 < t_b \ll t_{cd}$

It is interesting to note that this problem has a direct application in biology, to study cell differentiation. It has been observed in the development of Drosophila that about 25% of the cells forming the embryo evolve to the state of neuroblast, while the remaining 75% do not. How can we explain this differentiation and the observed fraction since, at the beginning of the process all cells can be assumed equivalent? A possible mechanism [28] is that some competition takes place between the adjacent biological cells. In other words, each cell produces some substance *S* but the production rate is inhibited by the amount of *S* already present in the neighboring cells. Differentiation occurs when a cell reaches a level of *S* above a given threshold.

The competition CA model we propose to describe this situation is the following. Because of the analogy with the biological system, we shall consider a hexagonal lattice, which is a reasonable approximation of the cell arrangement observed in the Drosophila embryo (see Fig. 6). We assume that the values of *S* can be 0 (inhibited) or 1 (active) in each lattice cell.

- A $S = 0$ cell will grow (i. e. turn to $S = 1$) with probability p_{grow} provided that all its neighbors are 0. Otherwise, it stays inhibited.
- A cell in state $S = 1$ will decay (i. e. turn to $S = 0$) with probability p_{decay} if it is surrounded by at least one active cell. If the active cell is isolated (all the neighbors are in state 0) it remains in state 1.



Cellular Automata Modeling of Physical Systems, Figure 6
The hexagonal lattice used for the competition-inhibition CA rule. *Black cells* are cells of state 1 (winners) and *white cells* are cells of state 0 (losers). The two possible final states with a fully regular structure are illustrated with density 1/3 and 1/7 of a winner, respectively

The evolution stops (stationary process) when no $S=1$ cell feels any more inhibition from its neighbors and when all $S=0$ cells are inhibited by their neighborhood. Then, with our biological interpretation, cells with $S=1$ are those which will differentiate.

What is the expected fraction of these $S=1$ cells in the final configuration? Clearly, from Fig. 6, the maximum value is 1/3. According to the inhibition condition we imposed, this is the close-packed situation on the hexagonal lattice. On the other hand, the minimal value is 1/7, corresponding to a situation where the lattice is partitioned in blocks with one active cell surrounded by six inhibited cells. In practice we do not expect any of these two limits to occur spontaneously after the automaton evolution. On the contrary, we should observe clusters of close-packed active cells surrounded by defects, i. e. regions of low density of active cells.

CA simulations show indeed that the final fraction s of active cells is a mix of the two limiting situations of Fig. 6

$$.23 \le s \le .24$$

almost irrespectively of the values chosen for p_{anihil} and p_{growth}.

This is exactly the value expected from the biological observations made on the Drosophila embryo. Thus, cell differentiation can be explained by a geometrical competition without having to specify the inhibitory couplings between adjacent cell and the production rate (i. e. the values of p_{anihil} and p_{growth}): the result is quite robust against any possible choices.

Traffic Models

Cellular automata models for road traffic have received a great deal of interest during the past few years (see [13,32,33,36,37,47,48,52] for instance).

One-dimensional models for single lane car motions are quite simple and elegant. The road is represented as a line of cells, each of them being occupied or not by a vehicle. All cars travel in the same direction (say to the right). Their positions are updated synchronously. During the motion, each car can be at rest or jump to the nearest neighbor site, along the direction of motion. The rule is simply that a car moves only if its destination cell is empty. This means that the drivers do not know whether the car in front will move or will be blocked by another car. Therefore, the state of each cell s_i is entirely determined by the occupancy of the cell itself and its two nearest neighbors s_{i-1} and s_{i+1}. The motion rule can be summarized by the following table, where all eight possible configurations $(s_{i-1}s_is_{i+1})_t \to (s_i)_{t+1}$ are given

$$\underbrace{(111)}_{1}\;\underbrace{(110)}_{0}\;\underbrace{(101)}_{1}\;\underbrace{(100)}_{1}\;\underbrace{(011)}_{1}\;\underbrace{(010)}_{0}\;\underbrace{(001)}_{0}\;\underbrace{(000)}_{0}\,. \tag{7}$$

This cellular automaton rule turns out to be Wolfram rule 184 [50,52].

These simple dynamics capture an interesting feature of real car motion: traffic congestion. Suppose we have a low car density ρ in the system, for instance something like

$$\ldots 0010000010010000010\ldots\,. \tag{8}$$

This is a *free* traffic regime in which all the cars are able to move. The average velocity $\langle v \rangle$ defined as the number of motions divided by the number of cars is then

$$\langle v_{\text{f}} \rangle = 1 \tag{9}$$

where the subscript f indicates a free state. On the other hand, in a high density configuration such as

$$\ldots 110101110101101110\ldots\,. \tag{10}$$

only six cars over 12 will move and $\langle v \rangle = 1/2$. This is a partially jammed regime.

If the car positions were uncorrelated, the number of moving cars (i. e. the number of particle-hole pairs) would be given by $L\rho(1-\rho)$, where L is the system size. Since the number of cars is ρL, the average velocity would be

$$\langle v_{\text{uncorrel}} \rangle = 1 - \rho \,. \tag{11}$$

However, in this model, the car occupancy of adjacent sites is highly correlated and the vehicles cannot move until a hole has appeared in front of them. The car distribution tries to self-adjust to a situation where there is one spacing between consecutive cars. For densities less than one-half, this is easily realized and the system can organize to have one car every other site.

Therefore, due to these correlations, Eq. (11) is wrong in the high density regime. In this case, since a car needs a hole to move to, we expect that the number of moving cars simply equals the number of empty cells [52]. Thus, the number of motions is $L(1-\rho)$ and the average velocity in the jammed phase is

$$\langle v_j \rangle = \frac{1-\rho}{\rho} \,. \tag{12}$$

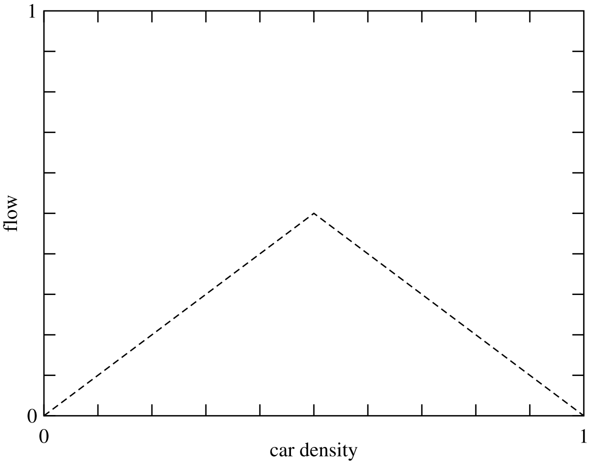
From the above relations we can compute the so-called fundamental flow diagram, i. e. the relation between the flow of cars $\rho\langle v \rangle$ as a function of the car density ρ: for $\rho \le 1/2$, we use the free regime expression and $\rho\langle v \rangle = \rho$. For densities $\rho > 1/2$, we use the jammed expression and $\rho\langle v \rangle = 1 - \rho$. The resulting diagram is shown in Fig. 7. As in real traffic, we observe that the flow of cars reaches a maximum value before decreasing.

A richer version of the above CA traffic model is due to Nagel and Schreckenberg [33,47,48]. The cars may have several possible velocities $u = 0, 1, 2, \ldots, u_{\max}$. Let u_i be the velocity of car i and d_i the distance, along the road, separating cars i and $i+1$. The updating rule is:

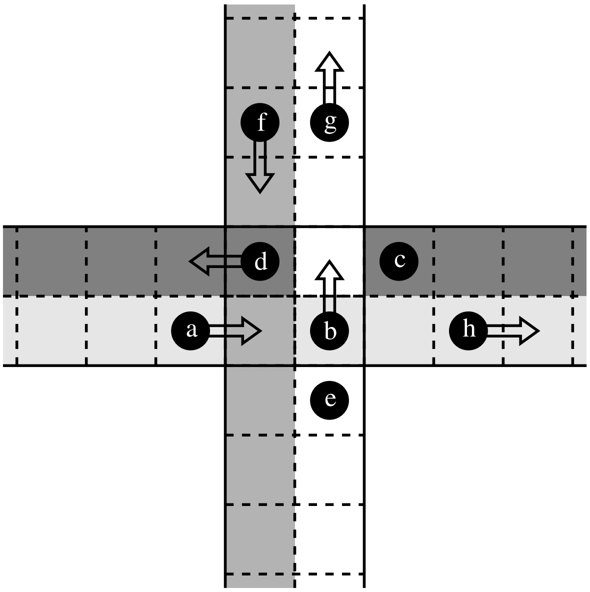
- The cars accelerate when possible: $u_i \to u_i' = u_i + 1$, if $u_i < u_{\max}$.
- The cars slow down when required: $u_i' \to u_i'' = d_i - 1$, if $u_i' \ge d_i$.
- The cars have a random behavior: $u_i'' \to u_i''' = u_i'' - 1$, with probability p_i if $u_i'' > 0$.
- Finally, the cars move u_i''' sites ahead.

This rule captures some important behaviors of real traffic on a highway: velocity fluctuations due to a nondeterministic behavior of the drivers, and “stop-and-go” waves observed in a high-density traffic regime.

We refer the reader to recent literature for the new developments of this topic. See for instance [24,25].

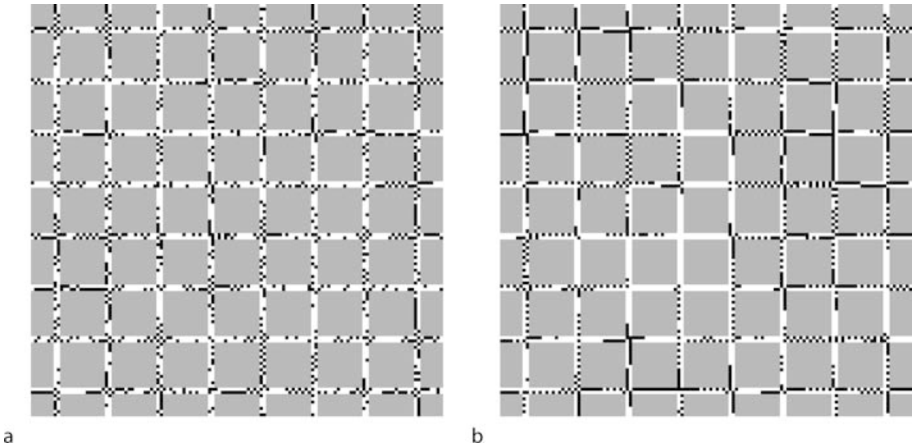


Cellular Automata Modeling of Physical Systems, Figure 7
Traffic flow diagram for the simple CA traffic rule

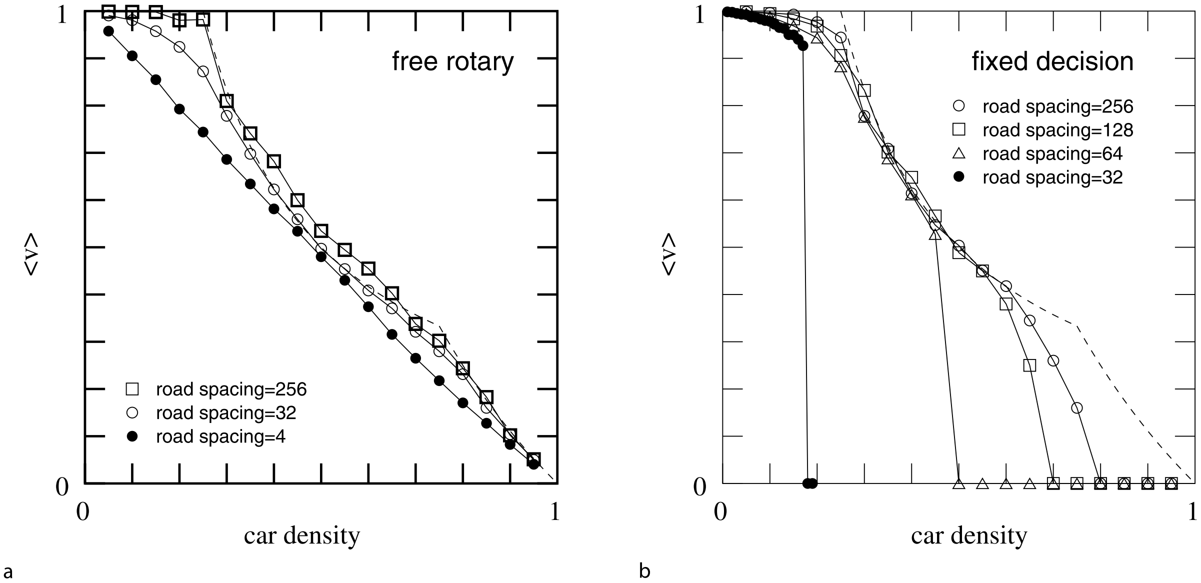


Cellular Automata Modeling of Physical Systems, Figure 8
The four central cells represent a roundabout which is traveled counterclockwise. The gray levels indicate the different traffic lanes: *white* is a northbound lane, *light gray* an eastbound lane, *gray* a southbound lane and, finally, *dark gray* is a westbound lane. The *dots* labeled *a*, *b*, *c*, *d*, *e*, *f*, *g*, and *h* are cars which will move to the destination cell indicated by the arrows, as determined by some local decision rule. Cars without an arrow are forbidden to move

Note that a street network can also be described using a CA. A possible approach is to model a road intersection as a roundabout. Cars in the roundabout have priority over those willing to enter. Figure 8 illustrates a simple four-way road junction.



Cellular Automata Modeling of Physical Systems, Figure 9
Traffic configuration after 600 iterations, for a car density of 30%. Streets are *white*, buildings *gray* and the *black pixels* represent the cars. Situation a corresponds to the roundabout junctions, whereas image b mimics the presence of traffic lights. In the second case, queues are more likely to form and the global mobility is less than in the first case



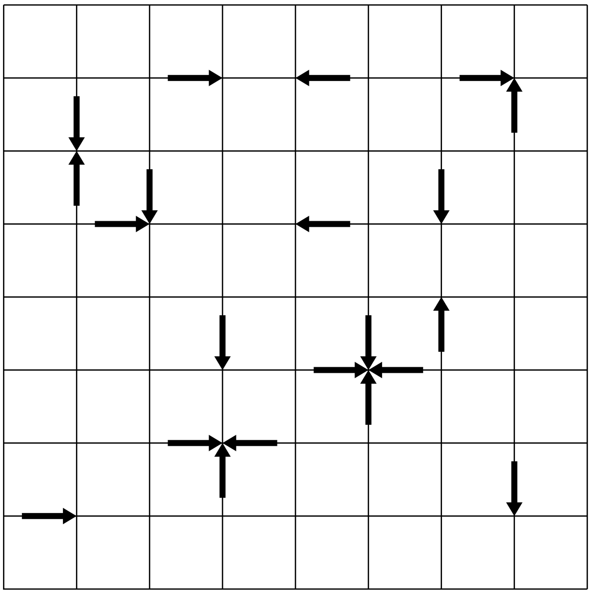
Cellular Automata Modeling of Physical Systems, Figure 10
Average velocity versus average density for the cellular automata street network, for a time-uncorrelated turning strategies and b a fixed driver decision. The different curves correspond to different distances *L* between successive road junctions. The *dashed line* is the analytical prediction (see [13]). Junction deadlock is likely to occur in b, resulting in a completely jammed state

Traffic lights can also be modeled in a CA by stopping, during a given number of iterations, the car reaching a given cell. Figure 9 illustrates CA simulation of a Manhattan-like city in which junctions are either controlled by a roundabout or by a traffic light. In both cases, the destination of a car reaching the junction is randomly chosen.

Figure 10 shows the fundamental flow diagram obtained with the CA model, for a Manhattan-like city governed by roundabouts separated by a distance *L*.

CA modeling of urban traffic has been used for real situations by many authors (see for instance [11]) and some cities in Europe and USA use the CA approach as a way to manage traffic.

Note finally that crowd motion is also addressed within the CA framework. Recent results [7,29] show that the approach is quite promising to plan evacuation strategies and reproduce several of the motion patterns observed in real crowds.



Cellular Automata Modeling of Physical Systems, Figure 11
Example of a configuration of HPP particles

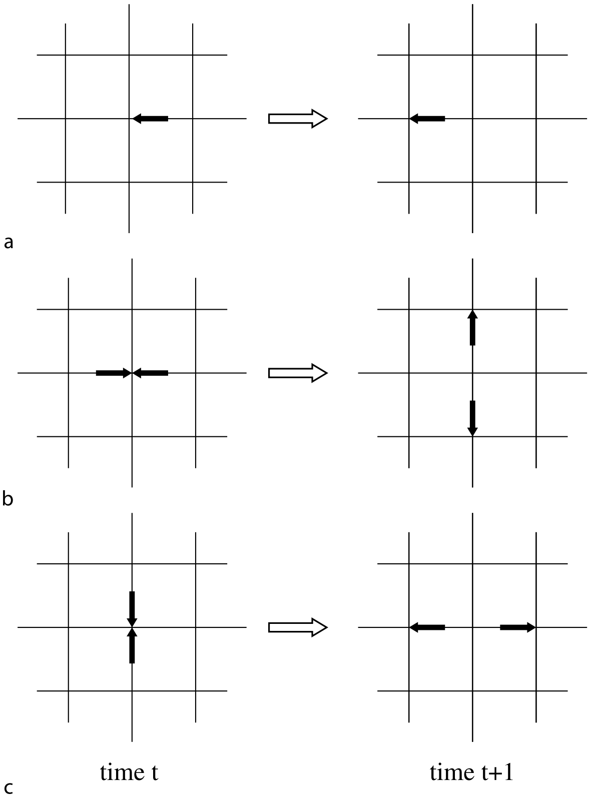
A Simple Gas: The HPP Model

The HPP rule is a simple example of an important class of cellular automata models: lattice gas automata (LGA). The basic ingredient of such models are point particles that move on a lattice, according to appropriate rules so as to mimic fully discrete "molecular dynamics."

The HPP lattice gas automata is traditionally defined on a two-dimensional square lattice. Particles can move along the main directions of the lattice, as shown in Fig. 11. The model limits to 1 the number of particles entering a given site with a given direction of motion. This is the exclusion principle which is common in most LGA (LGA models without the exclusion principle are called multi-particle models [10]).

With at most one particle per site and direction, four bits of information at each site are enough to describe the system during its evolution. For instance, if at iteration t site $\mathbf{r}$ has the following state $s(\mathbf{r}, t) = (1011)$, it means that three particles are entering the site along direction 1, 3, and 4, respectively.

The cellular automata rule describing the evolution of $s(\mathbf{r}, t)$ is often split into two steps: collision and motion (or propagation). The collision phase specifies how the particles entering the same site will interact and change their trajectories. During the propagation phase, the particles actually move to the nearest neighbor site they are traveling to. This decomposition into two phases is a quite convenient way to partition the space so that the collision rule is purely local.



Cellular Automata Modeling of Physical Systems, Figure 12
The HPP rule: a a single particle has a ballistic motion until it experiences a collision; b and c the two nontrivial collisions of the HPP model: two particles experiencing a head on collision are deflected in the perpendicular direction. In the other situations, the motion is ballistic, that is the particles are transparent to each other when they cross the same site

Figure 12 illustrates the HPP rules. According to our Boolean representation of the particles at each site, the collision part for the two head on collisions are expressed as

$$(1010) \to (0101) \qquad (0101) \to (1010) \tag{13}$$

all the other configurations being unchanged. During the propagation phase, the first bit of the state variable is shifted to the east neighbor cell, the second bit to the north and so on.

The aim of this rule is to reproduce some aspect of the real interactions between particles, namely that momentum and particle number are conserved during a collision. From Fig. 12, it is easily checked that these properties are obeyed: a pair of zero momentum particles along a given direction is transformed into another pair of zero momentum along the perpendicular axis.

It is easy to express the HPP model in a mathematical form. For this purpose, the so-called occupation numbers $n_i(\mathbf{r}, t)$ are introduced for each lattice site $\mathbf{r}$ and each time step t. The index i labels the lattice directions (or the possible velocities of the particles). In the HPP model, the lattice has four directions (North, West, South, and East) and i runs from 1 to 4.

By definition and due to the exclusion principle, the n_i's are Boolean variables

$$n_i(\mathbf{r}, t) = \begin{cases} 1 & \text{if a particle is entering site } \mathbf{r} \text{ at time } t \\ & \text{along lattice direction } i \\ 0 & \text{otherwise .} \end{cases}$$

From this definition it is clear that, for HPP, the n_i's are simply the components of the state s introduced above

$$s = (n_1, n_2, n_3, n_4) .$$

In an LGA model, the microdynamics can be naturally expressed in terms of the occupation numbers n_i as

$$n_i(\mathbf{r} + \mathbf{v_i}\Delta_t, t + \Delta_t) = n_i(\mathbf{r}, t) + \Omega_i(n(\mathbf{r}, t)) \qquad (14)$$

where $\mathbf{v_i}$ is a vector denoting the speed of the particle in the ith lattice direction. The function Ω is called the collision term and it describes the interaction of the particles which meet at the same time and same location.

Note that another way to express Eq. (14) is through the so-called collision and propagation operators C and P

$$n(t + \Delta_t) = PCn(t) \qquad (15)$$

where $n(t)$ describe the set of values $n_i(\mathbf{r}, t)$ for all i and $\mathbf{r}$. The quantities C and P act over the entire lattice. They are defined as

$$(Pn)_i(\mathbf{r}) = n_i(\mathbf{r} - \mathbf{v_i}\Delta_t) \qquad (Cn)_i(\mathbf{r}) = n_i(\mathbf{r}) + \Omega_i .$$

More specifically, for the HPP model, it can be shown [10] that the collision and propagation phase can be expressed as

$$\begin{aligned} &n_i(\mathbf{r} + \mathbf{v_i}\Delta_t, t + \Delta_t) \\ &\quad = n_i - n_i n_{i+2}(1 - n_{i+1})(1 - n_{i+3}) \\ &\qquad + n_{i+1}n_{i+3}(1 - n_i)(1 - n_{i+2}) . \end{aligned} \qquad (16)$$

In this equation, the values $i + m$ are wrapped onto the values 1 to 4 and the right-hand term is computed at position $\mathbf{r}$ and time t.

The HPP rule captures another important ingredient of the microscopic nature of a real interaction: invariance under time reversal. Figure 12b and c show that, if at some given time, the directions of motion of all particles are reversed, the system will just trace back its own history. Since the dynamics of a deterministic cellular automaton is exact, this fact allows us to demonstrate the properties of physical systems to return to their original situation when all the particles reverse their velocity.

Figure 13 illustrates the time evolution of an HPP gas initially confined in the left compartment of a container. There is an aperture on the wall of the compartment and the gas particles will flow so as to fill the entire space available to them. In order to include a solid boundary in the system, the HPP rule is modified as follows: when a site is a wall (indicated by an extra bit), the particles no longer experience the HPP collision but bounce back from where they came. Therefore, particles cannot escape a region delimited by such a reflecting boundary.

If the system of Fig. 13 is evolved, it reaches an equilibrium after a long enough time and no macroscopic trace of its initial state is visible any longer. However, no information has been lost during the process (no numerical dissipation) and the system has the memory of where it comes from. Reversing all the velocities and iterating the HPP rule makes all particles go back to the compartment in which they were initially located.

Reversing the particle velocity can be described by an operator R which swaps occupation numbers with opposite velocities

$$(Rn)_i = n_{i+2} .$$

The reversibility of HPP stem from the fact that the collision and propagation operators obey

$$PRP = R \qquad CRC = R .$$

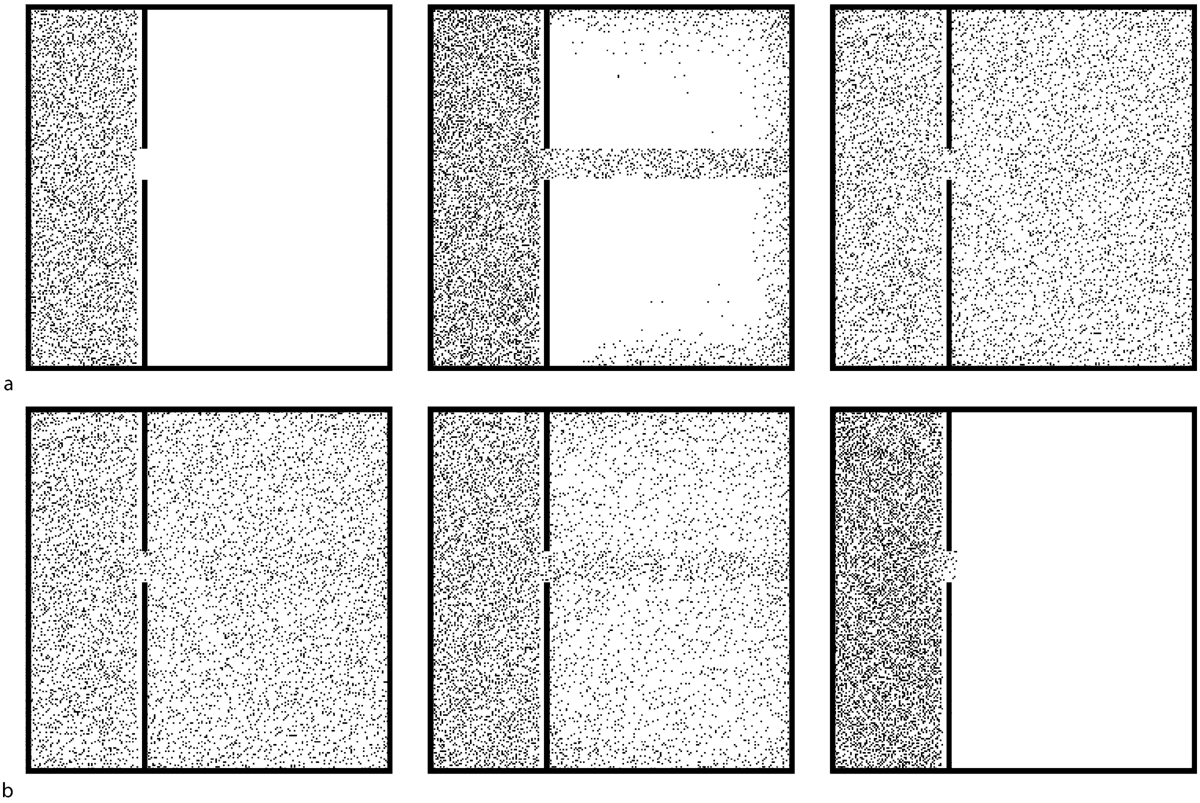
Thus

$$(PC)^n PR(PC)^n = R$$

which shows that the system will return to its initial state (though with opposite velocities) if the first n iterations are followed by a velocity change $\mathbf{v_i} \to -\mathbf{v_i}$, a propagation and again n iterations.

This time-reversal behavior is only possible because the dynamics are perfectly exact and no numerical errors are present in the numerical scheme. If one introduces externally some errors (for instance, one can add an extra particle in the system) before the direction of motion of each particle is reversed, then reversibility is lost.

Note that this property has inspired a new symmetric cryptography algorithm called Crystal [30], which exploits and develops the existing analogy between discrete physical models of particles and the standard diffusion-confusion paradigm of cryptography proposed by Shannon [39].



Cellular Automata Modeling of Physical Systems, Figure 13
Time evolution of an HPP gas. a From the initial state to equilibrium. b Illustration of time reversal invariance: in the rightmost image of a, the velocity of each particle is reversed and the particles naturally return to their initial position

The FHP Model

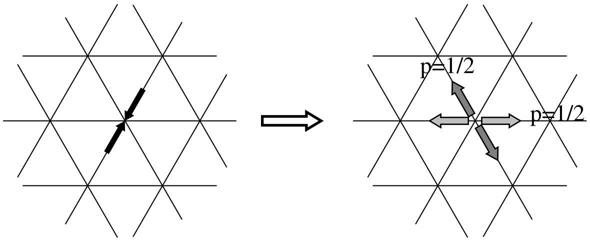
The HPP rule is interesting because it illustrates the basic ingredients of LGA models. However, the capability of this rule to model a real gas of particles is poor, due to a lack of isotropy and spurious invariants. A remedy to this problem is to use a different lattice and a different collision model.

The FHP rule (proposed by Frisch, Hasslacher, and Pomeau [18] in 1986) was the first CA whose behavior was shown to reproduce, within some limits, a two-dimensional fluid.

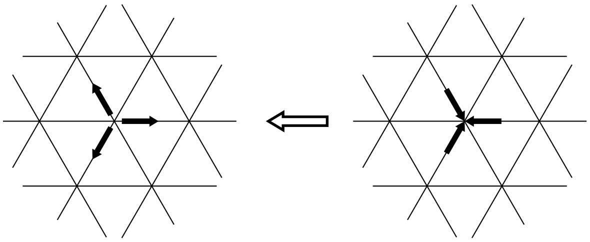
The FHP model is an abstraction, at a microscopic scale, of a fluid. It is expected to contain all the salient features of a real fluid. It is well known that the continuity and Navier–Stokes equations of hydrodynamics express the local conservation of mass and momentum in a fluid. The detailed nature of the microscopic interactions does not affect the form of these equations but only the values of the coefficients (such as the viscosity) appearing in them. Therefore, the basic ingredients one has to include in the microdynamics of the FHP model is the conservation of particles and momentum after each updating step. In addition, some symmetries are required so that, in the macroscopic limit, where time and space can be considered as continuous variables, the system be isotropic.

As in the case of the HPP model, the microdynamics of FHP is given in terms of Boolean variables describing the occupation numbers at each site of the lattice and at each time step (i. e. the presence or the absence of a fluid particle). The FHP particles move in discrete time steps, with a velocity of constant modulus, pointing along one of the six directions of the lattice.

Interactions take place among particles entering the same site at the same time and result in a new local distribution of particle velocities. In order to conserve the number of particles and the momentum during each interaction, only a few configurations lead to a nontrivial collision (i. e. a collision in which the directions of motion have changed). For instance, when exactly two particles enter the same site with opposite velocities, both of them are deflected by 60 degrees so that the output of the collision is



Cellular Automata Modeling of Physical Systems, Figure 14
The two-body collision in the FHP model. On the *right part of the figure*, the two possible outcomes of the collision are shown in *dark* and *light gray*, respectively. They both occur with probability one-half



Cellular Automata Modeling of Physical Systems, Figure 15
The three-body collision in the FHP model

still a zero momentum configuration with two particles. As shown in Fig. 14, the deflection can occur to the right or to the left, indifferently. For symmetry reasons, the two possibilities are chosen randomly, with equal probability.

Another type of collision is considered: when exactly three particles collide with an angle of 120 degrees between each other, they bounce back (so that the momentum after collision is zero, as it was before collision). Figure 15 illustrates this rule.

For the simplest case we are considering here, all interactions come from the two collision processes described above. For all other configurations (i. e. those which are not obtained by rotations of the situations given in Figs. 14 and 15) no collision occurs and the particles go through as if they were transparent to each other.

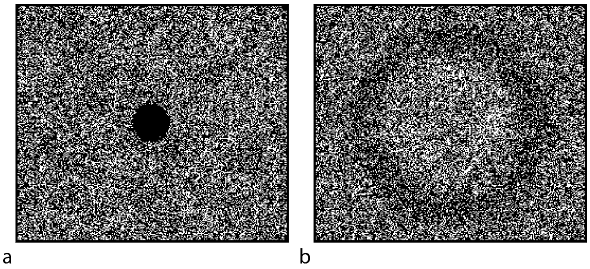
Both two- and three-body collisions are necessary to avoid extra conservation laws. The two-particle collision removes a pair of particles with a zero total momentum and moves it to another lattice direction. Therefore, it conserves momentum along each line of the lattice. On the other hand, three-body interactions deflect particles by 180 degrees and cause the net momentum of each lattice line to change. However, three-body collisions conserve the number of particles within each lattice line.

The FHP model has played a central role in computational physics because it can be shown (see for instance [10]) that the density ρ, defined as the average number of particles at a given lattice site and $\mathbf{u}$ the average velocity of these particles, obey Navier–Stokes equation

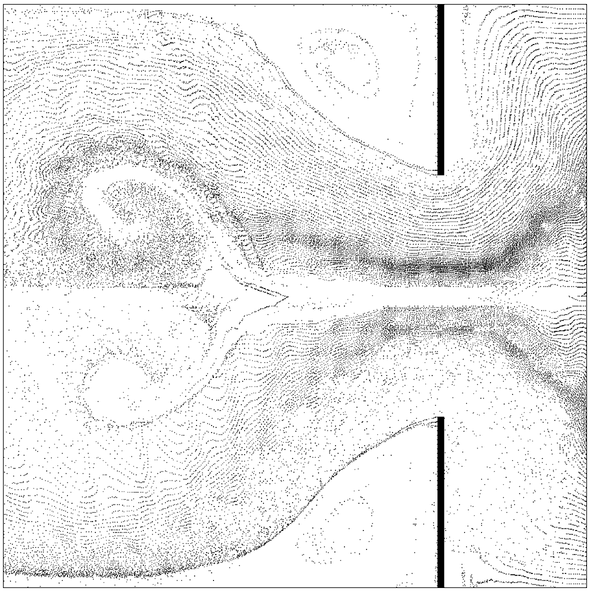
$$\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu\nabla^2\mathbf{u} \tag{17}$$

where $p = c_s^2\rho$ is the scalar pressure, with c_s the speed of sound and ν is the kinematic viscosity. Note that here both ν and c_s are quantities that emerge from the FHP dynamics. The speed of sound reflects the lattice topology whereas the viscosity reflects the details of the collision process.

As an illustration, Fig. 16 shows the propagation of a density wave in a FHP model. Figure 17 shows the eddies that form when a FHP-fluid flow against an obstacle.

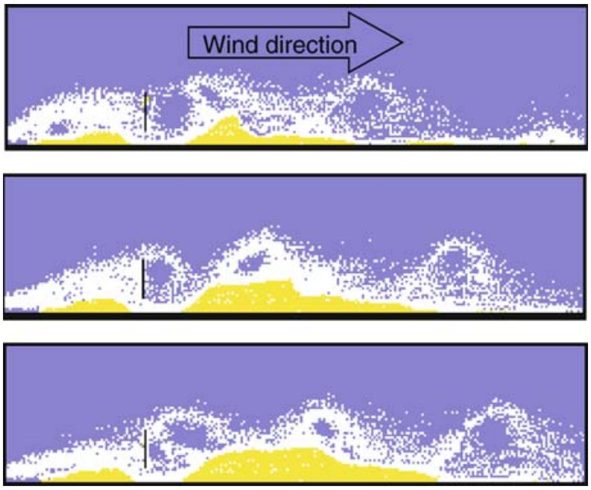


Cellular Automata Modeling of Physical Systems, Figure 16
Development of a sound wave in a FHP gas, due to an over particle concentration in the middle of the system



Cellular Automata Modeling of Physical Systems, Figure 17
Flow pattern from a simulation of a FHP model

Cellular Automata Modeling of Physical Systems, Figure 18
A CA snow transport and deposition model



Cellular Automata Modeling of Physical Systems, Figure 19
A lattice Boltzmann snow transport and deposition model. The three panels show the time evolution of the snow deposit (*yellow*) past a fence. Airborne snowflakes are shown as *white dots*

More complex models can be built by adding new processes on top of a FHP-fluid. For instance, Fig. 18 shows the result of a model of snow transport and deposition by wind. In addition to the wind flow, obtained from a FHP model, snow particles are traveling due to the combined effect of wind and gravity. Upon reaching the ground, they pile up (possible after toppling) so as to form a new boundary condition for the wind. In Fig. 19 an extension of the model (using the lattice Boltzmann approach described in Sect. "Lattice Boltzmann Models") shows how a fence with ground clearance creates a snow deposit.

Lattice Boltzmann Models

Lattice Boltzmann models (LBM) are an extension of the CA-fluid described in the previous section. The main conceptual difference is that in LBM, the CA state is no longer Boolean numbers n_i but real-valued quantity f_i for each lattice directions i. Instead of describing the presence or absence of a particle, the interpretation of f_i is the density distribution function of particles traveling in lattice directions i.

As with LGA (see Eq. (16)), the dynamics of LBM can be expressed in terms of an equation for the f_i's. The fluid quantities such as the density ρ and velocity field $\mathbf{u}$ are obtained by summing the contributions from all lattice directions

$$\rho = \sum_i f_i \qquad \rho\mathbf{u} = \sum_i f_i \mathbf{v_i}$$

where $\mathbf{v_i}$ denotes the possible particle velocities in the lattice.

From a numerical point of view, the advantages of suppressing the Boolean constraint are several: less statistical noise, more numerical accuracy and, importantly, more flexibility to choose the lattice topology, the collision operator and boundary conditions. Thus, for many practical applications, the LBM approach is preferred to the LGA one.

The so-called BGK (or "single-time relaxation") LBM

$$f_i(\mathbf{r} + \Delta_t \mathbf{v_i}, t + \Delta_t) = f_i(\mathbf{r}, t) + \frac{1}{\tau}\left(f_i^{\mathrm{eq}}(\rho, \mathbf{u}) - f_i\right) \quad (18)$$

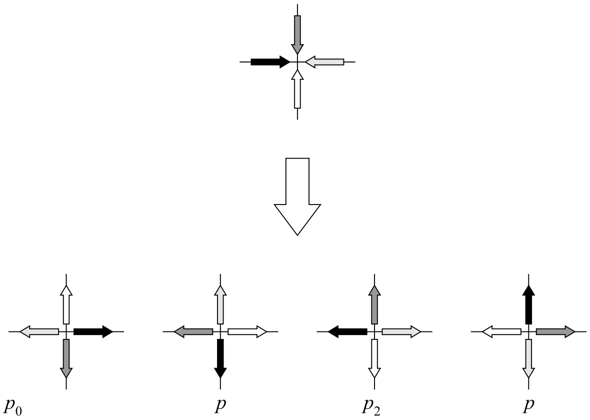
where f^{eq} is a given function, has been used extensively in the literature to simulate complex flows. The method is now recognized as a serious competitor of the more traditional approach based on the computer solution of the Navier–Stokes partial differential equation.

It is beyond the scope of this article to discuss the LBM approach in more detail. We refer the reader to several textbooks on this topic [8,10,40,41,49]. In addition to some rather technical aspects, one advantage of the LBM over the Navier–Stokes equation is its extended range of validity when the Knudsen number is not negligible (e. g. in microflows) [2].

Note finally that the LBM approach is also applicable to reaction-diffusion processes [1] or wave equation [10] by simply choosing an appropriate expression for f_i^{eq}.

Diffusion Processes

Diffusive phenomena and reaction processes play an important role in many areas of physics, chemistry, and biology and still constitute an active field of research. Systems in which reactive particles are brought into contact by a diffusion process and transform, often give rise to very complex behaviors. Pattern formation [31,34], is a typical example of such a behavior. CA provide an interesting framework to describe reaction-diffusion phenomena.



Cellular Automata Modeling of Physical Systems, Figure 20
How the entering particles are deflected at a typical site, as a result of the diffusion rule. The four possible outcomes occur with respective probabilities p_0, p_1, p_2, and p_3. The figure shows four particles, but the mechanism is data-blind and any one of the arrows can be removed when fewer entering particles are present

The HPP rule we discussed in the Sect. "A Simple Gas: The HPP Model" can be easily modified to produce many synchronous random walks. Random walk is well known to be the microscopic origin of a diffusion process.

Thus, instead of experiencing a mass and momentum conserving collision, each particle now selects, at random, a new direction of motion among the possible values permitted by the lattice. Since several particles may enter the same site (up to four, on a two-dimensional square lattice), the random change of directions should be such that there are never two or more particles exiting a site in the same direction. This would otherwise violate the exclusion principle.

The solution is to shuffle the directions of motion or, more precisely, to perform a random permutation of the velocity vectors, independently at each lattice site and each time step. Figure 20 illustrates this probabilistic evolution rule for a 2D square lattice.

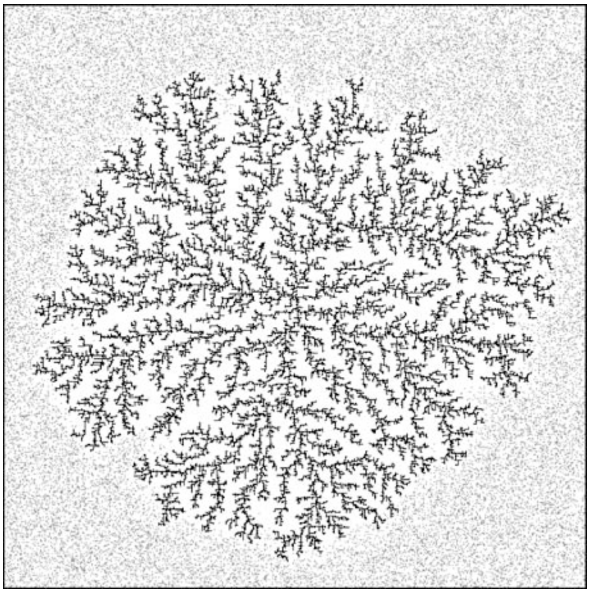
It can be shown that the quantity ρ defined as the average number of particle at site $\mathbf{r}$ and time t obeys the diffusion equation [10]

$$\partial_t \rho + \operatorname{div}\left[-D \operatorname{grad} \rho\right] = 0$$

where D is the diffusion constant whose expression is

$$D = \frac{\Delta_\mathrm{r}^2}{\Delta_t}\left(\frac{1}{4(p+p_2)} - \frac{1}{4}\right) = \frac{\Delta_\mathrm{r}^2}{\Delta_t}\left(\frac{p+p_0}{4[1-(p+p_0)]}\right) \tag{19}$$

where Δ_t and Δ_r are the time step and lattice spacing, respectively. For the one- and three-dimensional cases, a similar approach can be developed [10].



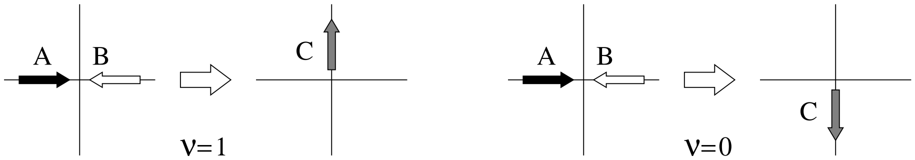
Cellular Automata Modeling of Physical Systems, Figure 21
Two-dimensional cellular automata DLA-like cluster (*black*), obtained with $p_\mathrm{s} = 1$, an aggregation threshold of 1 particle and a density of diffusing particle of 0.06 per lattice direction. The gray dots represent the diffusing particles not yet aggregated. The fractal dimension is found to be $d_\mathrm{f} = 1.78$

As an example of the use of the present random walk cellular automata rule, we discuss an application to growth processes. In many cases, growth is governed by an aggregation mechanism: like particles stick to each other as they meet and, as a result, form a complicated pattern with a branching structure.

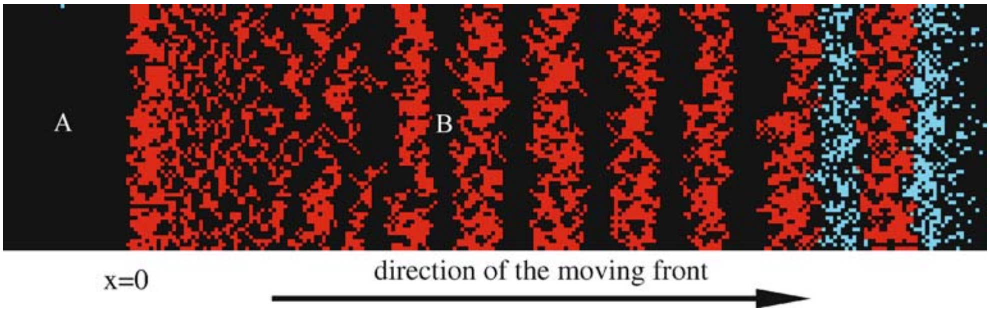
A prototype model of aggregation is the so-called DLA model (diffusion-limited aggregation), introduced by Witten and Sander [46] in the early 1980s. Since its introduction, the DLA model has been investigated in great detail. However, diffusion-limited aggregation is a far from equilibrium process which is not described theoretically by first principle only. Spatial fluctuations that are typical of the DLA growth are difficult to take into account and a numerical approach is necessary to complete the analysis.

DLA-like processes can be readily modeled by our diffusion cellular automata, provided that an appropriate rule is added to take into account the particle-particle aggregation. The first step is to introduce a rest particle to represent the particles of the aggregate. Therefore, in a two-dimensional system, a lattice site can be occupied by up to four diffusing particles, or by one "solid" particle.

Figure 21 shows a two-dimensional DLA-like cluster grown by the cellular automata dynamics. At the beginning of the simulation, one or more rest particles are intro-



Cellular Automata Modeling of Physical Systems, Figure 22
Automata implementation of the $A + B \rightarrow C$ reaction process. The reaction takes place with probability k. The Boolean quantity ν determines in which direction the C particle is moving after its creation



Cellular Automata Modeling of Physical Systems, Figure 23
Example of the formation of Liesegang bands in a cellular automata simulation. The *red bands* correspond to the precipitate which results from the $A + B$ reaction front (in *blue*)

duced in the system to act as aggregation seeds. The rest of the system is filled with particles with average concentration ρ. When a diffusing particle becomes nearest neighbor to a rest particle, it stops and sticks to it by transforming into a rest particle. Since several particles can enter the same site, we may choose to aggregate all of them at once (i. e. a rest particle is actually composed of several moving particles), or to accept the aggregation only when a single particle is present.

In addition to this question, the sticking condition is important. If any diffusing particle always sticks to the DLA cluster, the growth is very fast and can be influenced by the underlying lattice anisotropy. It is therefore more appropriate to stick with some probability p_s.

Reaction-Diffusion Processes

A reaction term can be added on top of the CA diffusion rule. For the sake of illustration let us consider a process such as

$$A + B \overset{K}{\rightarrow} C \tag{20}$$

where A, B, and C are different chemical species, all diffusing in the same solvent, and K is the reaction constant. To account for this reaction, one can consider the following mechanism: at the "microscopic" level of the discrete lattice dynamics, all the three species are first governed by a diffusion rule. When an A and a B particle enter the same site at the same time, they disappear and form a C particle.

Of course, there are several ways to select the events that will produce a C when more than one A or one B are simultaneously present at a given site. Also, when Cs already exist at this site, the exclusion principle may prevent the formation of new ones. A simple choice is to have A and B react only when they perform a head-on collision and when no Cs are present in the perpendicular directions. Figure 22 displays such a process.

Other rules can be considered if we want to enhance the reaction (make it more likely) or to deal with more complex situations ($2A + B \rightarrow C$, for instance).

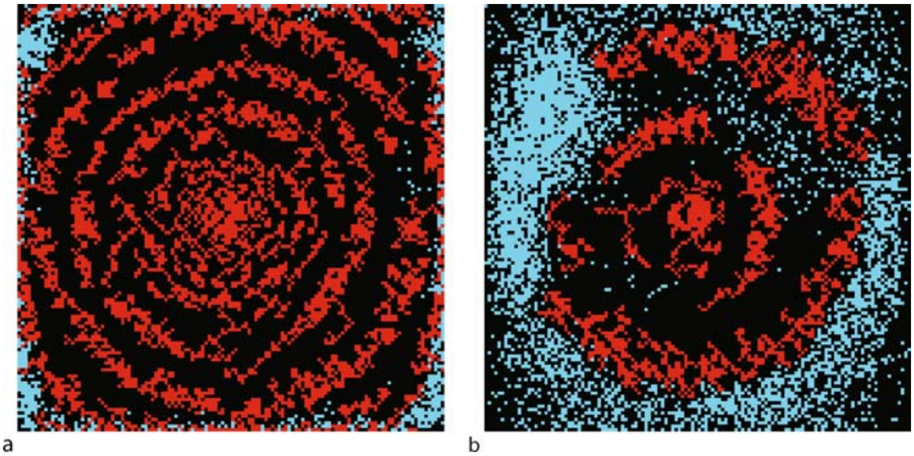
A parameter k can be introduced to tune the reaction rate K by controlling the probability of a reaction taking place. Using an appropriate mathematical tool [10], one can show that the idealized microscopic reaction-diffusion behavior implemented by the CA rule obeys the expected partial differential equation

$$\partial_t \rho_A = D\nabla^2 \rho_A - K\rho_A\rho_B \tag{21}$$

provided k is correctly chosen [10].

As an example of a CA reaction-diffusion model, we show in Fig. 23 the formation of the so-called Liesegang patterns [22].

Liesegang patterns are produced by precipitation and aggregation in the wake of a moving reaction front. Typ-



Cellular Automata Modeling of Physical Systems, Figure 24
Formation of Liesegang rings in a cellular automata simulation. The *red spots* correspond to the precipitate created by the $A + B$ reaction front (in *blue*)

ically, they are observed in a test tube containing a gel in which a chemical species B (for example $AgNO_3$) reacts with another species A (for example HCl). At the beginning of the experiment, B is uniformly distributed in the gel with concentration b_0. The other species A, with concentration a_0 is allowed to diffuse into the tube from its open extremity. Provided that the concentration a_0 is larger than b_0, a reaction front propagates in the tube. As this $A + B$ reaction goes on, formation of consecutive bands of precipitate (AgCl in our example) is observed in the tube, as shown in Fig. 23. Although this figure is from a computer simulation [12], it is very close to the picture of a real experiment.

Figure 24 shows the same process but in a different geometry. Species A is added in the middle of a 2D gel and diffuses radially. Rings (a) or spirals (b) result from the interplay between the reaction front and the solidification.

Excitable Media

Excitable media are other examples of reaction processes where unexpected space-time patterns are created. As opposed to the reaction-diffusion models discussed above, diffusion is not considered here explicitly. It is assumed that reaction occurs between nearest neighbor cells, making the transport of species unnecessary. The main focus is on the description of chemical waves propagating in the system much faster and differently than any diffusion process would produce.

An excitable medium is basically characterized by three states [5]: the resting state, the excited state, and the refractory state. The resting state is a stable state of the system. But a resting state can respond to a local perturbation and become excited. Then, the excited state evolves to a refractory state where it no longer influences its neighbors and, finally, returns to the resting state.

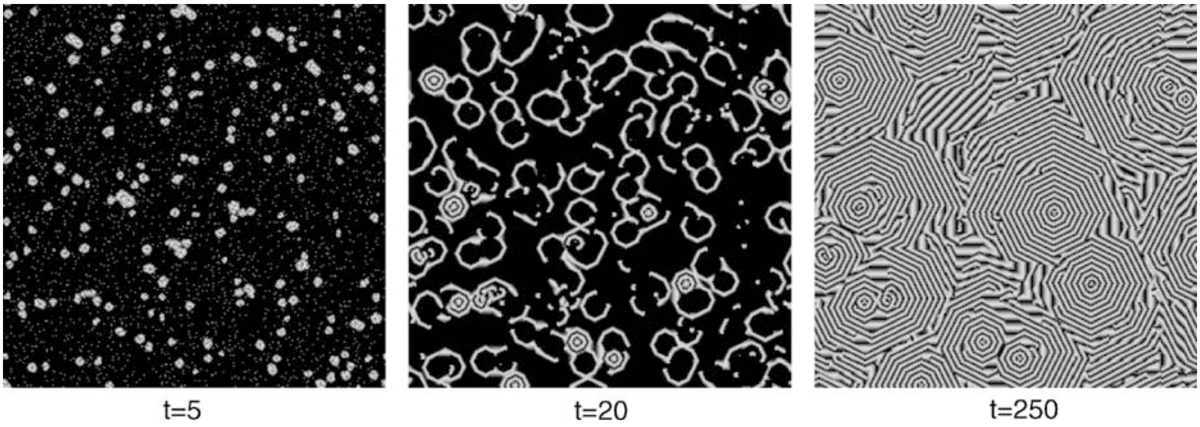
A generic behavior of excitable media is to produce chemical waves of various geometries [26,27]. Ring and spiral waves are a typical pattern of excitations. Many chemical systems exhibit an excitable behavior. The Selkov model [38] and the Belousov–Zhabotinsky reaction are examples. Chemical waves play an important role in many biological processes (nervous systems, muscles) since they can mediate the transport of information from one place to another.

The Greenberg–Hasting model is an example of a cellular automata model of an excitable media. This rule, and its generalization, have been extensively studied [17,20].

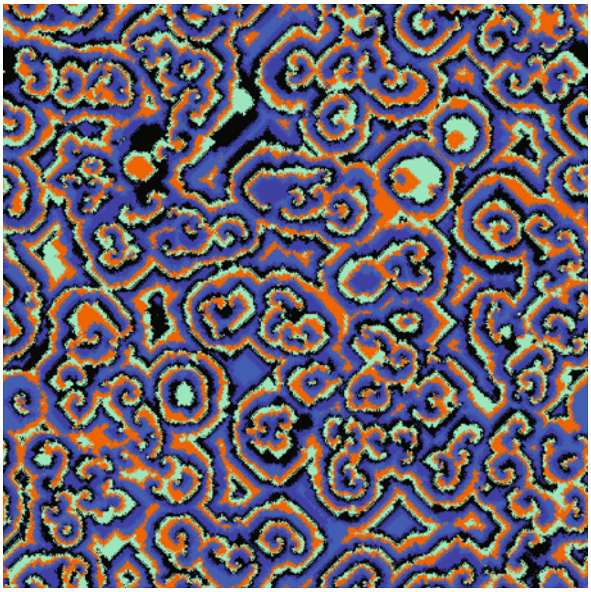
The implementation we propose here for the Greenberg–Hasting model is the following: the state $\psi(\mathbf{r}, t)$ of site $\mathbf{r}$ at time t takes its value in the set $\{0, 1, 2, \dots, n-1\}$. The state $\psi = 0$ is the resting state. The states $\psi = 1, \dots, n/2$ (n is assumed to be even) correspond to excited states. The rest, $\psi = n/2 + 1, \dots, n-1$ are the refractory states. The cellular automata evolution rule is the following:

1. If $\psi(\mathbf{r}, t)$ is excited or refractory, then $\psi(\mathbf{r}, t + 1) = \psi(\mathbf{r}, t) + 1 \mod n$.
2. If $\psi(\mathbf{r}, t) = 0$ (resting state) it remains so, unless there are at least k excited sites in the Moore neighborhood of site $\mathbf{r}$. In this case $\Psi(\mathbf{r}, t + 1) = 1$.

The n states play the role of a clock: an excited state evolves



Cellular Automata Modeling of Physical Systems, Figure 25
Excitable medium: evolution of a configuration with 5% of excited states $\Psi = 1$, and 95% of resting states (*black*), for $n = 8$ and $k = 3$



Cellular Automata Modeling of Physical Systems, Figure 26
The tube-worms rule for an excitable media

through the sequence of all possible states until it returns to 0, which corresponds to a stable situation.

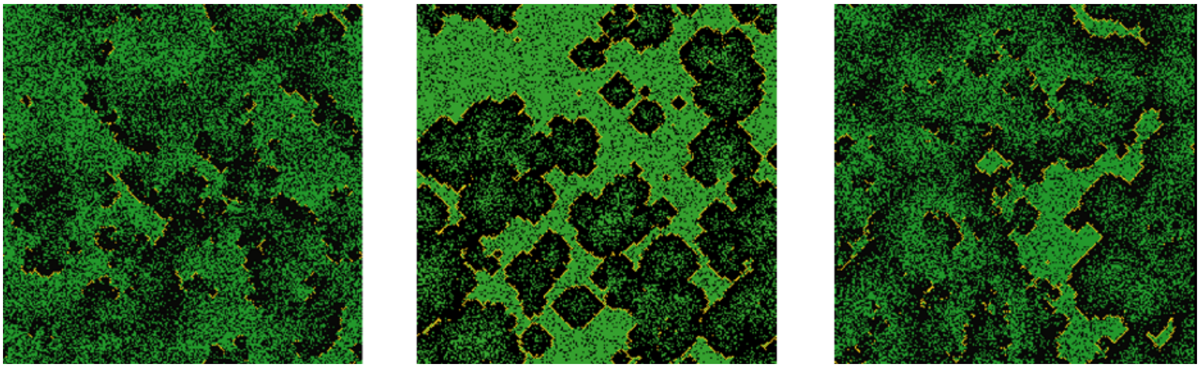
The behavior of this rule is quite sensitive to the value of *n* and the excitation threshold *k*. Figure 25 shows the evolution of this automaton for a given set of the parameters *n* and *k*. The simulation is started with a uniform configuration of resting states, perturbed by some excited sites randomly distributed over the system. Note that if the concentration of perturbation is low enough, excitation dies out rapidly and the system returns to the rest state. Increasing the number of perturbed states leads to the formation of traveling waves and self-sustained oscillations may appear in the form of ring or spiral waves.

The Greenberg–Hasting model has some similarity with the "tube-worms" rule proposed by Toffoli and Margolus [42]. This rule is intended to model the Belousov–Zhabotinsky reaction and is as follows. The state of each site is either 0 (refractory) or 1 (excited) and a local timer (whose value is 3, 2, 1, or 0) controls the refractory period. Each iteration of the rule can be expressed by the following sequence of operations: (i) where the timer is zero, the state is excited; (ii) the timer is decreased by 1 unless it is 0; (iii) a site becomes refractory whenever the timer is equal to 2; (iv) the timer is reset to 3 for the excited sites which have two, or more than four, excited sites in their Moore neighborhood.

Figure 26 shows a simulation of this automaton, starting from a random initial configuration of the timers and the excited states. We observe the formation of spiral pairs of excitations. Note that this rule is very sensitive to small modifications (in particular to the order of operations (i) to (iv)).

Another rule which is also similar to Greenberg–Hasting and Margolus–Toffoli tube-worm models is the so-called forest-fire model. This rule describes the propagation of a fire or, in a different context, may also be used to mimic contagion in the case of an epidemic. Here we describe the case of a forest-fire rule.

The forest-fire rule is a probabilistic CA defined on a *d*-dimensional cubic lattice. Initially, each site is occupied by either a tree, a burning tree, or is empty. The state of the system is parallel updated according to the following rule: (1) a burning tree becomes an empty site; (2) a green



Cellular Automata Modeling of Physical Systems, Figure 27

The forest fire rule: *green sites* correspond to a grown tree, *black pixels* represent burned sites and the *yellow color* indicates a burning tree. The snapshots given here represent three situations after a few hundred iterations. The parameters of the rule are $p = 0.3$ and $f = 6 \times 10^{-5}$

tree becomes a burning tree if at least one of its nearest neighbors is burning; (3) at an empty site, a tree grows with probability p; (4) A tree without a burning neighbor becomes a burning tree with probability f (so as to mimic an effect of lightning).

Figure 27 illustrates the behavior of this rule, in a two-dimensional situation. Provided that the time scales of tree growth and burning down of forest clusters are well separated (i. e. in the limit $f/p \to 0$), this model has self-organized critical states [15]. This means that in the steady state, several physical quantities characterizing the system have a power law behavior.

Surface Reaction Models

Other important reaction models that can be described by a CA are surface-reaction models where nonequilibrium phase transitions can be observed. Nonequilibrium phase transitions are an important topic in physics because no general theory is available to describe such systems and most of the known results are based on numerical simulations.

The so-called Ziff model [54] gives an example of the reaction A–B_2 on a catalyst surface (for example CO–O$_2$).

The system is out of equilibrium because it is an open system in which material continuously flows in and out. However, after a while, it reaches a stationary state and, depending on some control parameters, may be in different phases.

The basic steps are

- A gas mixture with concentrations X_{B_2} of B_2 and X_A of A sits above a surface on which it can be adsorbed. The surface is divided into elementary cells and each cell can adsorb one atom only.
- The B species can be adsorbed only in the atomic form. A molecule B_2 dissociates into two B atoms only if two adjacent cells are empty. Otherwise the B_2 molecule is rejected.
- If two nearest neighbor cells are occupied by different species they chemically react and the product of the reaction is desorbed. In the example of the CO–O$_2$ reaction, the desorbed product is a CO$_2$ molecule.

This final desorption step is necessary for the product to be recovered and for the catalyst to be regenerated. However, the gas above the surface is assumed to be continually replenished by fresh material so that its composition remains constant during the whole evolution.

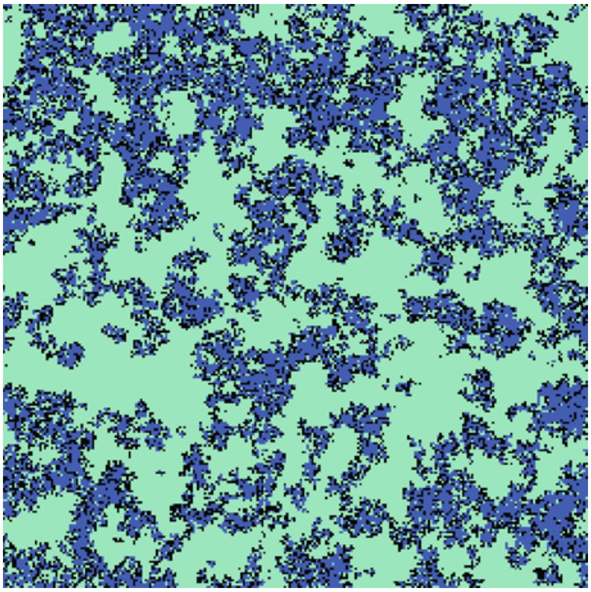
It is found by *sequential* numerical simulation [54] that a reactive steady state occurs only in a window defined by

$$X_1 < X_A < X_2$$

where $X_1 = 0.389 \pm 0.005$ and $X_2 = 0.525 \pm 0.001$ (provided that $X_{B_2} = 1 - X_A$). This situation is illustrated in Fig. 28, though for the corresponding cellular automata dynamics and $X_{B_2} \neq 1 - X_A$.

Outside this window of parameter, the steady state is a "poisoned" catalyst of pure A (when $X_A > X_2$) or pure B (when $X_A < X_1$). For $X_1 < X_A < X_2$, the coverage fraction varies continuously with X_A and one speaks of a continuous (or second-order) nonequilibrium phase transition. At $X_A = X_2$, the coverage fraction varies discontinuously with X_A and one speaks of a discontinuous (or first-order) nonequilibrium phase transition. Figure 30 displays this behavior.

The asymmetry of behavior at X_1 and X_2 comes from the fact that A and B atoms have a different adsorption rule: two vacant adjacent sites are necessary for B to stick on the surface, whereas one empty site is enough for A.



Cellular Automata Modeling of Physical Systems, Figure 28
Typical microscopic configuration in the stationary state of the CA Ziff model, where there is coexistence of the two species over time. The simulation corresponds to the generalized model described by rules R1, R2, R3, and R4 below. The *blue* and *green dots* represent, respectively, the *A* and *B* particles, whereas the empty sites are *black*

In a CA approach the elementary cells of the catalyst are mapped onto the cells of the automaton. In order to model the different processes, each cell j can be in one of four different states, denoted $|\psi_j\rangle = |0\rangle, |A\rangle, |B\rangle$ or $|C\rangle$.

The state $|0\rangle$ corresponds to an empty cell, $|A\rangle$ to a cell occupied by an atom A, and $|B\rangle$ to a cell occupied by an atom B. The state $|C\rangle$ is artificial and represents a precursor state describing the conditional occupation of the cell by an atom B. Conditional means that during the next evolution step of the automaton, $|C\rangle$ will become $|B\rangle$ or $|0\rangle$ depending upon the fact that a nearest neighbor cell is empty and ready to receive the second B atom of the molecule B_2. This conditional state is necessary to describe the dissociation of B_2 molecules on the surface.

The main difficulty when implementing the Ziff model with a fully synchronous updating scheme is to ensure that the correct stoichiometry is obeyed. Indeed, since all atoms take a decision at the same time, the same atom could well take part in a reaction with several different neighbors, unless some care is taken.

The solution to this problem is to add a vector field to every site in the lattice [53], as shown in Fig. 29. A vector field is a collection of arrows, one at each lattice site, that can point in any of the four directions of the lattice. The directions of the arrows at each time step are assigned randomly. Thus, a two-site process is carried out only on those pairs of sites in which the arrows point toward each other (matching nearest-neighbor pairs (MNN)). This concept of *reacting matching pairs* is a general way to partition the parallel computation in local parts.

In the present implementation, the following generalization of the dynamics is included: an empty site remains empty with some probability. One has then two control parameters to play with: X_A and X_{B_2} that are the arrival probability of an A and a B_2 molecule, respectively.

Thus, the time evolution of the CA is given by the following set of rules, fixing the state of the cell j at time $t+1$, $|\psi_j\rangle(t+1)$, as a function of the state of the cell j and its nearest neighbors (von Neumann neighborhood) at time t. Rules R1, R4 describe the adsorption–dissociation mechanism while rules R2, R3 (illustrated in Fig. 29) describe the reaction–desorption process.

<u>R1</u>: If $|\psi_j\rangle(t) = |0\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |A\rangle & \text{with probability } X_A \\ |C\rangle & \text{with probability } X_{B_2} \\ |0\rangle & \text{with probability } 1 - X_A - X_{B_2} \end{cases} \tag{22}$$

<u>R2</u>: If $|\psi_j\rangle(t) = |A\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |0\rangle & \text{if the MNN of } j \\ & \text{was in the state } |B\rangle \text{ at time } t \\ |A\rangle & \text{otherwise} \end{cases} \tag{23}$$

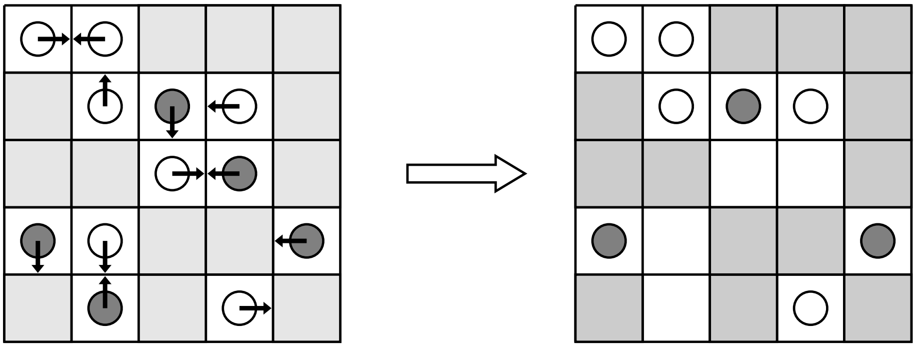
<u>R3</u>: If $|\psi_j\rangle(t) = |B\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |0\rangle & \text{if the MNN of } j \\ & \text{was in the state } |A\rangle \text{ at time } t \\ |B\rangle & \text{otherwise} \end{cases} \tag{24}$$

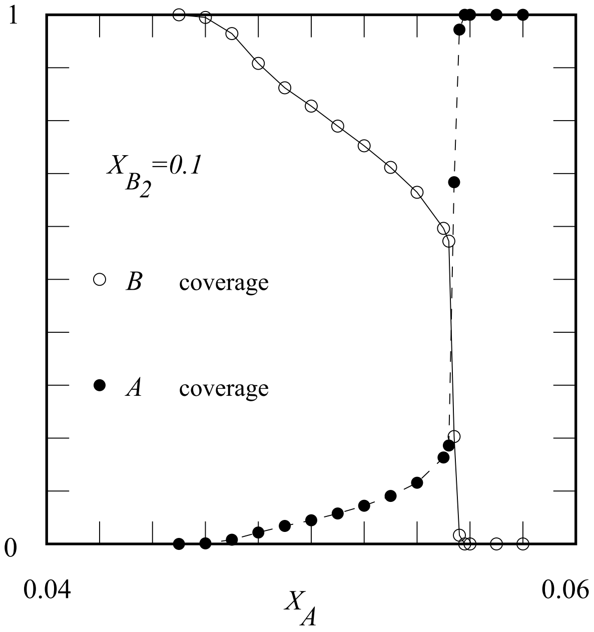
<u>R4</u>: If $|\psi_j\rangle(t) = |C\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |B\rangle & \text{if MNN is in the state } |C\rangle \\ & \text{at time } t \\ |0\rangle & \text{otherwise} \end{cases} \tag{25}$$

Figure 28 shows typical stationary configurations obtained with a cellular automata version of the Ziff model. At time $t = 0$, all the cells are empty and a randomly prepared mixture of gases with fixed concentrations X_A and X_{B_2} sits on top of the surface. The rules are iterated until



Cellular Automata Modeling of Physical Systems, Figure 29
Illustration of rules R2 and R3. The *arrows* select which neighbor is considered for a reaction. *Dark* and *white particles* represent the *A* and *B* species, respectively. The *shaded region* corresponds to cells that are not relevant to the present discussion such as, for instance, cells occupied by the intermediate *C* species



Cellular Automata Modeling of Physical Systems, Figure 30
Stationary state phase diagram corresponding to the CA Ziff model

a stationary state is reached. The stationary state is a state for which the mean coverage fractions X_A^a and X_B^a of atoms of type A or B does not change in time, although microscopically the configurations of the surface changes.

The phase diagram obtained for this generalized CA Ziff model is given in Fig. 30, with the value $X_{B_2} = 0.1$. This phase diagram is topologically similar to the sequential updating case (with $X_{B_2} = 1 - X_A$) since we observe a first and a second order transition surrounding a region of coexistence of both species. However, the locations of the critical points are different, illustrating the nonuniversal character of these quantities.

Future Directions

Cellular Automata are clearly an active field of research. CA are easy to implement on personal computers or on large-scale parallel machines. They are very well suited to discuss many complex systems and dynamical phenomena with space inhomogeneities. Thus, CA provide a very interesting framework for modeling and simulating various processes, for a wide range of application domains where space and time are playing a central role. One can expect that such developments will continue in the future.

In addition, to provide numerical predictions, CA also offer a very intuitive and efficient language to describe various processes or systems. In a period where interdisciplinary applications are very important, it is crucial to be able to communicate between different scientific communities and to develop models that incorporate the knowledge of different fields. Partial differential equations are often an obstacle for several researchers outside Mathematics or Physics. CA then provide a different tool to describe new systems, for instance in ecology, social behavior or in many fields for which a standard mathematical framework is still missing.

The concept of CA can certainly be extended to correspond better to the future scientific challenges. The Lattice Boltzmann method is a well-known example where the Boolean nature of the CA state has been removed. As a result, powerful flow solvers have been obtained.

The nature of the cellular space can also be extended. Graphs are more and more studied for their ability to describe many organizations, whether social, economical, or biological. Therefore, a natural evolution of CA is to be able to include irregular and/or dynamic lattices of cells, so as to capture the behaviors that come from specificities of the space topology.

Another important direction in computational science are the so-called multiscale and multiscience simulations, aiming at coupling several processes spanning very different spatial or temporal scales. Such a combination of processes is for instance quite common in the simulation of biomedical problems. Recently, the concept of Complex Automaton [23] was proposed as a way to integrate in one single simulation many different components, each described by a different CA.

Bibliography

Primary Literature

1. Alemani D, Chopard B, Buffle J, Galceran J (2006) Two grid refinement methods in the Lattice Boltzmann framework for reaction-diffusion processes. Phys Chem Chem Phys 8:35
2. Ansumali S, Karlin I, Arcidiacono S, Abbas A, Prasianakis N (2007) Hydrodynamics beyond Navier–Stokes: Exact solution to the lattice boltzmann hierarchy. Phys Rev Lett 98:124502
3. Banks E (1971) Information processing and transmission in cellular automata. Tech rep., MIT, MAC TR-81
4. Boon JP (ed) (1992) Advanced Research Workshop on Lattice Gas Automata Theory, Implementations, and Simulation. J Stat Phys 68(3/4):347–672
5. Boon JP, Dab D, Kapral R, Lawniczak A (1996) Lattice gas automata for reactive systems. Phys Rep 273:55–148
6. Burks A (1970) Von Neumann's self-reproducing automata. In: Burks A (ed) Essays on Cellular Automata. University of Illinois Press, Chicago, pp 3–64
7. Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. Physica A 295:506–525
8. Chen S, Doolen G (1998) Lattice Boltzmann methods for fluid flows. Annu Rev Fluid Mech 30:329
9. Chopard B, Droz M (1987) Cellular automata approach to non equilibrium phase transitions in a surface reaction model: static and dynamic properties. J Phys A 21:205
10. Chopard B, Droz M (1998) Cellular Automata Modeling of Physical Systems. Cambridge University Press, Cambridge
11. Chopard B, Dupuis A (2003) Cellular automata simulations of traffic: a model for the city of geneva. Netw Spat Econ 3:9–21
12. Chopard B, Luthi P, Droz M (1994) Reaction-diffusion cellular automata model for the formation of Liesegang patterns. Phys Rev Lett 72(9):1384–1387
13. Chopard B, Luthi PO, Queloz PA (1996) Cellular automata model of car traffic in two-dimensional street networks. J Phys A 29:2325–2336
14. Doolen G (ed) (1990) Lattice Gas Method for Partial Differential Equations. Addison-Wesley, Redwood City
15. Drossel B, Schwabl F (1992) Self-organized critical forest-fire model. Phys Rev Lett 69:1629
16. Farmer D, Toffoli T, Wolfram S (eds) (1984) Cellular Automata. Proceedings of an Interdisciplinary Workshop, Los Alamos. Physica D, vol 10. North-Holland, Amsterdam
17. Fisch R, Gravner J, Griffeath D (1991) Threshold-range scaling of excitable cellular automata. Stat Comput 1:23
18. Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the navier–stokes equation. Phys Rev Lett 56:1505
19. Gardner M (1970) The fantastic combinations of john conway's new solitaire game life. Sci Am 220(4):120
20. Gravner J, Griffeath D (1993) Threshold grouse dynamics. Trans Amer Math Soc 340:837
21. Gunton J, Droz M (1983) Introduction to the Theory of Metastable and Unstable States. Springer, Berlin
22. Henisch H K (1988) Crystals in Gels and Liesegang Rings. Cambridge University Press, Cambridge
23. Hoekstra A, Lorenz E, Falcone JL, Chopard B (2007) Towards a complex automata framework for multi-scale modeling: Formalism and the scale separation map. In: Shi Y et al (ed) Computational Sciences ICCS 2007. LNCS, vol 4487. Springer, Berlin, pp 922–939
24. Kanai M, Nishinari K, Tokihiro T (2005) Stochastic optimal velocity model and its long-lived metastability. Phys Rev E 72:035102(R)
25. Kanai M, Nishinari K, Tokihiro T (2006) Stochastic cellular automaton model for traffic flow. In: Yacoubi SE, Chopard B, Bandini S (eds) Cellular Automata: 7th ACRI conference. LNCS, vol 4173. Springer, Berlin, pp 538–547
26. Kapral R, Showalter K (eds) (1995) Chemical Waves and Patterns. Kluwer, Dordrecht
27. Keener J, Tyson J (1992) The dynamics of scroll waves in excitable media. SIAM Rev. 34:1–39
28. Luthi P O, Preiss A, Ramsden JJ, Chopard B (1998) A cellular automaton model for neurogenesis in drosophila. Physica D 118:151–160
29. Marconi S, Chopard B (2002) A multiparticle lattice gas automata for a crowd. In: Bardini S et al (ed) Proceedings of ACRI 2002 Geneva, Oct, 2002. Lecture notes in computer science, vol 2493. Springer, Berlin, p 230
30. Marconi S, Chopard B (2006) Discrete physics, cellular automata and cryptography. In: Yacoubi SE, Chopard B, Bandini S (eds) Cellular Automata: 7th ACRI conference. LNCS, vol 4173. Springer, Berlin, pp 617–626
31. Muray J (1990) Mathematical Biology. Springer, Berlin
32. Nagel K, Herrmann H (1993) Deterministic models for traffic jams. Physica A 199:254
33. Nagel K, Schreckenberg M (1992) Cellular automaton model for freeway traffic. J Physique I 2:2221
34. Pearson JE (1993) Complex patterns in a simple system. Science 261:189–192
35. Rothman D, Zaleski S (1997) Lattice-Gas Cellular Automata: Simple Models of Complex Hydrodynamics. Collection Aléa. Cambridge University Press, Cambridge
36. Schadschneider A, Schreckenberg M (1993) Cellular automaton models and traffic flow. J Phys A 26:L679
37. Schreckenberg M, Schadschneider A, Nagel K, Ito N (1995) Discrete stochastic models for traffic flow. Phys. Rev. E 51:2939
38. Selkov E (1968) Self-oscillation in glycolysis: A simple kinetic model. Eur J Biochem 4:79
39. Shannon C (1949) Communication theory of secrecy systems. Bell Syst Tech J 28:656–715
40. Succi S (2001) The Lattice Boltzmann Equation, For Fluid Dynamics and Beyond. Oxford University Press, Oxford
41. Sukop M, Thorne D (2005) Lattice Boltzmann Modeling: an Introduction for Geoscientists and Engineers. Springer, Berlin

42. Toffoli T, Margolus N (1987) Cellular Automata Machines: a New Environment for Modeling. MIT Press, Cambridge
43. Tolman S, Meakin P (1989) Off-lattice and hypercubic-lattice models for diffusion-limited aggregation in dimension 2–8. Phys Rev A 40:428–37
44. Vichniac G (1984) Simulating physics with cellular automata. Physica D 10:96–115
45. Vicsek T (1989) Fractal Growth Phenomena. World Scientific, Singapore
46. Witten T, Sander L (1983) Diffusion-limited aggregation. Phys Rev B 27:5686
47. Wolf D, Schreckenberg M et al (eds) Traffic and Granular Flow '97. Springer, Singapore
48. Wolf D, Schreckenberg M, Bachem A (eds) (1996) Traffic and Granular Flow. World Scientific, Singapore
49. Wolf-Gladrow D A (2000) Lattice-Gas Cellular Automata and Lattice Boltzmann Models: an Introduction. Lecture Notes in Mathematics, vol 1725. Springer, Berlin
50. Wolfram S (1986) Theory and Application of Cellular Automata. World Scientific, Singapore
51. Wolfram S (1994) Cellular Automata and Complexity. Addison-Wesley, Reading
52. Yukawa S, Kikuchi M, Tadaki S (1994) Dynamical phase transition in one-dimensional traffic flow model with blockage. J Phys Soc Jpn 63(10):3609–3618
53. Ziff R, Fichthorn K, Gulari E (1991) Cellular automaton version of the ab_2 reaction model obeying proper stoichiometry. J Phys. A 24:3727
54. Ziff R, Gulari E, Barshad Y (1986) Kinetic phase transitions in an irreversible surface-reaction model. Phys Rev Lett 56:2553

Books and Reviews

Chopard B, Droz M (1998) Cellular AutomataModeling of Physical Systems. Cambridge University Press, Cambridge

Deutsch A, Dormann S (2005) Cellular Automaton Modeling of Biological Pattern Formation. Birkhäuser, Basel

Gaylord RJ, Nishidate K (1996) Modeling Nature with Cellular Automata using Mathematica. Springer, Berlin

Ilachinski A (2001) Cellular Automata: a discrete universe. World Scientific, Singapore

Rivet JP, Boon JP (2001) Lattice Gas Hydrodynamics. Cambridge University Press, Cambridge

Rothman D, Zaleski S (1994) Lattice-gas models of phase separation: interface, phase transition and multiphase flows. Rev Mod Phys 66:1417–1479

Weimar JR (1998) Simulation with Cellular Automata. Logos, Berlin

Wolfram S (2002) A new kind of science. Wolfram Sciences, Champaign

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations

Carter Bays
Department of Computer Science and Engineering,
University of South Carolina, Columbia, USA

Article Outline

Typically, cellular automata ("CA") are defined in Cartesian space (e. g. a square grid). Here we explore characteristics of CA in triangular and other non-cartesian grids. Methods for programming CA for these non-cartesian grids are briefly discussed.

Glossary

Cellular automaton (CA) a structure comprising a grid with individual cells that can have two or more states; these cells evolve in discrete time units and are governed by a rule, which usually involves neighbors of each cell.

Game of life a particular cellular automaton discovered by John Conway in 1968.

Neighbor a neighbor of cell "x" is typically a cell that is in close proximity to (frequently touching) cell "x".

Oscillator a periodic shape within a specific cellular automaton rule.

Glider a translating oscillator that moves across the grid of a CA.

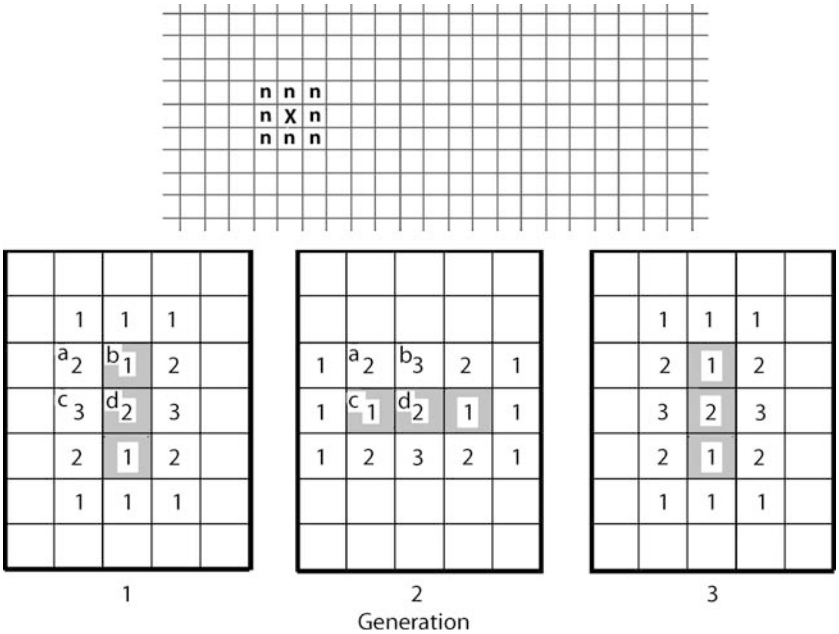
Generation the discrete time unit which depicts the evolution of a cellular automaton.

Rule determines how each individual cell within a cellular automaton evolves.

Definition of the Subject

A tessellation or tiling is composed of a specific shape that is repeated endlessly in a plane, with no gaps or overlaps. Examples of simple tessellations are the square grid, the triangular grid (a plane completely covered by identical triangles), etc. Hereafter, we shall also use "grid" when refering to tessellations.

Cellular automata (CA) can be explained most effectively with an example. Start with an infinite grid of squares; each square represents a cell, which is either "alive" or "dead". Time progresses in discrete units called "generations"; at every generation we evaluate simultane-



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 1
Top: Each cell in a grid has 8 "neighbors". The cells containing "n" are neighbors of the cell containing the "X". Any cell in the grid can be either "dead" or "alive". *Bottom*: Here we have outlined a specific area of what is presumably a much larger grid. At the *left* we have installed an initial shape. *Shaded* cells are alive; all others are dead. The number within each cell gives the quantity of live neighbors for that cell. (Cells containing no numbers have zero live neighbors.) Depicted are three generations, starting with the configuration at generation 1. Generations 2 then 3 show the result when we apply the following cellular automata rule: "Live cells with exactly 2 or 3 live neighbors remain alive (otherwise they die); dead cells with exactly 3 live neighbors come to life (otherwise they remain dead)". Let us now evaluate the transition from generation 1 to generation 2. In our diagram, cell "a" is dead. Since it does not have exactly 3 live neighbors, it remains dead. Cell "b" is alive, but it needs exactly 2 or 3 live neighbors to remain alive; since it only has 1, it dies. Cell "c" is dead; since it has exactly 3 live neighbors, it comes to life. And cell "d" has 2 live neighbors; hence it will remain alive. And so on. Notice that the form repeats every two generations. Such forms are called oscillators

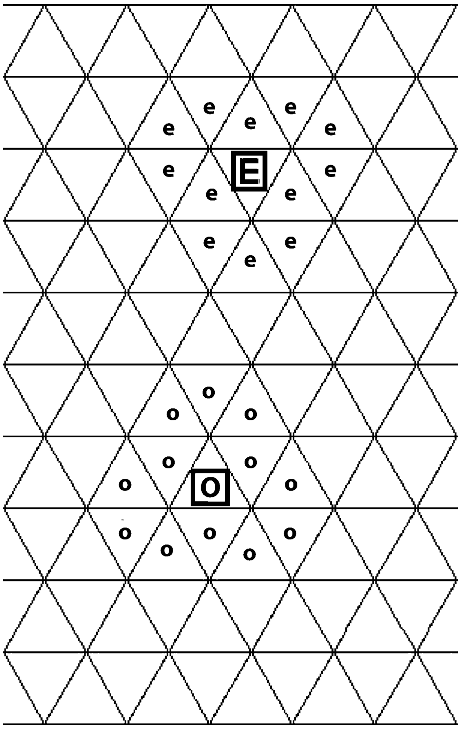
ously the fate for each cell at the next generation by examining neighboring cells (called "neighbors") – in this case, we shall consider as neighbors any cell touching the candidate cell (eight neighbors in all). This is sometimes called the Moore neighborhood. We apply a "rule" to determine the next generation status of our candidate cell. For example, our rule might state, (a) "If our candidate cell is currently alive, then it will remain alive next generation if it touches either two or three live neighbors, otherwise it dies", and (b) "If our candidate cell is not alive then it will come to life next generation if and only if it is touching exactly three live neighbors."

Figure 1 illustrates a simple configuration to which this CA rule has been applied. Notice that this particular object repeats itself indefinitely. Such an object is called an "oscillator"; this particular oscillator has a "period" of two. Other configurations can have much larger periods, or can behave in a more chaotic fashion. Motionless patterns can be thought of as oscillators whose period is one.

Needless to say, there are a huge number of rules that can be applied, and each rule will cause a distinct action. The rule given above – the most famous cellular automaton of all – specifies the "Game of Life", discovered by John Horton Conway in 1968. Game of Life (GL) rules must satisfy the following informal criteria.

1. All neighbors must be touching the candidate cell and all are treated the same.
2. There must exist at least one translating oscillator (called a "glider").
3. Random configurations must eventually stabilize into zero or more oscillators..

For a more formal description of GL rule requirements see [1]. It is important to note that CA can be represented in one, two, three or higher dimensions, but most work has been done in one or two. Furthermore, neighbors can be defined in many ways; for example we might only consider as neighbors those cells touching the sides of



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 2

The neighborhoods for cells in the triangular grid. Note that the candidate cells can have two orientations – "E" and "O". The neighbors are indicated by "e" and "o" respectively

a candidate cell and not the corners. Or we might expand our neighborhood to include cells within a given distance of a candidate cell. This is typically done for one-dimensional CA.

Some Convenient Notation for Describing CA Rules

We shall write CA rules using the following notation,

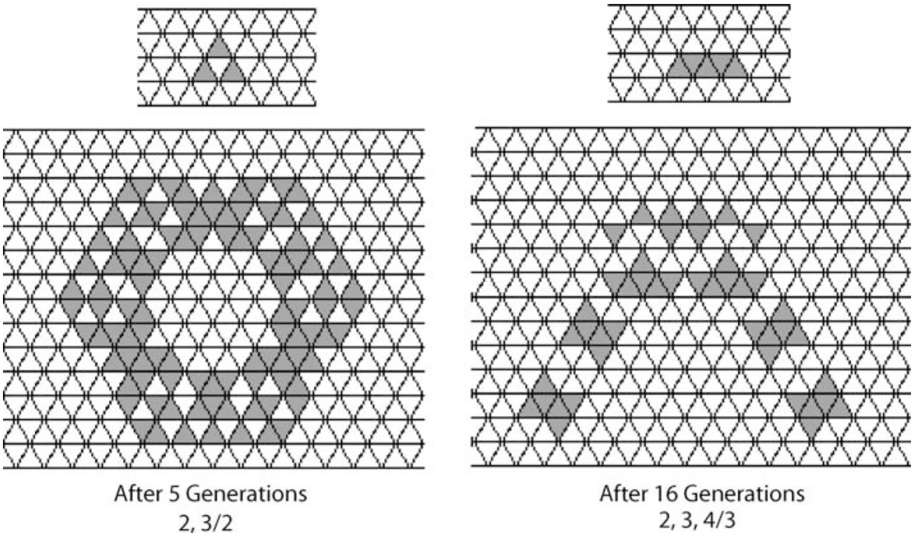
$$E_1, E_2, \ldots / F_1, F_2, \ldots$$

where the E_i specify the number of live neighbors required to keep a living cell alive, and the F_i give the number required to bring a non-living cell to life. The E_i and F_i will be listed in ascending order; hence if $i > j$ then $E_i > E_j$ etc. Thus the rule for Conway's Game of Life is written 2,3/3.

We shall also use a convenient shorthand when appropriate: E_i–E_j denotes $E_i, E_{i+1}, \ldots, E_{i+j-i}$ etc. Thus, 2,3,4,5,6/2,3,4 can also be written 2–6/2–4.

Introduction

Almost all CA research in two dimensions has been done using rectangular (Cartesian) coordinates, and hence typically utilizes the square grid. But there is no reason to limit ourselves to this tessellation; the number of different possible grids is almost endless. Here we shall briefly investigate CA behavior in only three – triangular, hexagonal and pentagonal.



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 3

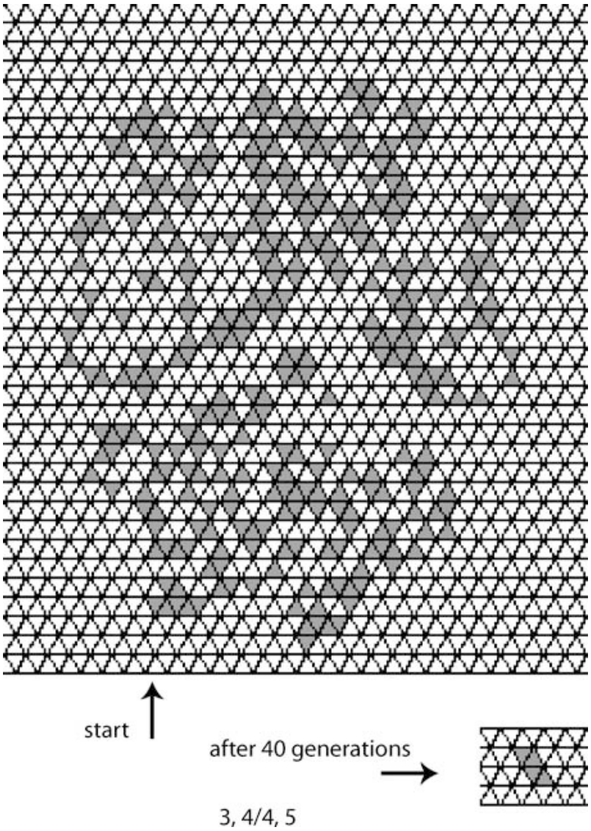
Examples of expanding rules. The starting configurations are at the top

Two Dimensional Cellular Automata in the Triangular Grid

Throughout this article we shall consider as neighbors only those cells that touch the candidate cell; hence for a grid composed of triangles, each cell would have 12 neighbors (Fig. 2). These non-cartesian grids for CA have been investigated from time to time; most notably by Preston [3] and Bays [1,2]. Recently work relating to hexagonal CA has appeared on the internet occasionally.

With 12 touching neighbors instead of 8 (as in the square grid), we can write more than 16 million distinct rules, most of which are probably of only marginal interest. Many however exhibit behavior worthy of investigation.

Some rules will generate a continually expanding collection of live cells – we shall call such rules "expanding" or "unstable" rules. Thus 2,3/2; 2,3/3,4; 2,3,4/3 each produce an ever increasing area of live cells – even with extremely small starting configurations (Fig. 3). A few expanding rules "barely" expand; i. e. several generations are required and the initial live configuration must be fairly large in order to observe instability. For example 2,3,6/4,5 can produce unbounded growth, while 2,3,8/4,5 always eventually stabilizes. The fate of configurations under 2,3,7/4,5 is uncertain, but the rule appears to produce unbounded growth. Many rules will ultimately lead to a stable pattern (Fig. 4), or no live cells at all.



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 4
An example of a stable rule. The starting random configuration eventually stabilizes into the shape shown at the *lower right*; interestingly this shape happens to be an oscillator with a period of two

For some rules we can start with bounded forms whose innards churn endlessly forever; these rules can, for example, be used to generate random numbers (Fig. 5). Such rules differ somewhat from expanding rules in that all finite patterns are bounded and will not expand indefinitely, but an infinite grid of random live cells will never stabilize.

Game of Life Rules in the Triangular Grid

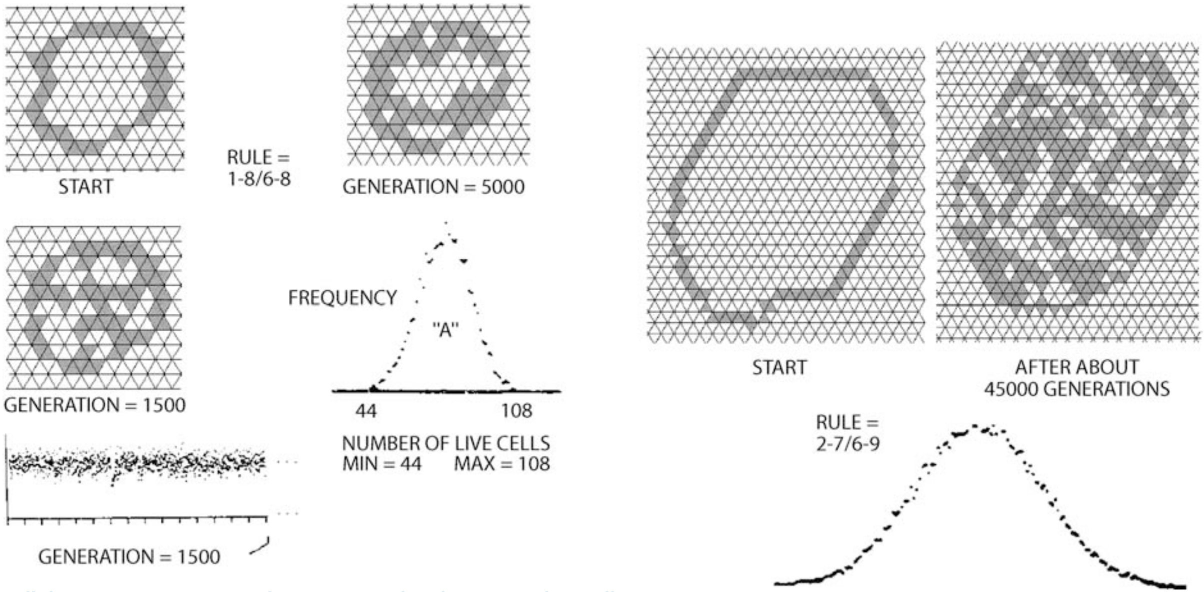
As mentioned above, the most famous GL rule is Conway's game, which utilizes a square grid. But GL rules are not limited to squares; quite a few exist in the triangular grid. Among these are 4,5,6/4; 3,4/4,5; 4,5/4,5,6; 2,3/4,5; 3,4/4,5,6; 2/3; 2,4/4,6; 3,5/4; 2,4,6/4,6; 2,7/3; 2,7,8/3. Further information about these and other rules can be found in [1] and [2].

The Hexagonal Grid

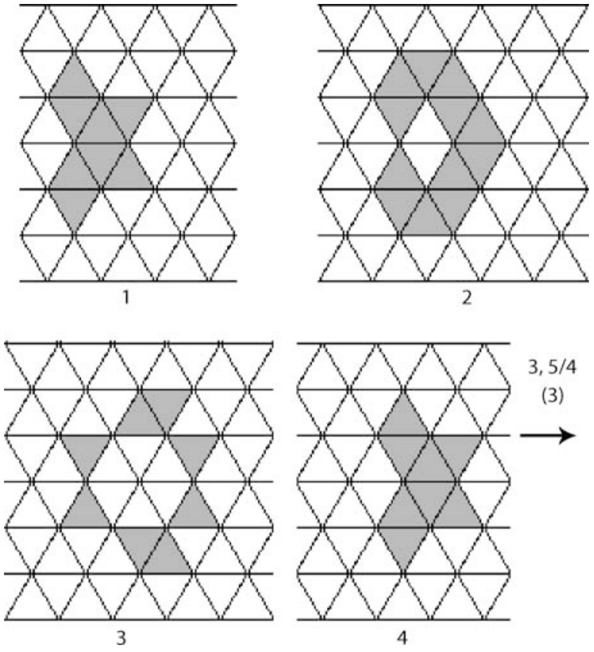
The neighborhood for the hexagonal grid is only half the size of that for the triangular grid and is symmetric – each neighbor is identical in the manner of contact with the cell in question. This symmetry can be important for some applications. Unfortunately, the hexagonal grid has a limited number of possible rules – there are only about 4000, many of which are of little interest. For many years past attempts to find a GL rule in the hexagonal grid have failed, although gliders were discovered by defining rules where the spatial relationship between neighbors was a factor [3]. Recently however the GL rule 3/2 was discovered. It supports the glider shown in Fig. 12. Another glider has also turned up; its rule is 3/2,4,5. Unfortunately this rule is not a GL rule, as it will very slowly exhibit unbounded growth, given a sufficiently large starting pattern.

The Pentagonal Grid

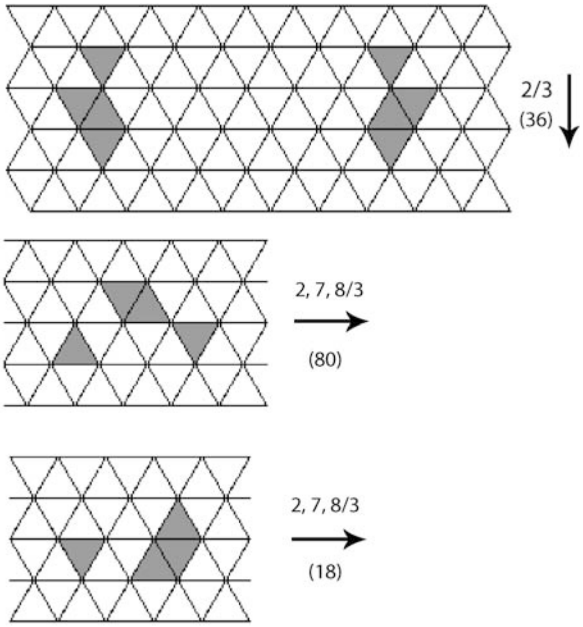
Regular pentagons cannot be formed into a grid, but by varying the angles and side lengths, we can create several tessellations from identical convex pentagons. A classification system has been devised, wherein 14 different types of tilings have been identified (Fig. 14). Of these, 12 are



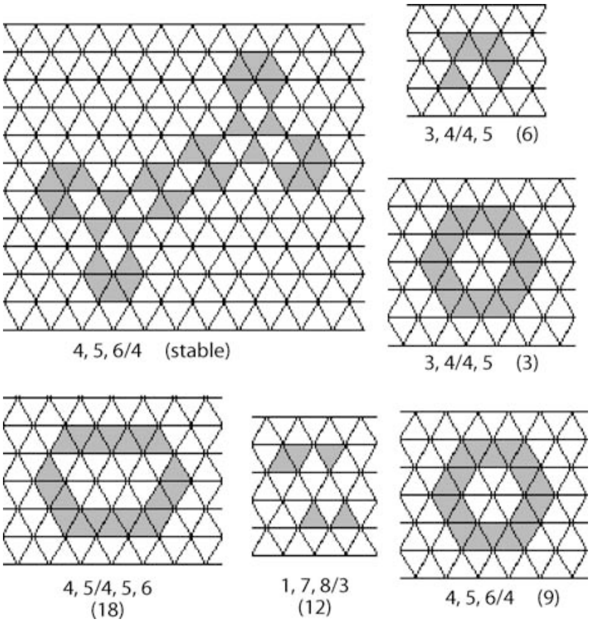
Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 5
Examples of bounded rules that churn endlessly. The total number of live cells can be employed as pseudorandom numbers that approximate a normal distribution. Many candidate rules can be used. Naturally the randomlike patterns eventually repeat, but with a sufficiently large initial shape, the period will be quite large. The plot at the *lower left* gives the number of live cells at each generation. These values exhibit a normal distribution (plot "A"). Note however that there are some gaps. This is because the rule 1-8/6-8 tends to have "clumps" of living (and fairly large "holes" of non-living) cells. Hence, before using this technique for generating random numbers, the candidate rule should be carefully investigated



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 6
A simple glider for the GL rule 3,5/4. It has a period of three (indicated in parentheses) after which it will have moved one cell to the right. Many gliders are not this well behaved, with much longer periods and irregular structure (see next figure)



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 7
Some gliders exhibit rather spectacular evolution. The period 80 2,7,8/3 glider, swells to 60 live cells during its swaggering trip across the grid, and at the 81st generation, will have moved 12 cells to the right. The gliders move in the direction given by the arrows. It should be noted that gliders have also been found for non-GL rules but since these rules are unstable they have not been investigated



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 8

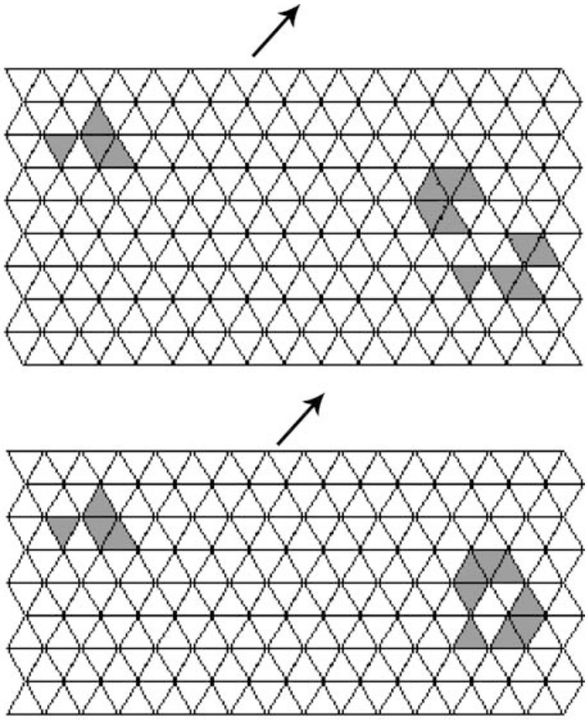
Hundreds of oscillators exist for the GL (and other) rules in the triangular grid. A few interesting ones are illustrated here. The stationary 4,5,6/4 form is representative of an infinite number of such objects that can be created for this rule by the careful positioning of live cells. The different oscillators at the lower right happen to share one identical shape. The oscillator at the *upper right* "rotates" clockwise, as does the period 12 oscillator at the bottom. Unfortunately rule 1,7,8/3 is not a GL rule

topologically distinct; these twelve varieties will behave in different ways under CA rules. We have chosen to investigate one of the most pleasing, the "Cairo tiling", so named because of its alleged use in parts of that city. It's appeal derives from the fact that the pentagons are both equilateral and isoseles.

Under the Cairo grid, there are rules that behave in the manner already described for the triangular grid – some rules expand, some stabilize, others contain a bounded, churning mass, etc. Interestingly, a GL rule has been discovered; its glider is depicted in Fig. 15. There is much opportunity for discovery in this and other pentagonal grids, as very little work has been done.

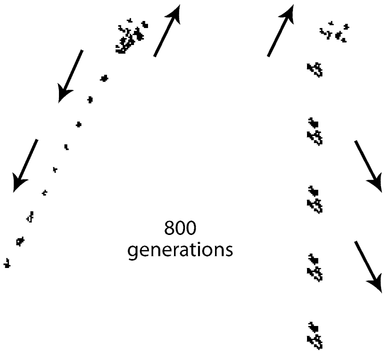
Programming Tips

We can speed up the scan of any grid by storing within each cell its current number of live neighbors along with a tag that indicates whether it is alive or not. Thus when we scan the entire grid for the next generation, we update



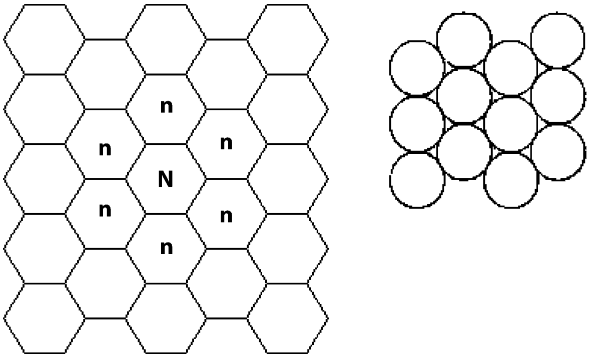
Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 9

The GL rule 2,7,8/3 is of special interest. It is the only known GL rule besides Conway's rule that supports a "glider gun" – a configuration that spews out an endless stream of gliders. In fact, there are probably several such patterns under that rule. Here we illustrate two guns; the top one generates period 18 gliders and the bottom one creates period 80 gliders. These configurations move in the direction shown, sending a stream of gliders out behind them (see next figure)

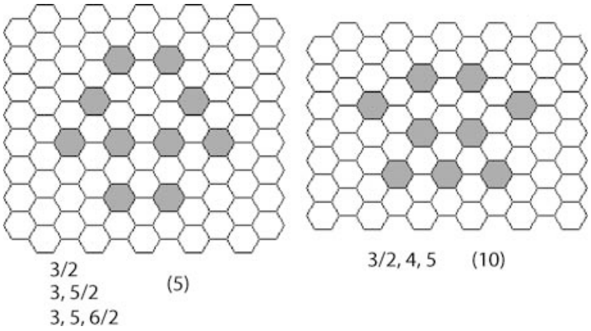


Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 10

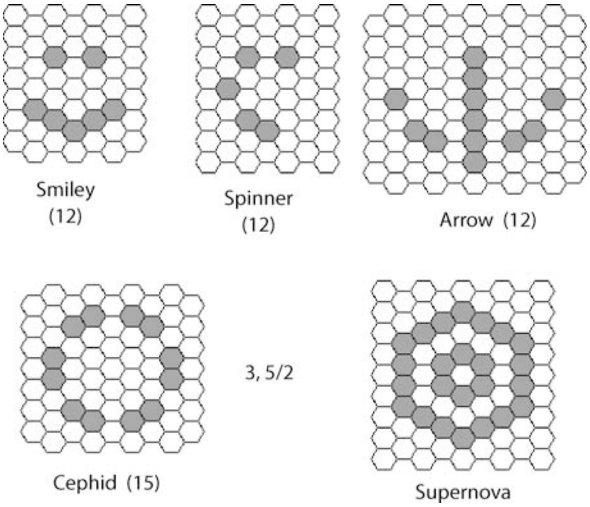
After 800 generations, the two guns will have produced the output shown. Motion is in the direction given by the arrows. The gun at the left yields period 18 gliders, one every 80 generations, and the gun at the right produces a period 80 glider every 160 generations



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 11
The symmetric hexagonal neighborhood. This rather "natural" grid can also be illustrated with *circles* (*upper right*) and, just as the square grid can be expanded to cubes in 3 dimensions, the hexagonal grid lends itself to "densely packed spheres" in three dimensions, where each sphere has exactly 12 touching neighbors



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 12
At least two gliders have been found. The GL rule 3/2 supports a period 5 glider and the non-GL rule 3/2,4,5 supports a period 10 glider. Note that the 3/2 glider also works for GL rules 3,5/2 and 3,5,6/2



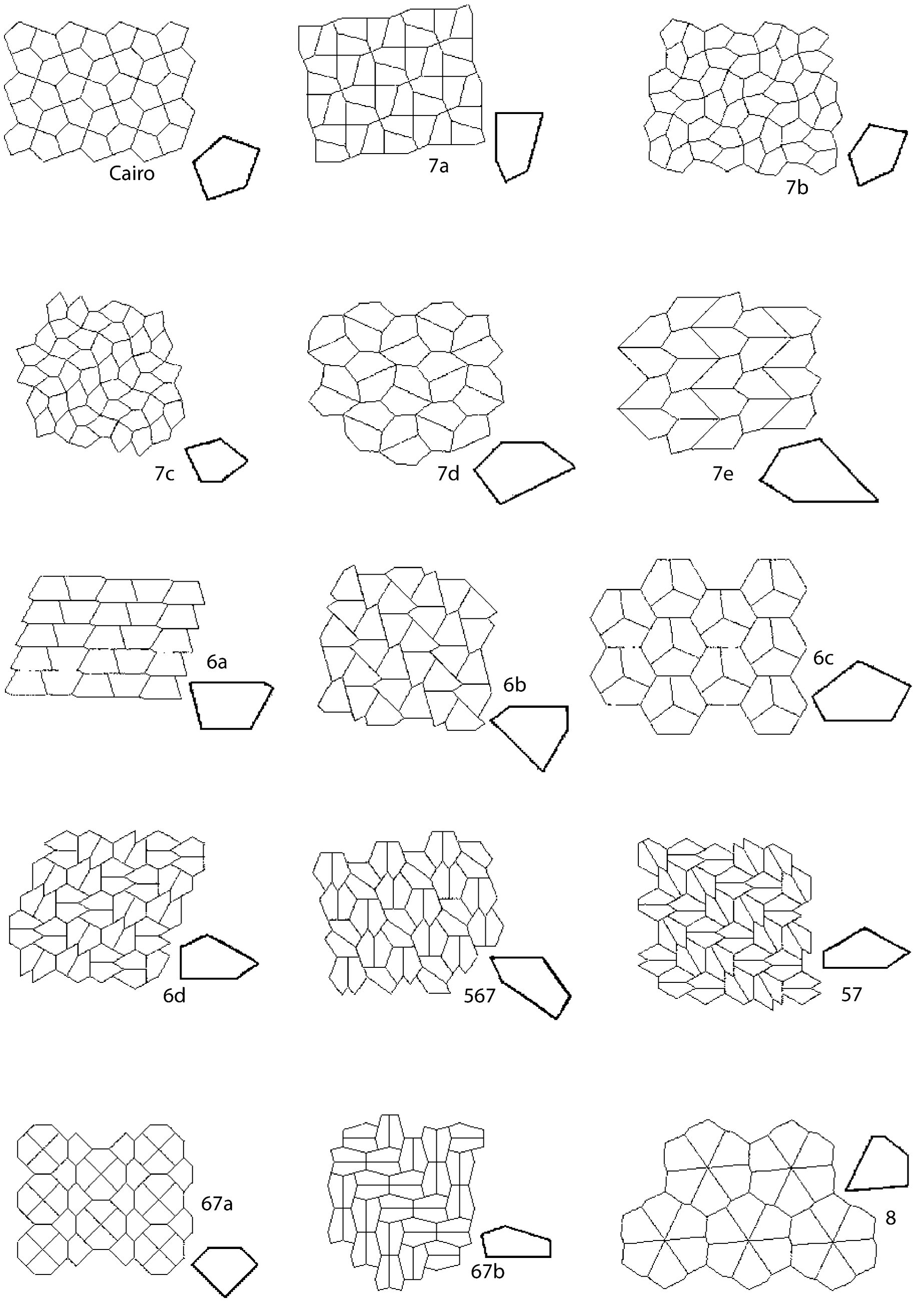
Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 13
Several interesting oscillators have been discovered for GL rule 3/2. They have been given rather whimsical names, a custom dating back to the early days of Conway's rule. After 65 generations the "supernova" pattern leaves a period 3 "neutron star" remnant. These patterns also work under rules 3,5/2 and 3,5,6/2

the status of cells that have changed since last generation (by examining their new neighbor counts) and, for each cell whose status has changed, we fix the neighbor counts for its neighbors; these cells are candidates for updating at the next generation iteration. This method employs two arrays – a "current" array, *A*, and a "next" array, *B*. And, rather than moving *B* back to *A* for the next iteration, we switch between the two; i. e. if array *A* is the array we are examining, then we copy it into array *B*, changing the status and neighbor counts of cells as needed. Array *B* then becomes array *A* for the next generation, etc. This trick allows us to rapidly scan over all non-changed cells. For further speed we can utilize hashing techniques and only store cells whose status is going to change. For this method, the speed of evaluation will depend only upon the number of cells that change between generations, and not the size of the grid, nor the total number of live cells. Furthermore, with a clever plotting algorithm, we can get away with replotting only those changed cells and not the entire grid.

We can program practically any grid or tiling in rectangular (square) coordinates by using templates to locate the neighbor cells as depicted in Fig. 16. The operation of finding the correct neighbors via templates adds a very small amount of time to the overall "next generation" evaluation; hence we would expect calculations on any type of grid to execute almost as fast as on the standard square grid.

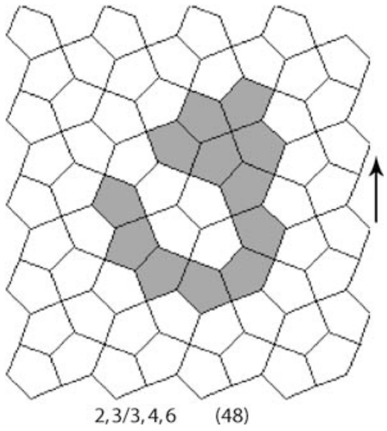
Future Directions

The triangular grid yields 12 touching neighbors and hence an ample supply of rules to investigate – many more than the 8 neighbor square grid. The hexagonal grid affords a more natural approach to CA than does the traditional 8 neighbor square grid, since neighbors all touch in the same way. Furthermore, when we expand this grid into three dimensions, we obtain a universe of dense packed spheres, which probably gives the best methodology for emulating 3D applications, as each cell has 12 touching neighbors and all touch in the same way. The fact that



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 14

The 14 distinct convex pentagonal tilings [see 1,5]. They are based upon certain relationships between the angles and lengths of the sides of the particular pentagon that constitutes the tiling. A sample of the pentagon for that tiling is displayed at the right of each. The tilings have been arranged to depict the number of touching neighbors for each cell. Where more than one number is given, there are some cells with each of those neighbor counts. For example the "67b" tiling is the second tiling where some cells have 6 neighbors and others 7. The Cairo tiling is at the *upper left* and is topologically equivalent to 7a and 7b. Note that 7c and 7d are also topologically equivalent

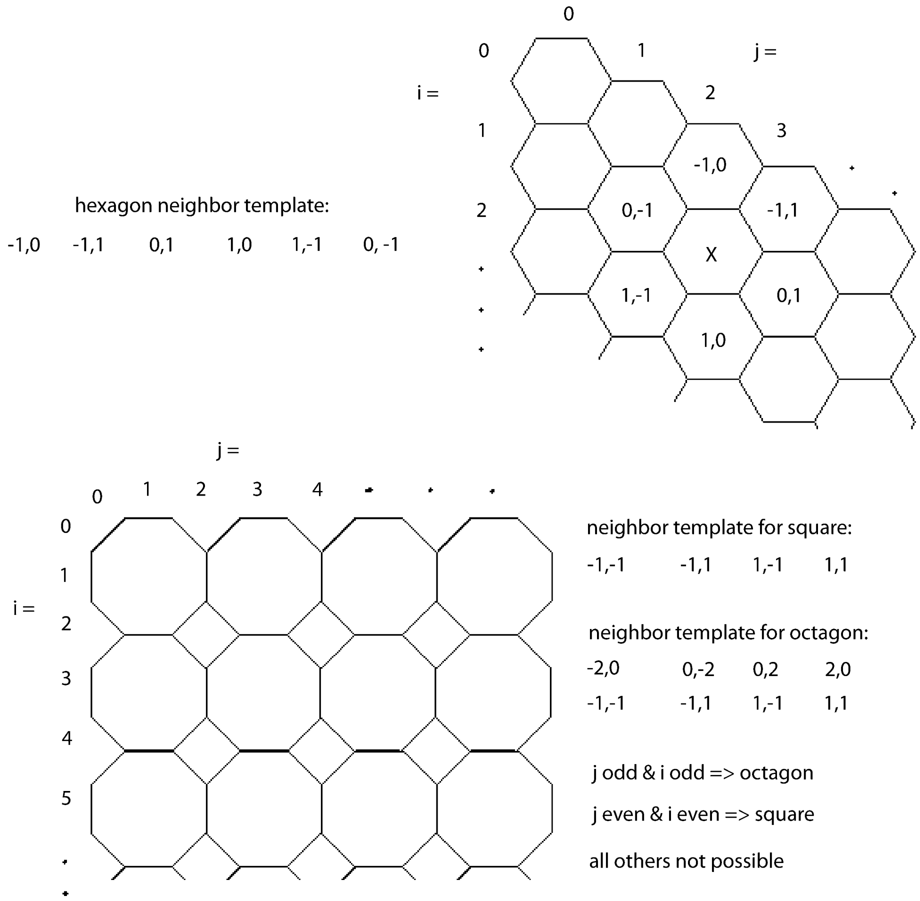


Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 15
The GL rule 2,3/3,4,6 supports the period 48 glider shown. It is asymmetric, though the second half of its period is a mirror image of the first. This characteristic is common amongst many gliders

GL rules have been found in a pentagonal grid undoubtedly means that other such rules can probably be found in many different tessellations – pentagonal and otherwise. The ultimate conclusion is that there is room for much work in the area of non-cartesian CA.

Bibliography

1. Bays C (2005) A Note on the Game of Life in Hexagonal and Pentagonal Tessellations. Complex Syst 15:245–252
2. Bays C (1994) Cellular Automata in the Triangular Tessellation. Complex Syst 8:127–150
3. Preston K Jr, Duff MJB (1984) Modern Cellular Automata. Plenum Press, New York
4. Sugimoto T, Ogawa T (2000) Tiling Problem of Convex Pentagon. Forma 15:75–79
5. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign



Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 16
Templates can be used to simulate any grid with rectangular coordinates. For example, if we are evaluating the neighbors for a hexagonal cell at (i, j) (see "X") they would be found at $(i-1, j)$; $(i-1, j+1)$; $(i, j+1)$; etc. We can even simulate grids made up of different types of polygons. Here, we determine the polygon type by examining the subscripts of the cell in question. Of course, appropriate graphics procedures must be employed in order to view our grid

Cellular Automata, Universality of

Jérôme Durand-Lose
Laboratoire d'Informatique Fondamentale d'Orléans,
Université d'Orléans, Orléans, France

Article Outline

Glossary

Cellular automata They are dynamical systems that are continuous, local, parallel, synchronous and space and time uniform. Cellular automata are used to model phenomena where the space can be regularly partitioned and where the same rules are used everywhere, for example: flow dynamics or percolation in physics, systolic arrays in computer science, epidemics in biology...

The configurations are infinite arrays of cells. Each cell has a state chosen inside a finite set. The dynamics is given by replacing the state of each cell according to it and the states of the cells at a bounded distance. Since there are finitely many neighboring cells, there are finitely many state patterns/inputs. The mapping to the new state is called the local function. The same local function is used for all the cells. They are all updated simultaneously.

Computational universality Computability is defined by Turing machine, μ-recursive functions or λ-calculus. All these approaches (and many more) end up defining the same set of functions over $\mathbb{N}$ (or on words, i. e. finite sequences over a finite alphabet): the computable functions. They defined, according to the Church–Turing thesis, what can be computed by any reasonable device.

A machine is computation universal if it is able to compute any computable function (indicated as a part of the entry). This corresponds also to the common approach of computer, the hardware is universal and the program to be executed is stored in main memory (like the data to process) and is part of the input as far as the hardware / operating system is concerned.

Intrinsic universality It is the capability to simulate any machine in a class of machine. If one think of Turing machines or an equivalent model of computation, this folds back to the classical computational universality. The interest of this notion lays with machines that are not equivalent to Turing machines. This is the case of cellular automata: they update infinite configurations, there are uncountably many possible configurations thus they cannot be encoded in a countable set, say $\mathbb{N}$.

An intrinsically universal CA "represents" all the CA since it can exhibit any phenomenon any other one can.

Definition of the Subject

Cellular automata (CA) and the subject are briefly defined before two kinds of universality are considered: computational universality and intrinsic universality. A more involving section on advanced topics ends this chapter.

Computational universality deals with the capability to carry out any computation as defined by Turing machines (in computability Theory) while *intrinsic universality* deals with the capability to simulate any other machine of the same class (here cellular automata). This distinction is fundamental here because while computational universality refers to finite inputs and relates to our understanding of computing with computers, intrinsic universality encompasses infinite configurations and relates to our understanding of the physical world. These universalities are presented as simply as possible and an example of universal CA is presented in each case. The last section is devoted to the history and advanced topics such as various definitions of simulation between CA, restriction to reversible CA and different underlying lattices.

Introduction

A cellular automaton is a discrete dynamical system composed of regularly displayed cells which satisfies the following properties:

1. *Local finiteness*: the set of states available for any cell is finite and there is finitely many cells in any bounded region of space;
2. *Locality of computation*: the next state of a cell only depends on the cells around it. There is no global data nor unbounded effect;
3. *Uniformity* in both space and time: the dynamics of the cells are identical (space) and never change (time);
4. *Parallelism*: all cells are updated at each iteration; and
5. *Synchrony*: all cells are iterated at the same instant.

Local finiteness and locality of computation ensure that the next state of any cell can be computed with finite information. This and uniformity ensure that a finite description exists. Parallelism and synchrony ensure that the system is deterministic.

Locality of computation also means that a finite part of a configuration can be isolated in order to see how its central part evolves. It also means that the system is continuous according to the product topology [21,48]. Uniformity also means that the date is not relevant and that if, starting from some local pattern, a phenomenon appends, when the same pattern appears again, the same phenomenon appends, whatever the iteration and location.

Universality. Being universal is somehow to represent all, to be capable to achieve anything possible. This notion is twofold: on the one hand it can be *absolute*, not related to anything in particular, and on the other hand, it can be *relative* to a specific domain. In the case of cellular automata, these two cases are: computational universality and intrinsic universality.

The first one deals with the capability to compute any computable function and relates to simulating Turing machines. Computability Theory and the Church–Turing thesis tell that the set of computable functions does not rely on one specific computing system. On the many textbooks on the subject, (see Part 2 in [51]) is among the best ones for a computer scientist approach.

The second one deals with the capability to simulate any other CA starting from any initial configuration. The distinction is real because since CA do not halt, the notion of the result of a computation is quite meaningless and because CA handle infinite configurations. Even if Turing machines with infinite entries are considered, they only update a limited part of the tape in finite time while CA update the whole (infinite) configuration at each iteration! Since there are computation universal CA in dimension one and simulation is expected to preserve this property, all intrinsically universal CA are also computation universal.

For more information on cellular automata and universality, the reader might be interested in the following surveys and books: [11,20,24,26,58,61].

Outline of this chapter. Section "Computational Universality" deals with computational universality and Sect. "Intrinsic Universality" deals with intrinsic universality. In each of these sections, definition and results as well as the construction of an example of a universal CA are provided.

Section "Advanced Topics" starts with some history on the subject and then presents advanced and more involving results on existing definitions of simulations between CA, results on the restriction to the reversible CA and on variations on different underlying lattices. Section "Future Directions" presents some insight on future researches.

Formal Definition of a Cellular Automaton

A configuration is a made of cells regularly displayed on $\mathbb{Z}^d$; d is called the *dimension*. Each cell is in a state chosen among a finite set of states Q. A *configuration* is then an element of the set:

$$C = Q^{\mathbb{Z}^d} ,$$

which is referred to as the *set of configurations*. The computation is local, each cell evolves according to the states of the cells whose coordinates differ by at most r (*radius*) on any coordinate. The local evolution is given by a *local transition function* f that maps the states of the cell and the neighboring ones to the new state of the cell. The *global transition function*, $\mathcal{G}$, maps configurations into configurations; each state is replaced by the image of the states of the cell and the neighboring ones by the local transition function. The subset of $\mathbb{Z}^d$, $\mathcal{N} = [\![-r, r]\!]^d$ is called the *(complete) neighborhood*. It represents the relative positions of neighboring cells. The neighborhood is not necessarily of this form, in fact it can be any finite subset of $\mathbb{Z}^d$. The local and global functions are defined by:

$$\begin{array}{rcl} Q^{\mathcal{N}} & \xrightarrow{f} & Q \end{array} \qquad \begin{array}{rcl} C & \xrightarrow{\mathcal{G}} & C \\ c & \mapsto & \mathcal{G}(c) \ \text{ s.t. } \ \forall x \in \mathbb{Z}^d , \\ & & (\mathcal{G}(c))_x = f(c|_{x+\mathcal{N}}) , \end{array} \tag{1}$$

where $c|_{x+\mathcal{N}}$ denotes the restriction of the configuration c to the positions in $x + \mathcal{N}$.

Definition 1 A *cellular automaton* (CA) is designated by: (d, Q, r, f). When the (finite) neighborhood does not correspond to some $[\![-r, r]\!]^d$, this is emphasized by using $\mathcal{N}$ instead of r. The *space-time diagram*, $\mathcal{D}\colon \mathbb{Z}^d \times \mathbb{N} \to Q$, or orbit of a CA is just the infinite sequence of the configurations as the CA is iterated.

Examples of space-time diagrams are provided on Figs. 2 (left) and 4. For 1-dimensional CA, the space-time diagram can be seen as a tiling of the plane. In dimension 2, each configuration can be considered as a tiling. This approach leads to many undecidability results. It is developed in the ▶ Tiling Problem and Undecidability in Cellular Automata.

Definition 2 If any, the *quiescent state* of a CA, $q_\#$, is a distinguished state such that the uniform configuration $\overline{q_\#}$ is map onto itself (which is equivalent to $f(q_\#, q_\#, \ldots, q_\#) = q_\#$). A configuration is *finite* if only finitely many cells are not in the quiescent state.

Computational Universality

In this section, cellular automaton are proved to be able to perform any computation in the acceptation of computability Theory and thus that there exist computation universal CA. Only useful concepts and definitions are presented.

The computable functions can be defined by μ-recursion, λ-calculus, Turing machines or any other equivalent model. The Church–Turing thesis asserts that one gets the same functions up to some encoding/representation with any reasonable mean of computation. This is important since, for example, recursive functions address functions over natural integers while Turing machines deal with computations over words (finite sequences over a finite set of symbols). A model of computation is *computation universal* if any computable function can be computed by an instance/machine of the model. A machine is *computation universal* if it can compute any computable function as long as it is provided with the description of the function to compute along with the corresponding input. Computability Theory guarantees that such a universal machine exists.

The typical techniques to prove that a model of computation, here cellular automata, is computation universal are by:

Induction like recursion Theory. This would be proving that some functions over integer (e. g. $n \mapsto n+1$) are computable and then that the functions computable in the model are closed under some operations (e. g. composition);

Simulation of a generic instance of a computation universal model of computation. This would be to prove that any, say, Turing machine could be simulated by a cellular automata; and

Simulation of a computation universal instance of a computation universal model of computation.

These approaches are very different. The first one relies on defining functions but is not concerned by the way they are computed. The second one deals with effective means of computation (allowing the implementation of algorithms and the measure of complexity). The third one is the modern vision of the all-purposes computer: a laptop can do anything from picture manipulation to music playing going through text edition and programming. There is only one hardware and according to the need, one uses a program or another. This is the duality of data and code in modern computers.

Generally, the second case is more simple to tackle. There are mainly two cases when a specific universal machine is transformed:

- When the situation is so complex that using a specific machine with few instructions becomes easier; and
- To get a computation universal CA with specific properties or qualities, typically as few states as possible.

The latter is done by designing a special simulation with a good property and then applying it to a particularly well chosen instance.

The term "computation universal" is not synonymous of "Turing equivalent" which means that whatever computes the model can be computed by, say, a Turing machine. In the case of cellular automata, this is meaningless for two reasons:

- The output is not defined since as a dynamical system, a CA never stops; and
- When considering infinite configurations, there is just no way to manipulate them with Turing machines (uncountably many configurations for countably many words).

The first reason leads to a notion of simulation in an not-ending computation rather than an input-output function definition approach. The second reason can be bypassed when considering restrictions of configurations: finite or periodic. Another canonical thing is to consider models of computation able to handle uncountably many different inputs… like cellular automata. This idea leads to intrinsic universality presented in Sect. "Intrinsic Universality".

A Computation Universal Cellular Automaton

Here Turing machines and their executions are defined and a simulation by cellular automata is provided.

A Turing machine is a very simple device: a finite automaton that can read and write on an unbounded memory (indexed by $\mathbb{N}$). The memory is organized as an infinite sequence of cells called the *tape*. Each cell has a value from a finite set of *symbols* (the alphabet). Only a finite part of the tape is not empty at any step of the computation. The automaton is equipped with a head that can read or write a single cell of the tape and move the head one position forward of backward on the tape.

Definition 3 A *Turing machine* is defined by $(\Sigma, \#, Q, q_i, \delta)$, where

- Σ is a finite *alphabet*;
- $\# \in \Sigma$ is a special symbol use to indicate *empty* part of the tape;
- Q is the *set of states* of the automaton;
- $q_i \in Q$ is the *initial state*, and
- $\delta\colon Q \times \Sigma \to Q \times \Sigma \times \{\leftarrow, \rightarrow\}$ is the *transition function*.

δ	a	b	#
q_i	q_i, b, $\rightarrow$	q_i, a, $\rightarrow$	r, #, $\leftarrow$
r	r, a, $\leftarrow$	r, b, $\leftarrow$	r, #, $\leftarrow$

x	y	z	$f(x,y,z)$
α	q_i, a	β	b
α	q_i, b	β	a
q_i, a	α	β	q_i, α
q_i, b	α	β	q_i, α
α	q_i, #	β	#
α	β	q_i, #	r, β
α	r, β	γ	β
α	β	r, γ	r, β

$\alpha, \beta, \gamma \in \{\mathtt{a}, \mathtt{b}, \#\}$

	b	
α	q_i, a	β

Cellular Automata, Universality of, Figure 1
Transition function of a Turing machine and the simulating CA

The transition function works as follows: in a state q, reading a symbol a, if $\delta(q, \mathtt{a}) = (r, \mathtt{b}, \rightarrow)$ then the new state is r, the symbol b is written instead of a and the head moves one step on the right/forward.

Definition 4 A *computation* of a TM starts with the input written on the tape (completed by #s) and the automaton in state q_i. The computation goes on as defined by the transition function δ. The computation ends when the head tries to leave the tape on the left. The result is what is written on the tape.

This is not the usual definition which involves an halting state, although it is correct. This formalization stresses that stopping is somehow an incident from the dynamical system point of view. As far as cellular automata are concerned, they do not stop; an operator might stop a CA when some condition is fulfilled but this is external to the CA.

The Turing machine considered as an example is very simple: starting on a word on $\{\mathtt{a}, \mathtt{b}\}$, it replaces each a by a b and vice-versa. The only symbols on the tape are a, b and #. There are only two states: q_i and r. The transition function is given on the left of Fig. 1.

This Turing machine is simulated by a CA of radius 1 (i. e. only cells at distance at most 1 are taken into account for computing the next state of a cell). The set of states of the CA is $\Sigma \cup \Sigma \times Q$, that is, a tape symbol alone or together with a state of the TM. The CA has 9 states, the table of its local function has $729 (= 9^3)$ entries! In the table on the right of Fig. 1 only the cases where the state of the central cell changes are indicated. On the far right, the first transition rule is represented as it appears on space-time diagrams: the bottom line represents the cell and its two closest neighbors and on top the next state of the cell (at the next iteration). In all the space-time diagrams, time is evolving upward.

The dynamics is presented on Fig. 2. On the left the whole computation of the Turing machine on the entry aab is given; the output of the function is bba as written on the tape when the machine stops. On the right, the corresponding iterations of the simulating CA are displayed. Since the CA works on a bi-infinite lattice, the configuration is completed with # on the left. As mentioned, a CA never stops so that at some point the computation has been carried out but the system goes on.

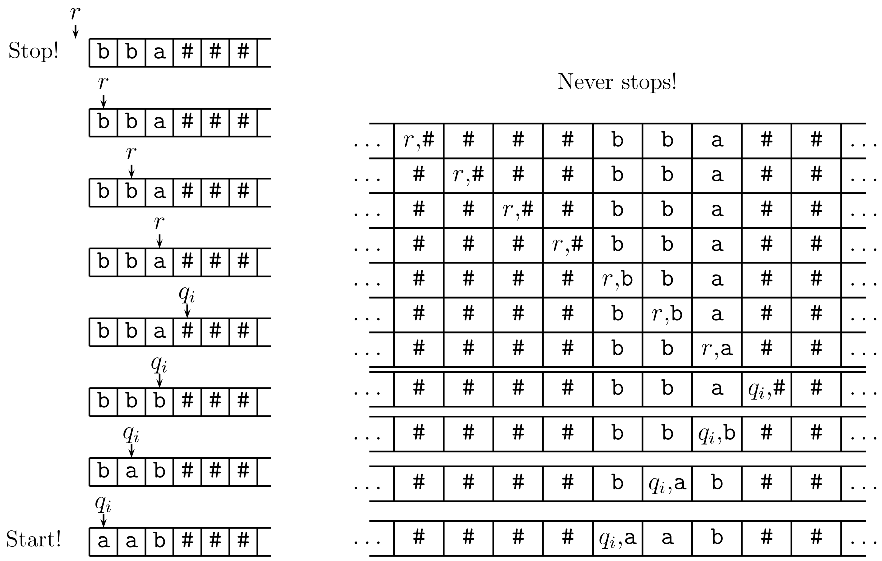
A "halting" condition has to be provided, especially when one remembers that the halting of a Turing machine (the famous *Halting problem*) is not decidable. Usually, something very simple to test is chosen, for example that some state appears somewhere. Here this would be the first time the closest #-cell on the left is not in state #.

This example can easily be extended to a general method to simulate any Turing machine or specifically a computation universal one. Starting from a computation universal TM, a computation universal CA is generated.

Other Ways to Achieve Computational Universality

Among the various systems achieving computational universality that have been used to prove computational universality of CA, a brief classification can be made.

Machines. (Turing machines, counter automata and random access machines) These systems are very simple, an automaton together with a memory. In general, the whole evolution/orbit of the Turing machine is encoded inside the space-time diagram of the cellular automata like in the example. For counter automata [36], there are finitely many counters but they can hold any natural integer, any counter can be accessed at any time. Random access machines are like counter automata but with infinitely many register and an indirect access mode which allows to access any register. The random access machines model is the closest to modern computer architecture, but to sim-



...	r,#	#	#	#	b	b	a	#	#	...
...	#	r,#	#	#	b	b	a	#	#	...
...	#	#	r,#	#	b	b	a	#	#	...
...	#	#	#	r,#	b	b	a	#	#	...
...	#	#	#	#	r,b	b	a	#	#	...
...	#	#	#	#	b	r,b	a	#	#	...
...	#	#	#	#	b	b	r,a	#	#	...
...	#	#	#	#	b	b	a	q_i,#	#	...
...	#	#	#	#	b	b	q_i,b	#	#	...
...	#	#	#	#	b	q_i,a	b	#	#	...
...	#	#	#	#	q_i,a	a	b	#	#	...

Cellular Automata, Universality of, Figure 2
Iterations of a Turing machine and of the simulating CA

ulate it, one has to consider indirect addressing and infinitely many register. Counter automata are more simple to simulate: since there exists a computation universal 2-counter automaton, only 6 instructions have to be implemented.

Boolean circuits. The idea is to encode Boolean logic and then to say that any value can be encoded in binary and any function can then be computed with the binary encoding. For example, the transition function of a TM (resp. 2-counter automata) can be encoded and the value of the tape (resp. the counters) stored in an infinite memories. This is generally done by providing a way to encode bits, then logical gates and finally wiring to connect them. It is usually done in dimension 2, since ensuring a correct wire crossing is more complicated in dimension 1. A typical example of this is the computational universality of the Game of Life [6,19]. This is developed in the ▶ Gliders in Cellular Automata.

Rewriting systems. These systems work on words by removing some sub-word and adding some other sub-word. Starting with a word encoding an entry, the system is expected to stop with the output encoded in the final word. Here are some examples of computation universal systems:

- Type-0 grammar in Chomsky's hierarchy: one may replace sub-words by others according to rewriting rules; and
- Tag-systems: a prefix of the word is removed and according to it and some rules, a suffix is added.

The proof of the universality of the 2-states 3-neighbors 1-dimension CA referred as the elementary cellular automaton 110 by Cook [9] is done with some tag-system. The construction relies on an intermediate level of simulation that ensures signal transmission and updating. It is too involving and lengthy to be presented here. Signals and particles/solitons [1,53] are often used to carry bits/information around.

Consequences of Computational Universality

To have computation universality provides two things. On the positive side:

- Any computation can be carried out, so that if something has to be computed, whether as a final result or to use in a broader scheme, it can be done;
- Since any computation may take place, complex and interesting behaviors may appear;

and on negative side: many questions one may ask about the system become undecidable. The second point comes from the undecidability of the *Halting problem* (whether a given computation of a TM halts). Here are a few examples of this – many more can be found in, for example the ▶ Tiling Problem and Undecidability in Cellular Automata –, given a CA and a finite initial configuration:

- Will some state ever appear?
- Will the CA ever enter a stable configuration?
- Will the configuration grow infinitely?
- Will some given configuration be reached?

"Quality" of a Computation Universal Cellular Automata

Computability Theory is rather disappointing: one the one side what is computing and computable becomes clear but on the other side it mostly provides negative results like this and that are not computable. In this subsection slight differences on computability/simulation are addressed as well as complexity (of computing) issues.

For Turing machines, the written part of the tape is always finite. The tape is *potentially infinite*, it is extended by # as much as needed, it is never infinite. For computing with a CA, it can be assumed that only a finite part of the space is used for computing since otherwise it would take an infinite time for states to interact (it takes $l/2r$ for cells at distance l to interact). There is a definition of finiteness for CA (Def. 2): outside a finite part is it the quiescent state $q_{\#}$ (which plays the same role as #). When an infinite configuration (even if it is periodic after some point) is used to achieve computational universality, one talks about *weak computational universality*; more insights on the weakness notion on Turing machines can be found in [62].

Another important criterion is the one of the efficiency of computation. For example, the simulating a Turing machine by a 2-counter automaton suffers an exponential slowdown so that one may not be interested in such a computing device or any device where computational universality derives from a 2-counter automaton. For example, the proof of the computational universality of rule 110 by Cook [9] has an exponential slowdown but Neary and Wood [42] proved that in fact that it can be done with a polynomial slowdown. Polynomial slowdown is the classic simulation mode between reasonable models of computation (and this one of the key to the definition – and stability – of the complexity class P –polynomial time solvable problems).

Another thing worth mentioning is that CA are inherently parallel, the universality proofs more or less directly leads to the Turing machines which is the canonical sequential model. Various approaches have been made to consider CA as a computing system on its own, independently of any other model of computation, for example:

- *Iterative automaton* where the input is given symbol by symbol to a distinguished cell;
- As word recognizers with only one cell per symbol [55] (this cannot be computation universal because the memory is bounded);
- If finitely many cells are considered but they are all equipped with a stack, then any computation can be made [27];
- A algorithmic on CA based on signals has been developed [12,32,35]; and
- Martin devised an intrinsically universal CA together $S-m-n$ theorem for CA (as computing devices over infinite configurations) to provide a *acceptable programming system* point of view [31].

This directly links to the notion of universality developed in the next section.

Intrinsic Universality

Definition

Previous section deals with universality as the capability to perform any computation as defined in computability Theory. This theory deals with numbers/words and there exist only countably many configurations. But there are uncountably many configurations for any CA (unless, of course, if there is only one state). It is worth inquiring about another kind of universality that would take this into account and not involve any other model of computation. (It can thus be used for any class of dynamical systems.) It is some kind of *inner-universality*. The idea behind *intrinsic universality* is somehow the counterpart of universality for a Turing machine: being able to simulate any other CA (of the same dimension) on any (infinite) configuration.

Definition 5 A cellular automaton $\mathcal{A}$ *simulates* another CA $\mathcal{B}$ if there is an injective (one-to-one) function, ι, from the configurations of $\mathcal{B}$ to the ones of $\mathcal{A}$ and an integer, τ, such that the following diagram commutes:

$$\begin{tikzcd}
C_{\mathcal{B}} \arrow[r, "\iota"] \arrow[d, "\mathcal{G}_{\mathcal{B}}"'] & C_{\mathcal{A}} \arrow[d, "\mathcal{G}_{\mathcal{A}}^{\tau}"] \\
C_{\mathcal{B}} \arrow[r, "\iota"] & C_{\mathcal{B}}
\end{tikzcd}$$

A cellular automaton is *intrinsically universal* if it can simulate any other CA (of the same dimension).

The injectivity ensures that $\mathcal{B}$-configurations can be distinguished when mapped into $\mathcal{A}$-configurations. From this definition, it directly comes that:

$$\forall n \in \mathbb{N}\,, \quad \iota \circ \mathcal{G}_{\mathcal{B}}^{n} = \mathcal{G}_{\mathcal{A}}^{n.\tau} \circ \iota\,.$$

In an infinite run, $\mathcal{A}$ generates one iteration of $\mathcal{B}$ every τ iterations. Since the composition of injective functions is injective, the simulation relation is transitive. The definition is illustrated by the construction of an example in the rest of this section.

Other definitions of simulation can be found in Subsect. "Defining Simulation Among CA (For Intrinsic Universality)". In every case, a different intrinsic universality is generated.

The Way It Usually Works

It is assumed that the simulated CA has radius 1 and is 1-dimensional (how to deal with higher dimension is explained at the end of this section). If it is not the case, it is easy to simulate it by one of radius 1 (and then to use transitivity). Let $\mathcal{A} = (1, Q, r, f)$ be such that the radius is greater than 1. The idea is to group cells r by r and define $\mathcal{B} = (1, Q^r, 1, f_{\mathcal{B}})$ such that $f_{\mathcal{B}}$ is:

$$\begin{aligned}
&\forall (c_1, c_2, \ldots, c_r), (c_{r+1}, c_{r+2}, \ldots, c_{2r}),\\
&(c_{2r+1}, c_{2r+2}, \ldots, c_{3r}) \in (Q^r)^3,\\
&f_{\mathcal{B}}((c_1, c_2, \ldots, c_r), (c_{r+1}, c_{r+2}, \ldots, c_{2r}),\\
&(c_{2r+1}, c_{2r+2}, \ldots, c_{3r}))\\
&\qquad = (f(c_1, c_2, \ldots, c_{2r+1}), f(c_2, c_3, \ldots, c_{2r+2}),\\
&\qquad\qquad \ldots, f(c_r, c_{r+1}, \ldots, c_{3r})) .
\end{aligned}$$

The injection ι is the canonical injection where states are grouped r by r:

$$\forall c \in C \, , \forall i \in \mathbb{Z} \, , (\iota(c))_i = (c_{ir}, c_{ir+1}, \ldots, c_{(i+1)r-1}) .$$

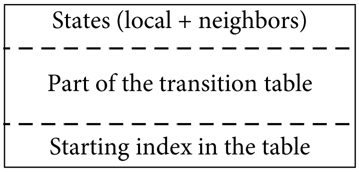
This simulation is just a rescaling of space. From now on, only radius 1 CA are considered. The construction of an intrinsically universal CA uses two scales:

- *Meta-cells scale* to manipulate the states and the transition function of the simulated CA; and
- *Bit scale* to implement the meta-cells.

A meta-cell gathers copies of the states of its two closest neighbors. Then it has to use the transition function to compute the new state. The transition function is tabulated in the form of a sequence of blocks $(\mathtt{a}, \mathtt{b}, \mathtt{c}, f(\mathtt{a}, \mathtt{b}, \mathtt{c}))$. (It could also have been be represented by, e. g., a boolean circuit with states binary encoded as in [45].)

Meta-Cells Scale Each meta-cell holds one block/entry of the transition table. The entries are infinitely repeating on both side and are endlessly shifting on the left so that in one period, the block corresponding to the update eventually appears. When one period has passed a new cycle starts. The architecture of meta-cells is presented on Fig. 3.

In the upper part of Fig. 4, a space-time diagram generated by some CA is given as well as the transition function of the CA in terms of a relation inside the space-time diagram between three states (at an iteration) and the state above (at the next iteration). In the lower part of Fig. 4, the simulation, at the meta-cell scale is presented. Each meta-cell corresponds to the presentation in Fig. 3. The transition table is set inside the initial configuration. The starting index is used by the meta-cell to detect the end of the cycle.


States (local + neighbors)
Part of the transition table
Starting index in the table

Cellular Automata, Universality of, Figure 3
Meta-cell

As it can be seen on Fig. 4, on the upper part of each cell; at start (lowest row) there are three copies of the state of the simulated cell then, copies are exchanged with the neighbors (first simulating iteration). Then the transition table starts moving. As soon as the entry is found in the table, the three states are replaced by three copies of the new state. The meta-cell then waits for its index to appear again to start this cycle again. The synchrony and uniformity of CA ensures that all meta-cells stay synchronized.

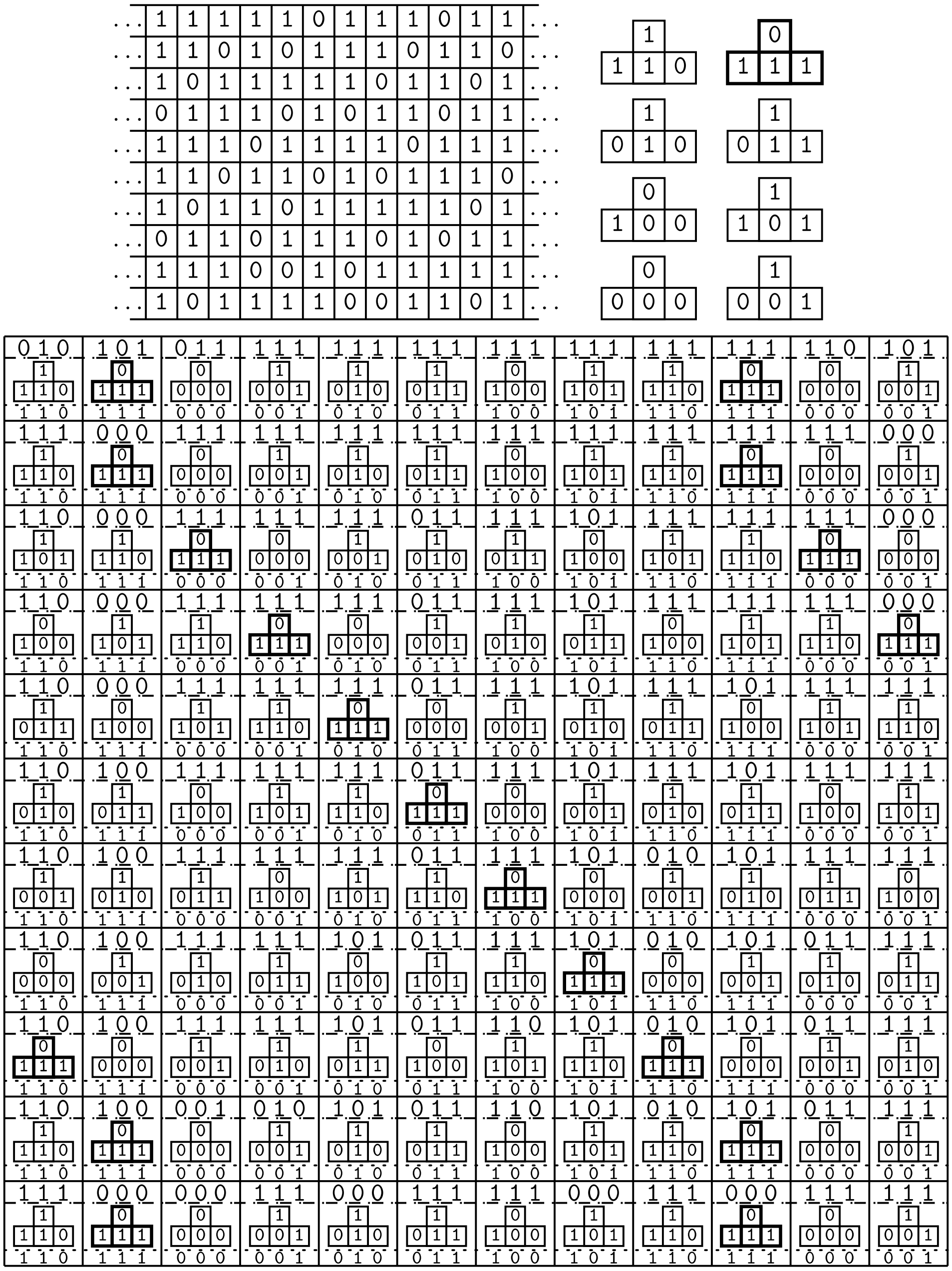
Bits Scale Since the encoding and simulation must work for any CA, it must be able to handle any set of states. The set of states of any CA if finite but unbounded, thus it is impossible to use a common set of states. A state a is thus binary encoded (denoted $(a)_2$). This way if the set of states is larger, then the meta-cell is larger, composed of more elementary cells.

On Fig. 5, a meta-cell is given and the layers named. The set of states on each layer for elementary cells are also given. Some layers are added:

- Structure: to delimit the different parts of the meta-cell,
- Control: to drive the movement, copy and test of bits, and
- Left and right: to carry bits around (to exchange states and to shift the transition table).

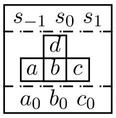
The left and right layers carry the bits very simply: unless noted otherwise the new value in left (resp. right) is the one that on the left (resp. right) layer of the right (resp. left) cell ensuring a left (resp. right) shift on the layer. When the simulation starts, s_{-1} and s_1 are equal to s_0 and (a, b, c) is equal to (a_0, b_0, c_0).

The meta-cell implementation works as follows: there is a single value in the control layer. It moves forth and back on the entire meta-cell like a signal. It manages bits from various layers and changes its state accordingly. The construction is only sketched; the states mentioned below are in an intermediate level between meta-cells and elementary cells. The universal CA is not detailed because although this is not complicated it would be quite lengthy and not very informative.



Cellular Automata, Universality of, Figure 4
Iteration of a CA and simulation at meta-cell scale of the first iteration

Starting in state q_0, a meta-cell first carries out the states exchange with the meta-cell on the left (the meta-cell on the right takes care of the other exchange) bit by bit using the left and right layers. In state q_1, the transition table is shifted by one entry. In state q_2, it moves through the entire meta-cell and checks whether the state and rule-in layers are identical. If the layers agree, the right transition is found and the state layer can be updated (the bits are copied) then it enters q_3 otherwise it restarts in q_1. In state q_3, the meta-cell has been updated; the transition



$*$	$-\ \cdots\ -$	$+$	$-\ \cdots\ -$	$+$	$-\ \cdots\ -$	structure $\in \{*,+,-\}$
q_0	$-\ \cdots\ -$		$-\ \cdots\ -$		$-\ \cdots\ -$	control $\in Q_{\mathcal{U}} \cup \{-\}$
	$0\ \cdots\ 0$		$0\ \cdots\ 0$		$0\ \cdots\ 0$	left $\in \{0,1\}$
	$0\ \cdots\ 0$		$0\ \cdots\ 0$		$0\ \cdots\ 0$	right $\in \{0,1\}$
	$(s_1)_2$		$(s_0)_2$		$(s_{-1})_2$	state $\in \{0,1\}$
	$(d)_2$		$(d)_2$		$(d)_2$	rule-out $\in \{0,1\}$
	$(a)_2$		$(b)_2$		$(c)_2$	rule-in $\in \{0,1\}$
	$(a_0)_2$		$(b_0)_2$		$(c_0)_2$	index $\in \{0,1\}$

Cellular Automata, Universality of, Figure 5
Binary encoding of a meta-cell with elementary cells

table is still shifted until a full shift has been done. This is indicated by identical rule-in and index layers. In this the case it enters q_0 and the simulation of a new iteration starts.

Various technical things like delays are added to keep meta-cells synchronized. The number of iterations needed to simulate an iteration is roughly speaking the length of a cell multiplied by the length of the transition table: $\log(|Q|).|Q|^3$.

Higher dimension. It is treated exactly in the same way. One layer gathers the information, another one deals with the transition function. The transition table shifting is done on one dimension only. Extra dimensions can be used to accelerate the process or to design circuitry for a more efficient encoding of the transition function.

Intrinsically universal CA are computation universal by composition (as long as the definitions are robust enough). There is no notion of semi-weakly simulation since the whole configuration has to be used to encode an infinite configuration.

Advanced Topics

This section gathers a brief history of the subject and various results on specific approaches. It is more involving and targeted to a learned reader.

A Bit of History

Since there is a lot of places where the history of CA is presented (as well as other chapters of this encyclopedia, the reader might be interested in the following survey: [50]), only the universality part is developed here.

Cellular automaton were introduced in the 50's by Ulam and von Neumann [59] to study self-reproduction [3,7,60]. Computational universality was used just to prove that any pattern can be built. Universality was investigated and proved without any explicit distinction between computational and intrinsic universality before [5].

The most famous CA is certainly Conway's Game of Life [19] from the early 70s. This 2-dimensional CA is both computation universal [6] and intrinsically universal [14].

Quest for Small Universal CA There have been an ongoing search for more than four decades for universal CA as small as possible. The interest is to know whether "small" CA are more simple and thus can be handled or there is no gap in complexity. Table 1 sums up results in this quest. Intrinsically universal CA are also computation universal but the converse is not true.

These results were achieved with various constructions. Some use quite different definitions of CA for commodity (but there are indeed CA):

- As partitioned CA [38], and
- As CA with Margolus's neighborhood (or partitioning CA): [30,58] (billiards) and [10] (spin model in Physics).

Computing universality can also be defined by the computational complexity of the sets of orbits of a CA as well as the reachability relation, relating to the Turing degrees of undecidability [54].

Defining Simulation Among CA (For Intrinsic Universality)

The definition used in Sect. "Computational Universality" is not the only existing one. Many papers prove results on simulation without providing a formalized definition of it, but by considering the construction anyone would say that it is a simulation, in an empirical fashion. There is no absolute definition commonly accepted and none contradicts the intuition of simulation. In Table 1, no distinction on the simulation is made and most of the time the construction fits in more than one definition.

Most of the presented definitions can be amended in order to cover cases where not all iterations are covered,

Cellular Automata, Universality of, Table 1
Historic bounds on computation and intrinsically universal CA

Year	Reference	Dimension	$\lvert\mathcal{N}\rvert$	$\lvert Q\rvert$	Computation	Intrinsic
1966	Neumann [60]	2	5	29	✘	✘
1968	Codd [8]	2	5	8	✘	✘
1970	Banks [5]	2	9	2	✘	✘
		1	3	18	✘	✘
		1	5	2	✘	✘
1971	Smith III [52]	2	7	7	✘	
1987	Albert and Culik [2]	1	3	14	✘	✘
1990	Lindgren and Nordahl [28]	1	3	7	✘	
2002	Ollinger [45]	1	3	6	✘	✘
2004	Cook [9]	1	3	2	✘	
2006	Richard [47]	1	3	4	✘	✘

say for example only the one out of 3. In the following, $\mathcal{A} = (Q_{\mathcal{A}}, r_{\mathcal{A}}, f_{\mathcal{A}})$ and $\mathcal{B} = (Q_{\mathcal{B}}, r_{\mathcal{B}}, f_{\mathcal{B}})$ denote cellular automata of the same dimension d.

Embedding of Hertling The term *embedding* is to be understood as simulation. It was introduced in order to prove that CA that are not onto cannot be simulated by onto (and specially reversible) ones of the same dimension.

Definition 6 (Embedding [23]) A mapping $\mu\colon Q_{\mathcal{A}}^{\mathbb{Z}^d} \to Q_{\mathcal{A}}^{\mathbb{Z}^d}$ is said to be a *morphism* if and only if for any shift of $\mathbb{Z}^d$, $\sigma_{\mathcal{A}}$, there exists a shift of $\mathbb{Z}^d$, $\sigma_{\mathcal{B}}$, such that: $\mu \circ \sigma_{\mathcal{A}} = \sigma_{\mathcal{B}} \circ \mu$.

$\mathcal{A}$ can be embedded in $\mathcal{B}$ if there are mapping $\mu\colon Q_{\mathcal{A}}^{\mathbb{Z}^d} \to Q_{\mathcal{A}}^{\mathbb{Z}^d}$ and $\nu\colon Q_{\mathcal{A}}^{\mathbb{Z}^d} \to Q_{\mathcal{A}}^{\mathbb{Z}^d}$ and an integer k such that:

$$\forall t \in \mathbb{N}\,, \quad G_{\mathcal{A}}^{t} = \nu \circ G_{\mathcal{B}}^{kt} \circ \mu\,. \tag{2}$$

The embedding is *strong* if μ is a continuous morphism, *weak* if it is a just a morphism and *set-theoretic* otherwise (i. e. not a morphism).

The interest of the morphism is to enforce the respect of the structure of $\mathbb{Z}^d$. Hertling proved, using the Axiom of Choice (hence the qualification) that any CA can be set-theoretically embedded in the one dimensional CA that does nothing but shift the value on the left. This is of course highly nonconstructive and contradict the intuition of what simulation could be.

Comparing to Def. 5, using ν to come back allows to have some garbage produced and discarded by ν.

Grouping Relation of Mazoyer and Rappaport This definition of a grouping relation was introduced to focus on the importance of space and time structure. It allows to consider iterations once in a while, periodically.

Definition 7 (Grouping [33,34]) A cellular automaton $\mathcal{A}$ is a *sub-automaton* of $\mathcal{B}$ (denoted $\mathcal{A} \subseteq \mathcal{B}$) if there is an injection $\phi\colon Q_{\mathcal{A}} \to Q_{\mathcal{B}}$ such that:

$$\overline{\phi} \circ G_{\mathcal{A}} = G_{\mathcal{B}} \circ \overline{\phi}\,,$$

where $\overline{\phi}$ is the component-wise extension of ϕ to $\mathcal{A}$-configurations. The nth grouping of an automaton is defined by grouping the cells n by n and consider only every nth iteration, $\mathcal{A}^n = (d, Q_{\mathcal{A}}^n, r_{\mathcal{A}}, f_{\mathcal{A}}^{(n)})$. The function $G_{\mathcal{A}^n}$ is the nth iterate of $G_{\mathcal{A}}$. Since the cells are grouped by n, the radius is not changed. The grouping relation is defined by:

$$\begin{aligned}&\mathcal{A} \leq_{\text{grouping}} \mathcal{B}\\ &\qquad \iff \exists n, m \in \mathbb{N}\,, \quad 0 < n, m\,, \quad \mathcal{A}^n \subseteq \mathcal{B}^m\,.\end{aligned}$$

This is a stronger form of simulation where the space-time diagrams should be included; up to some rescaling on both side, all the space-time diagrams of $\mathcal{A}$ should be exactly (up to an injection) generated. There is no shift involved and the space-time ratio should be preserved.

This relation is a pre-order. The authors proved that there is a bottom equivalence class for CA: the CA with only one state (and one configuration). The nilpotent ones (after a fixed number of iterations any configuration is turned into the same one: only quiescent state) are just above. They also proved that there is an unbounded infinite ascending chain, so that there is no top and thus no intrinsically universal CA for this definition.

Rescaling of Ollinger The previous definition is on the one hand interesting because it relays on space-time diagrams and a natural operation, grouping, over them, but on the other hand, there is no intrinsically universal CA, which have been provided for other definition. The following definition is a weakening of the first one that allows intrinsically universal CA.

Definition 8 (Rescaling [44]) For any k in $\mathbb{Z}^d$, let σ^k denotes the *shift* by k over configurations. For any m in $\mathbb{Z}^d$, let o^m denotes the *packing* of cells into packs of size m; it is a mapping from $Q^{\mathbb{Z}^d}$ into $(Q^m)^{\mathbb{Z}^d}$ (o^{-m} is the inverse, the unpacking function).

For $n, k \in \mathbb{Z}^d$ and $n \in \mathbb{N}$, $0 < n$, the $\langle m, n, k\rangle$-rescaling of $\mathcal{A}$ is the cellular automaton $\mathcal{A}^{\langle m,n,k\rangle}$ such that:

$$\mathcal{G}_{\mathcal{A}^{\langle m,n,k\rangle}} = \sigma^k \circ o^m \circ \mathcal{G}^n_{\mathcal{A}} \circ o^{-m} \; .$$

A cellular automaton $\mathcal{A}$ is simulated by $\mathcal{B}$ if there exists a rescaling of $\mathcal{A}$ which is a sub-automaton of a rescaling of $\mathcal{B}$. (The sub-automaton relation is defined as in Def. 7.)

This definition allows to include a shift and to treat independently the size of the blocks of cells and the iteration step. Comparing to the grouping definition (Def. 7), this allows to have enough time to locally mix information and compute the next state. Intrinsically universal CA exist (also with a meta-cell approach) and intrinsic universality of a 1-d CA is undecidable [46]. This result is still true on captive CA (the transition function may only output a state that is in the input) even though as the number of states grows larger, almost all captive CA is intrinsically universal [56,57].

Reversible Case

The reversible subset of cellular automaton (CA such that the global function is invertible, its inverse is then the one of a CA) also contains computation universal CA [13,38,40]. This topic is developed in the ▶ Reversible Cellular Automata.

There exist reversible CA that are intrinsically universal inside reversible CA [15,16,18]. This means able to simulate any other reversible CA (of the same dimension) and not just any CA. It is a strong embedding (Def. 6) and does not contradict Hertling's results [23] that non surjective CA cannot be simulated by reversible one.

Morita proved the any CA can be simulated over finite configurations by a reversible in [37,39] but garbage is produced in order to ensure reversibility and the simulation time varies as the simulation goes on.

There is also a particular result [17] including the simulation of the non-reversible CA, but the simulation goes by a different definition. It is centered (like the usual metric) and the simulated iterated configurations are displayed on parabolas. This twisting of space yields an infinite space to store the information for reversibility. It is not possible to recover a simulated iteration from finitely many simulating ones! The whole simulating space-time diagram is needed.

Variations on CA

Changing the Underlying Space Other 2-dimensional spaces have been considered. There exist reversible computation universal CA both on triangular lattices [25] and hexagonal lattice [41].

Róka studied simulation between CA on different lattices in a very general way: lattices are Cayley graph (it corresponds to a group, the arrows corresponds to generators). She proved the existence of simulations in the case of the existence of an homomorphism with a finite kernel and that all bi-dimensional planar structure are equivalent to $\mathbb{Z}^2$ [49].

There exist computation universal CA [22] and intrinsically universal CA [29] on the hyperbolic plane. This is more developed in the ▶ Cellular Automata in Hyperbolic Spaces.

Intrinsic Universality Among Quantum Cellular Automaton In the past decade, quantum computation theory has been tremendously developing. It relies on unitary gates which are of course reversible. The results on reversible CA has been "naturally" extended, for example, there is a 1-dimension Quantum CA which is intrinsically universal (among quantum CA) [4]. For more on the topic, please refer to the ▶ Quantum Cellular Automata.

Variable Neighborhood A new approach is to fix the states and the transition function and to have only the neighborhood (i. e. the relative localization of the entries of the transition function) varying. Somehow it can be considered that the neighborhood is not defined by the CA but by the configuration. Some simulation results exists [43,63]. The transition function of simulated CA is not given inside the simulating configuration but by the simulating neighborhood.

Future Directions

As mentioned just above, understanding the role played by the neighborhood in computing might be very enlightening.

It is known that 2 states and 2 neighbors is enough for computing with polynomial slowdown. It might be interesting to find constructions with limited slowdown and very concise encoding. Since CA are inherently parallel while Turing machines are sequential, a computing (and complexity) theory and algorithm that incorporate the parallelism of CA (local, uniform and synchronous) is worth enquiring.

As far as intrinsic universality is concerned, it relies on simulation between CA. The various definitions have to be linked and investigated.

Bibliography

1. Adamatzky A (ed) (2002) Collision based computing. Springer
2. Albert J, Čulik K II (1987) A simple universal cellular automaton and its one-way and totalistic version. Complex Syst 1:1–16
3. Arbib MA (1966) Simple self-reproducing universal automata. Inf Control 9(2):177–189
4. Arrighi P, Fargetton R (2007) Intrinsically universal one-dimensional quantum cellular automata. arXiv:0704.3961
5. Banks ER (1970) Universality in cellular automata. In: Eleventh annual symposium on switching and automata theory. IEEE, pp 194–215
6. Berlekamp E, Conway J, Guy R (1982) Winning ways for your mathematical plays (games in particular), vol 2. Academic Press
7. Burks A (1970) Essays on cellular automata. Univ of Illinois Press
8. Codd E (1968) Cellular automata. Academic Press
9. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15:1–40
10. Cordero P, Goles E, Hernández G (1992) Q2R + Q2R as a universal billiard. Int J Modern Phys C 3(2):251–266
11. Čulik K II, Pachl J, Yu S (1990) Computation theoretic aspects of cellular automata. Phys D 45(1–3):357–378
12. Delorme M, Mazoyer J (2002) Signals on cellular automata. In: Adamatzky A (ed) Collision-based computing. Springer, pp 234–275
13. Dubacq J-C (1995) How to simulate Turing machines by invertible 1d cellular automata. Int J Found Comput Sci 6(4):395–402
14. Durand B, Róka Z (1998) The game of life: Universality revisited. In: Delorme M, Mazoyer J (eds) Cellular Automata: A parallel model. Math Appl, vol 460. Kluwer, Dordrecht, pp 51–74
15. Durand-Lose J (1995) Reversible cellular automaton able to simulate any other reversible one using partitioning automata. In: LATIN'95. LNCS, vol 911. Springer, Berlin, pp 230–244
16. Durand-Lose J (1997) Intrinsic universality of a 1-dimensional reversible cellular automaton. In: STACS'97. LNCS, vol 1200. Springer, pp 439–450
17. Durand-Lose J (2000) Reversible space-time simulation of cellular automata. Theor Comput Sci 246(1–2):117–129
18. Durand-Lose J (2002) Computing inside the billiard ball model. In: Adamatzky A (ed) Collision-based computing. Springer, pp 135–160
19. Gardner M (1970) Mathematical games. The fantastic combinations of John Conway's new solitaire game "life". Sci Am 223:120–123
20. Gutowitz H (ed) (1991) Cellular automata, theory and experimentation. MIT/North-Holland
21. Hedlund GA (1969) Endomorphism and automorphism of the shift dynamical system. Math Syst Theory 3:320–375
22. Herrmann F, Margenstern M (2003) A universal cellular automaton in the hyperbolic plane. Theor Comput Sci 296:327–364
23. Hertling P (1998) Embedding cellular automata into reversible ones. In: Calude C, Casti J, Dinneen M (eds) Unconventional models of computation. DMTCS. Springer
24. Ilachinski A (2001) Cellular automata – A discrete universe. World Scientific
25. Imai K, Morita K (1998) A computation-universal two-dimensional 8-state triangular reversible cellular automaton. In: Margenstern M (ed) Universal machines and computations (UCM 98), vol 2. Université de Metz, pp 90–99
26. Kari J (2005) Theory of cellular automata: A survey. Theor Comput Sci 334:3–33
27. Kutrib M (2001) Efficient universal pushdown cellular automata and their application to complexity. In: Margenstern M, Rogozhin Y (eds) Machines, computations, and universality (MCU'01), Chisinau, Moldavia, 23–27 May 2001. LNCS, vol 2055. Springer, Berlin, pp 252–263
28. Lindgren K, Nordahl MG (1990) Universal computation in simple one-dimensional cellular automata. Complex Syst 4:299–318
29. Margenstern M (2006) An algorithm for buiding inrinsically universal automata in hyperbolic spaces. In: Arabnia HR, Burgin M (eds) International conference on foundations of computer science (FCS'06). pp 3–9
30. Margolus N (2002) Universal cellular automata based on the collisions of soft spheres. In: Adamatzky A (ed) Collision-based computing. Springer, pp 107–134
31. Martin B (1994) A universal cellular automaton in quasi-linear time and its S-n-m form. Theor Comput Sci 123:199–237
32. Mazoyer J (1996) Computations on one dimensional cellular automata. Ann Math Artif Intell 16:285–309
33. Mazoyer J, Rapaport I (1998) Inducing an order on cellular automata by a grouping operation. In: 15th annual symposium on theoretical aspects of computer science (STACS'98). LNCS, vol 1373. Springer, Berlin, pp 116–127
34. Mazoyer J, Rapaport I (1999) Inducing an order on cellular automata by a grouping operation. Discet Appl Math 218:177–196
35. Mazoyer J, Terrier V (1999) Signals in one-dimensional cellular automata. Theor Comput Sci 217(1):53–80
36. Minsky M (1967) Finite and infinite machines. Prentice Hall
37. Morita K (1992) Any irreversible cellular automaton can be simulated by a reversible one having the same dimension. Technical report of the IEICE Comp 92-45(1992-10):55–64
38. Morita K (1992) Computation-universality of one-dimensional one-way reversible cellular automata. Inform Process Lett 42:325–329
39. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comput Sci 148:157–163
40. Morita K, Harao M (1989) Computation universality of one-dimensional reversible (injective) cellular automata. Trans IEICE E 72(6):758–762
41. Morita K, Margenstern M, Imai K (1998) Universality of reversible hexagonal cellular automata. In: MFCS'98 Satellite workshop on frontiers between decidability and undecidability
42. Neary T, Woods D (2006) P-completeness of cellular automa-

ton rule 110. In: Bugliesi M, Preneel B, Sassone V, Wegener I (eds) International colloquium on automata languages and programming (ICALP'06). LNCS, vol 4051(1). Springer, Berlin, pp 132–143

43. Nishio H (2007) Changing the neighborhood of cellular automata. In: Durand-Lose J, Margenstern M (eds) Machine, computations and universality (MCU'07). LNCS, vol 4664. Springer, Berlin, pp 255–266
44. Ollinger N (2001) Two-states bilinear intrinsically universal cellular automata. In: FCT'01. LNCS, vol 2138. Springer, pp 369–399
45. Ollinger N (2002) The quest for small universal cellular automata. In: ICALP'02. LNCS, vol 2380. Springer, pp 318–329
46. Ollinger N (2003) The intrinsic universality problem of one-dimensional cellular automata. In: STACS'03. LNCS, vol 2607. Springer, Berlin, pp 632–641
47. Richard G (2006) A particular universal cellular automaton. oai:hal.ccsd.cnrs.fr:ccsd-00095821_v1
48. Richardson D (1972) Tessellations with local transformations. J Comput Syst Sci 6:373–388
49. Róka Z (1999) Simulation between cellular automata on cayley graphs. Theor Comput Sci 225:81–111
50. Sarkar P (2000) A brief history of cellular automata. ACM Comput Surv 32(1):80–107
51. Sipser M (1997) Introduction to the theory of computation. PWS Publishing Co, Boston
52. Smith AR III (1971) Simple computation-universal cellular spaces. J ACM 18(3):339–353
53. Steiglitz K, Kamal I, Watson A (1988) Embedding computation in one-dimensional automata by phase coding solitons. IEEE Trans Comput 37(2):138–145
54. Sutner K (2005) Universality and cellular automata. In: Margenstern M (ed) Machines, computations, and universality (MCU'04). LNCS, vol 3354. Springer, Berlin, pp 50–59
55. Terrier V (1996) Language not recognizable in real time by one-way cellular automata. Theor Comput Sci 156:281–287
56. Theyssier G (2004) Captive cellular automata. In: Fiala J, Koubek V, Kratochvíl J (eds) Mathematical foundations of computer science (MFCS'04), 29th international symposium, Prague, Czech Republic, 22–27 August. LNCS, vol 3153. Springer, Berlin, pp 427–438
57. Theyssier G (2005) How common can be universality for cellular automata? In: Diekert V, Durand B (eds) 22nd annual symposium on theoretical aspects of computer science (STACS'05), Stuttgart, Germany, 24–26 February. LNCS, vol 3404. Springer, Berlin, pp 121–132
58. Toffoli T, Margolus N (1987) Cellular automata machine – A new environment for modeling. MIT press, Cambridge
59. Ulam S (1952) Random processes and transformations. In: International congress of mathematics 1950, vol 2. pp 264–275
60. von Neumann J (1966) Theory of self-reproducing automata. University of Illinois Press, Urbana
61. Wolfram S (2002) A New kind of Science. Wolfram Media
62. Woods D, Neary T (2007) Small semi-waekly universal turing machines. In: Durand-Lolse J, Margenstern M (eds) Machines, computations and universality (MCA'07). LNCS, vol 4664. Springer, Berlin, pp 246–257
63. Worsch T, Nishio H (2007) Variations on neighborhoods in ca. In: Moreno-Diaz R (ed) EUROCAST'07. LNCS, vol 4739. Springer, Berlin, pp 581–588

Cellular Automaton Modeling of Tumor Invasion

HARALAMBOS HATZIKIROU[1], GEORG BREIER[2], ANDREAS DEUTSCH[1]

[1] Center for Information Services and High Performance Computing, Technische Universität Dresden, Dresden, Germany

[2] Institute of Pathology, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

Article Outline

Glossary

Cadherins Important class of transmembrane proteins. They play a significant role in cell-cell adhesion, ensuring that cells within tissues are bound together.

Chemotaxis Motion response to chemical concentration gradients of a diffusive chemical substance.

Extracellular matrix (ECM) Components that are extracellular and composed of secreted fibrous proteins (e. g. collagen) and gel-like polysaccharides (e. g. glycosaminoglycans) binding cells and tissues together.

Fiber tracts Bundle of nerve fibers having a common origin, termination, and function and especially one within the spinal cord or brain.

Haptotaxis Directed motion of cells along adhesion gradients of fixed substrates in the ECM, such as integrins.

"Slime trail motion" Cells secrete a non-diffusive substance; concentration gradients of the substance allow the cells to migrate towards already explored paths.

Somatic evolution Darwinian-type evolution that occurs on soma (as opposed to germ) cells and characterizes cancer progression [1].

Definition of the Subject

Cancer cells acquire characteristic traits in a step-wise manner during carcinogenesis. Some of these traits are autonomous growth, induction of angiogenesis, invasion and metastasis. In this chapter, the focus is on one of the late stages of tumor progression: tumor invasion. Tumor invasion has been recognized as a complex system, since its behavior emerges from the combined effect of tumor cell-cell and cell-microenvironment interactions. Cellular automata (CA) provide simple models of self-organizing complex systems in which collective behavior can emerge out of an ensemble of many interacting "simple" components. Recently, cellular automata have been used to gain a deeper insight in tumor invasion dynamics. In this chapter, we briefly introduce cellular automata as models of tumor invasion and we critically review the most prominent CA models of tumor invasion.

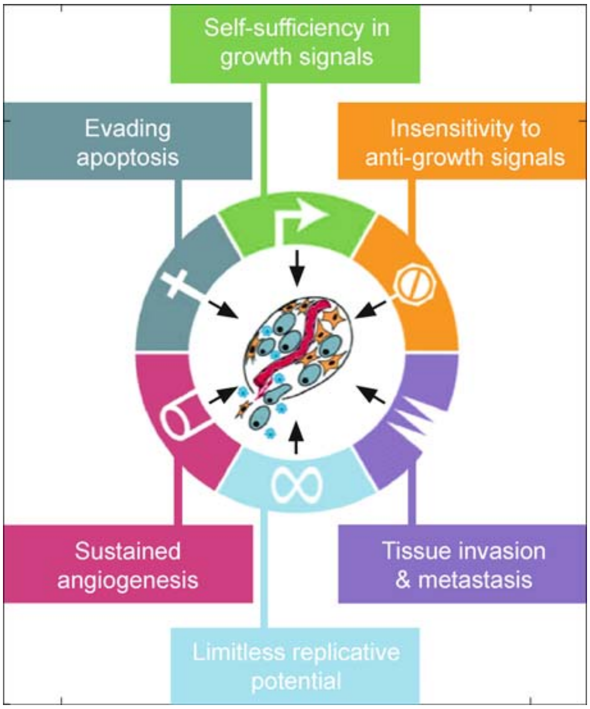
Introduction

Cancer describes a group of genetic and epigenetic diseases characterized by uncontrolled growth of cells, leading to a variety of pathological consequences and frequently death. Cancer has long been recognized as an evolutionary disease [2]. Cancer progression can be depicted as a sequence of traits or phenotypes that cells have to acquire if a neoplasm (benign tumor) is to become an invasive and malignant cancer. A phenotype refers to any kind of observed morphology, function or behavior of a living cell. Hanahan and Weinberg [3] have identified six cancer cell phenotypes: unlimited proliferative potential, environmental independence for growth, evasion of apoptosis, angiogenesis, invasion and metastasis.

In this chapter, we focus on the invasive phase of tumor growth. Invasion is the main feature that allows a tumor to be characterized as malignant. The progression of a benign tumor and delimited growth to a tumor that is invasive and potentially metastatic is the major cause of poor clinical outcome in cancer patients, in terms of therapy and prognosis. Understanding tumor invasion could potentially lead to the design of novel therapeutical strategies. However, despite the immense amounts of funds invested in cancer research, the intracellular and extracellular dynamics that govern tumor invasiveness in vivo remain poorly understood.

Biomedically, invasion involves the following tumor cell *processes*:

- Tumor cell migration, which is a result of down-regulation of cadherins, that is loss of cell-cell adhesion,
- Tumor cell-extracellular matrix (ECM) interactions, such as cell-ECM adhesion, and ECM degradation or remodeling, by means of proteolysis. These processes allow for the penetration of the migrating tumor cells into host tissue barriers, such as basement and interstitial stroma [4], and
- Tumor cell proliferation.


Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Cellular Automaton Modeling of Tumor Invasion, Figure 1
Hanahan and Weinberg have identified six possible types of cancer cell phenotypes: unlimited proliferative potential, environmental independence for growth, evasion of apoptosis, angiogenesis, invasion and metastasis (Reprinted from [3], with permission from the authors)

Tumor invasion facilitates the emergence of metastases, i. e. the spread of cancer cells to another part of the body and the formation of secondary tumors. Tumor invasion comprises a central aspect in cancer progression. However, invasive phenomena occur not only in pathological cases of malignant tumors but also during normal morphogenesis and wound healing.

Cancer research has been directed towards the understanding of tumor invasion dynamics and its implications in treatment design. In particular, research concentrates along the following *problems*:

Invasive tumor morphology: A wealth of empirical evidence links disease progression with tumor morphology [5]. The tumor morphology can indicate the degree of a tumor's malignancy. In particular, it is experimentally and clinically observed that a morphological instability is related to invasive solid tumors, producing finger-like spatial patterns. The question is which molecular and cellular mechanisms are responsible for this spatial pattern formation.

Cell migration and influence of the ECM: Important aspects of invading tumors are cell motion and the effect of the surrounding environment, especially the ECM [4].

Metabolism and acidosis: The multi-step process of carcinogenesis is often described by somatic evolution, wherein phenotypic properties are retained or lost depending on their contribution to the individual tumor cell survival and reproductive potential. One of the most prominent phenotypic changes involves the anaerobic glucose metabolism (glycolysis). A side-product of this metabolic activity is the production of H^+ ions that increase the pH of the tumor's microenvironment (acidosis). This gives rise to the questions: (i) Why does tumor evolution lead to this kind of metabolism, which is energetically deficient in comparison with the aerobic one? (ii) What are the advantages for the tumor? (iii) How do glycolytic tumor cells influence tumor invasion?

Emergence of invasion: Typically, tumor invasion appears during the late stages of carcinogenesis. Of ultimate importance is the question what are the mechanisms and the environmental conditions that trigger the progression from benign neoplasms to malignant invasive tumors.

Robustness: There are several questions concerning the stability and the resistance of tumor invasion such as: (i) Why are malignant tumor so robust (resistant) to perturbations (i. e. therapies)? (ii) Is it possible to design intelligent therapies (at the cellular level) that disturb the tumor's robustness? (iii) How can we investigate the tumor's robustness?

Mathematical modeling and analysis provide invaluable tools towards answering the above questions. Tumor invasion involves processes which occur at different spatio-temporal scales, including processes at the subcellular, cellular and tissue level. Mathematical models allow description and linking of these levels. One can distinguish *molecular*, *cellular* and *tissue* scales, respectively [6,7]:

- The molecular scale refers to phenomena at the subcellular level and concentrates on molecular interactions and resulting phenomena, such as alterations of signaling cascades and cell cycle control, gene mutations, etc. In tumor invasion, the down-regulation of cadherins provides an example of a molecular process.
- The cellular scale refers to cellular interactions and therefore to the most prominent dynamics of cell populations, e. g. adhesion, contact inhibition, chemotaxis etc.
- The tissue scale focuses on tissue level processes taking into account macroscopic quantities, such as volumes, flows etc. Continuum phenomena include cell convection and diffusion of nutrients and chemical factors, mechanical stress and the diffusion of metastases.

For example, genetic alterations may lead to invasive cells (molecular scale) that are able to migrate (cellular scale) and interact with diffusible or non-diffusible signals (tissue scale). Models that deal with phenomena at multiple scales are called multi-scaled.

Recently, a variety of mathematical models have been proposed to analyze different aspects of tumor invasion. Deterministic macroscopic models are used to model the spatio-temporal growth of tumors, usually assuming that tumor invasion is a wave propagation phenomenon [8,9,10,11,12]. Computational investigations of the invasiveness of glioma tumors illustrate that the ratio of tumor growth and spatial anisotropy in cell motility can quantify the degree of tumor invasiveness [13,14]. Whilst these models are able to capture the tumor structure at the tissue level, they fail to describe the tumor at the cellular and the sub-cellular levels. Lately, multi-scale approaches attempt to describe and predict invasive tumor morphologies, growth and phenotypical heterogeneity [15,16].

Cellular automata (CA), and more generally cell-based models, provide an alternative modeling approach, where a micro-scale investigation is allowed through a stochastic description of the dynamics at the cellular level [17]. In particular, CA define an appropriate modeling framework for tumor invasion since they allow for the following:

- CA rules can mimic the processes at the cellular level. This fact allows for the modeling of an abundance of experimental data that refer to cellular and sub-cellular processes related to tumor invasion.
- The discrete nature of CA can be exploited for investigations of the boundary layer of a tumor. Bru et al. [18] have analyzed the fractal properties of tumor surfaces (calculated by means of fractal scaling analysis) which can be compared with corresponding CA simulations to gain a better understanding of the tumor phenomenon. In addition, the discrete structure of CA facilitates the implementation of complicated environments without any of the computational problems characterizing the simulation of continuous models.
- Motion of tumor cells through heterogeneous media (e. g. ECM) involves phenomena at various spatial and temporal scales [19]. These cannot be captured in a purely macroscopic modeling approach. Alternatively, discrete microscopic models, such as CA, can in-

corporate different spatio-temporal scales and they are well-suited for simulating such phenomena.

- CA are paradigms of parallelizable algorithms. This fact makes them computationally efficient.

In the following section, we provide a definition of CA. In Sect. "Models of Tumor Invasion", we review the existing CA models for central processes of tumor invasion. Finally, the discussion, we critically discuss the use of CAs in tumor invasion modeling and we identify future research questions related to tumor invasion.

Cellular Automata

The notion of a *cellular automaton* originated in the works of John von Neumann (1903–1957) and Stanislaw Ulam (1909–1984). Cellular automata may be viewed as simple models of self-organizing complex systems in which collective behavior can emerge out of an ensemble of many interacting "simple" components. In complex systems, even if the basic and local interactions are perfectly known, it is possible that the global behavior obeys new laws that cannot be obviously extrapolated from the individual properties, as if the whole is more than the sum of the parts. This property makes cellular automata a very interesting approach to model complex systems in physics, chemistry and biology (examples are introduced in [17,20]). A CA can be defined as a 4-tuple $(\mathcal{L}, \mathcal{S}, \mathcal{N}, \mathcal{F})$, where:

- $\mathcal{L}$ is an infinite regular *lattice* of nodes (discrete space),
- $\mathcal{S}$ is a finite set of *states* (discrete states); each cell $i \in \mathcal{L}$ is assigned a state $s \in \mathcal{S}$,
- $\mathcal{N}$ is a finite set of *neighbors*,
- $\mathcal{F}$ is a deterministic or probabilistic map

$$\mathcal{F} : \mathcal{S}^{|\mathcal{N}|} \to \mathcal{S} \tag{1}$$

$$\{s_i\}_{i \in \mathcal{N}} \mapsto s \,, \tag{2}$$

which assigns a new state to a node depending on the state of all its neighbors indicated by $\mathcal{N}$ (*local rule*).

The evolution of a CA is defined by applying the function $\mathcal{F}$ synchronously to all nodes of the lattice $\mathcal{L}$ (*homogeneity* in space and time).

The above features can be extended, giving rise to several variants of the classical CA notion [21]. Some of these are:

Asynchronous CA: in such CA, the restriction of simultaneous update of all the nodes is revoked, allowing for asynchronous update.

Non-homogeneous CA: this variation allows the transition rules to depend on node position. Agent-based models are "relatives" of CA that lost the homogeneity property, i. e. each individual-particle may have its own set of rules.

Coupled-map lattices: in this case the constraint of discrete state space is withdrawn, i. e. the state variables are assumed to be continuous. An important type of coupled-map lattices are the so-called Lattice Boltzmann models [22].

Structurally dynamic CA: in these systems, the underlying lattice is no longer a passive static object but becomes a dynamic component. Therefore, the lattice, structure evolves depending on the values of the nodes state variables.

Models of Tumor Invasion

This section reviews the existing cellular automata models of tumor invasion. Categorizing these models is a nontrivial task. Moreover, existing CA models describe tumor invasion at more than one scale (sub-cellular, cellular and tissue). In this review, we distinguish models that analyze: (i) the invasive morphology, (ii) tumor cell migration and the influence of the ECM, (iii) metabolism and acidosis, and (iv) the emergence of tumor invasion.

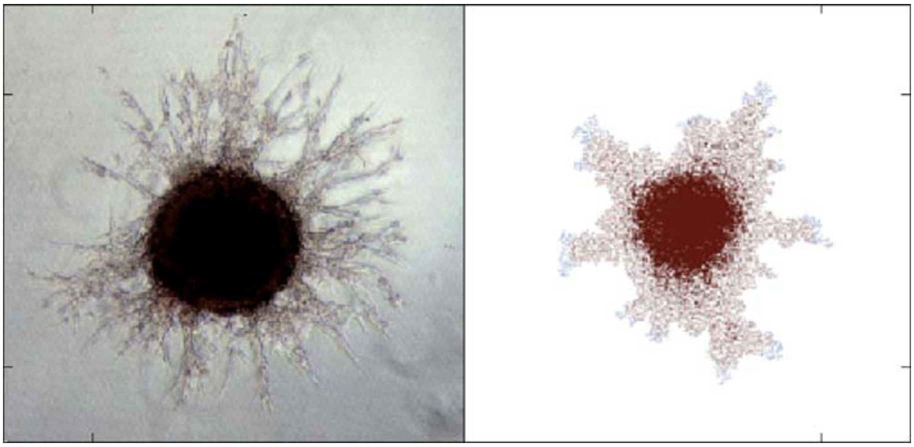
Invasive Tumor Morphology

The tumor morphology arising from the spatial pattern formation of the tumor cell population has been recognized as a very important aspect of tumor growth. Several researchers have attempted to reveal the mechanisms of spatial pattern formation of invasive tumors. Here, we present the most representative CA models for the invasive tumor morphology.

Effects of Directed Cell Motion Sander et al. [23] developed a CA model to investigate the branching morphology of invasive brain tumors. In the model tumor cell motion is influenced by two key processes: (i) chemotaxis, and (ii) "slime trail following". A typical example of a slime trail following mechanism is found in the motion of certain myxobacteria [24].

The authors show that the branching morphology of tumors can be explained as a result of chemotaxis and "slime trail following". In particular, simulations reproduce the branching pattern formation observed in vitro cultures of glioma cells. However, the assumption of slime trail following has not been proven biologically as yet.

Spatial Structure of Invasive Tumors Anderson [15,25] proposed a model to examine the effects of tumor cell heterogeneity (at the genetic level) on the spatial morphol-



Cellular Automaton Modeling of Tumor Invasion, Figure 2

***Left*: Microscopy image of a multicellular tumor spheroid, exhibiting an extensive branching system that rapidly expands into the surrounding extracellular matrix gel. These branches consist of multiple invasive cells. (Reprinted from [26] with permission). *Right*: Simulation of Anderson's model [15] reproducing the experimentally observed morphology of invasive tumors**

ogy and to analyze the importance of cell-cell and cell-ECM adhesion. The model assumes a non-diffusible, fixed configuration of ECM. The extracellular matrix can be degraded by diffusible enzymes, such as metallo-proteinases, produced by tumor cells. Moreover, cells are allowed to mutate and evolve their phenotype from proliferative to invasive. Finally, an oxygen concentration field plays the role of nutrients in the model.

Simulations of the model show that: (i) the ECM heterogeneity is mainly responsible for the tumor branching morphology (Fig. 2), (ii) cell-cell adhesion plays an important role only in the early stages of tumor development, (iii) invasive tumor cells are located at the boundary of the tumor, and (iv) the tumor is a phenotypically heterogeneous object.

Tumor Cell Migration and the Influence of the Extracellular Matrix

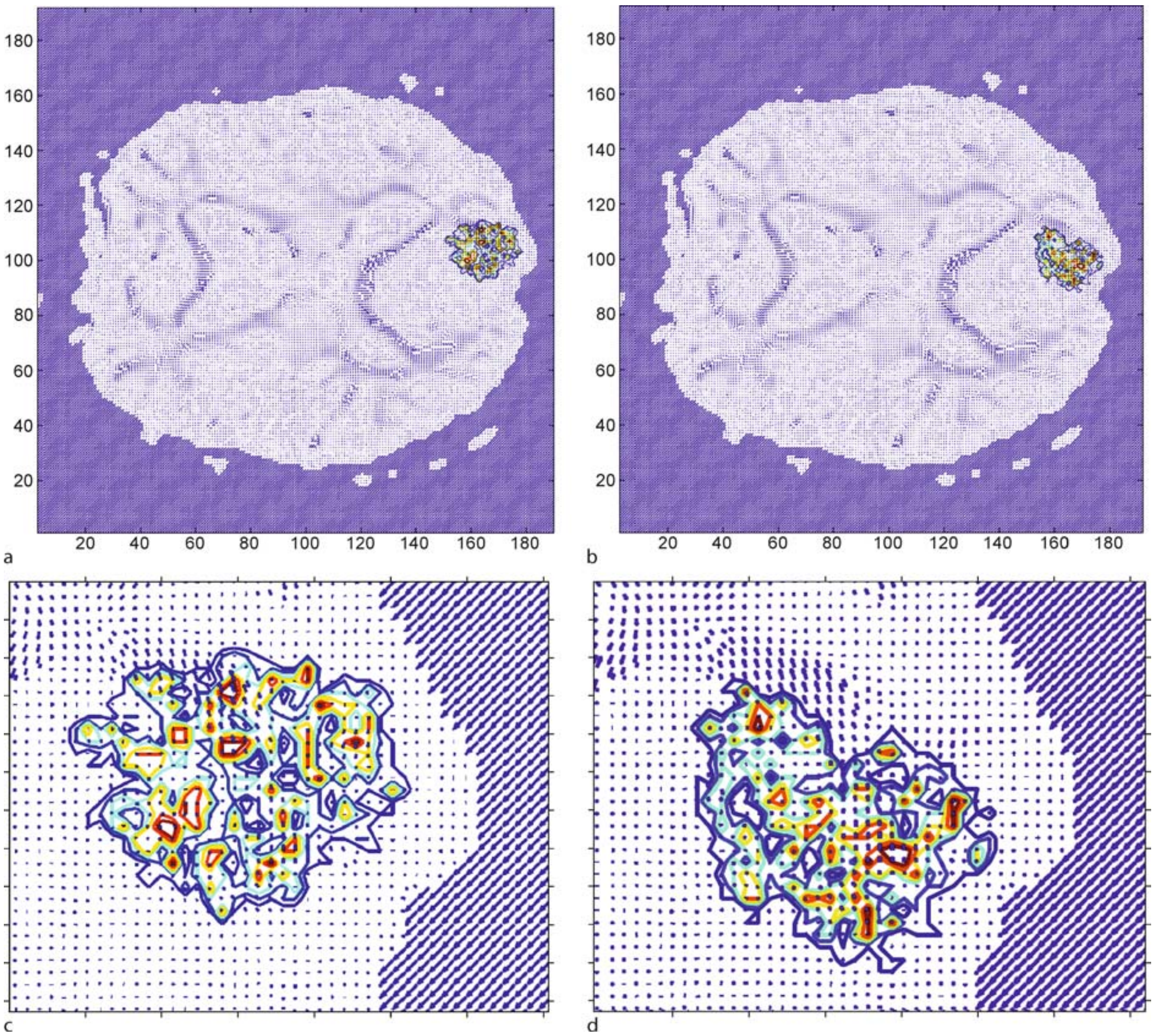
Cell migration and cell-ECM interactions are two of the most crucial invasion-related processes. Cellular automata provide an appropriate framework to model and analyze the effect of cell motility and cell-environment interactions of tumor cell migration.

The Role of Cell-Cell and Cell-ECM Adhesion Turner and Sheratt [27] proposed a cellular Potts model [28] to investigate how cell-cell and cell-ECM adhesion influence the tumor invasion depth and tumor morphology. A cellular Potts model can be viewed as an extension of the CA idea allowing to analyze phenomena that take into account specific cell shapes. Cells are assumed to move according to intercellular adhesive interactions and haptotactical gradients. Moreover, cells are allowed to proliferate, while mitotic probabilities depend on the strength of the adhesive interaction. Finally, cells are assumed to secrete proteolytic enzymes that degrade the ECM.

The authors show that adhesive dynamics can explain the "fingering" patterns observed in their simulations. Moreover, the authors demonstrate that the width of the invasion zone depends less on cell-cell adhesion and more on cell-ECM adhesion facilitated by haptotaxis and proteolysis.

Cellular Mechanisms of Glioma Cell Migration In the work of Aubert et al. [29], a CA model is introduced that allows for the investigation of tumor cell migration, based on experimentally observed density profiles of glioma cell cultures. The goal is to identify the mechanisms of tumor (glioma) cell motion, which play a crucial role in tumor invasion. The authors do not consider proliferation of tumor cells. Only the influence of tumor cell migration and intercellular interactions are studied. The authors introduce and test two distinct cell mechanisms: (i) cell-cell adhesion, and (ii) a kind of "inertia" in cell motion, i. e. the cells tend to maintain the direction of their motion.

The authors carefully scale the model according to the experimental setup and calibrate the corresponding model parameters. The simulation results indicate that cell-cell adhesion can explain the experimental results. It is concluded that cell-cell adhesion is an important process in glioma cell migration.



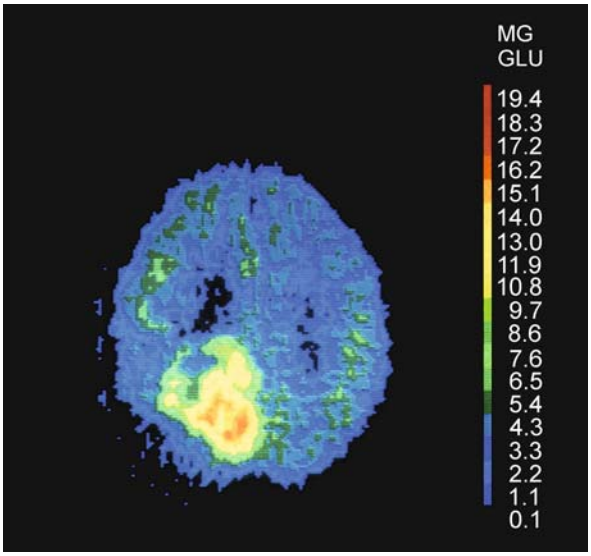
Cellular Automaton Modeling of Tumor Invasion, Figure 3

The effect of the brain's fiber tract on glioma growth. a A simulation is shown without taking into account the influence of fiber tracts. b The fiber tracts in the brain strongly drive the evolution of the tumor growth. c–d Figures display a close up of the tumor area of the top a–b simulations (Reprinted from [31])

Effects of Fiber Tracts on Glioma Invasion Wurzel et al. [30] model glioma tumor invasion with a lattice-gas cellular automaton (LGCA) [17]. The authors address the question of how fiber tracts found in the brain's white matter influence the spatio-temporal evolution and the invading front morphology of glioma tumors. Cells are assumed to move, proliferate and undergo apoptosis according to corresponding stochastic processes. Fiber tracts are represented as a local gradient field that enhances cell motion in a specific direction.

The authors develop and analyze different scenarios of fiber tract influence. A gradient field may increase the speed of the invading tumor front. For high field intensities the model predicts the formation of cancer islets at distances away from the main tumor bulk. The simulated invasion patterns qualitatively resemble clinical observations.

Effect of Heterogeneous Environments on Tumor Cell Migration Hatzikirou et al. [31] developed a LGCA model to investigate the influence of heterogeneous environments on tumor cell dispersal. This model is a simplified version of [30] which facilitates the mathematical analysis. In this study no proliferation or death of cells



Cellular Automaton Modeling of Tumor Invasion, Figure 4
Typically, tumors exhibit abnormal levels of glucose metabolism. Positron emission imaging (PET) techniques localize the regions of abnormal glycolytic activity and identify the tumor locus

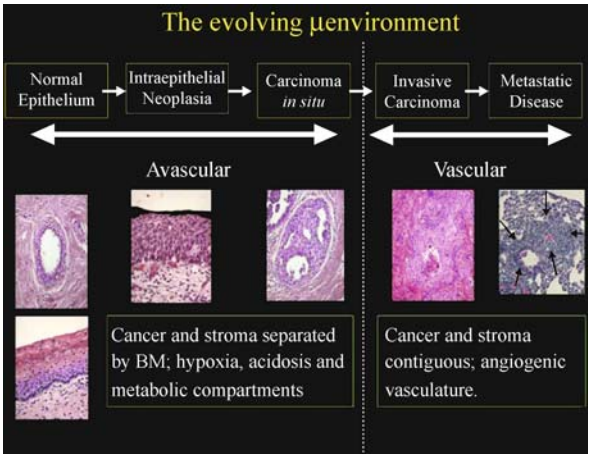
is considered. The authors distinguish two kinds of cell-ECM interactions (i) cell-ECM adhesion leading to haptotactical motion along integrin concentration gradients (environment with directional information) and (ii) contact guidance that promotes the alignment along ECM pores or fibres as seen in Fig. 3 (environment with orientational information).

In this study, it is investigated the impact of both types of cell-ECM interaction on tumor cell motion. In particular, macroscopic dispersal measures (such as mean cell flux) depending on cellular and environmental parameters are calculated. Accordingly, the models allow for prediction of cell motion in different environments.

Metabolism and Acidosis

In the course of cancer progression, tumor cells undergo several phenotypic changes in terms of motility, metabolism and proliferative rates. In particular, it is important to analyze the effect of the anaerobic metabolism of tumor cells and the acidification of the environment (as a side-product of glycolysis) on tumor invasion (Fig. 4).

Patel et al. [32] proposed a model of tumor growth to examine the roles of native tissue vascularity and anaerobic metabolism on the growth and invasion efficacy of tumors. The model assumes a vascularized host tissue. Anaerobic metabolism involves the consumption of glucose and the production of H^+ ions, leading to the acidification of the local tissue. The vascular network allows for the absorption of H^+ ions. Cells are assumed to be proliferative and non-motile. The pH level, i. e. is the H^+ concentration, and the glucose concentration determine the survival and death of the cells.


The evolving μenvironment

Normal Epithelium → Intraepithelial Neoplasia → Carcinoma *in situ* → Invasive Carcinoma → Metastatic Disease

Avascular

Vascular

Cancer and stroma separated by BM; hypoxia, acidosis and metabolic compartments

Cancer and stroma contiguous; angiogenic vasculature.

Cellular Automaton Modeling of Tumor Invasion, Figure 5
The evolving microenvironment of breast cancer. The multiple stages of breast carcinogenesis are shown progressing from left to right, along with histological representations of these stages. As indicated, the pre-invasive stages occur in an avascular environment, whereas cancer cells have direct access to vasculature following invasion. (Reprinted from [33])

Simulations of the model show: (i) high tumor H^+ ion production favors tumor invasion by the acidification of the neighboring host tissue, and (ii) there is an optimal density of microvessels that maximizes tumor growth and invasion, by minimizing the acidification effects on tumor cell proliferation (absorption of H^+ ions) and maximizing the negative effect of H^+ ions on the neighboring tissue.

Emergence of Tumor Invasion

Recently, several models have been proposed that concentrate on the evolutionary dynamics of tumors (Fig. 5). The main goal of these models is to understand under which environmental conditions particular phenotypes appear. Here, we review those models that focus on the mechanisms that allow the emergence of invasive behavior.

Influence of Metabolic Changes Smallbone et al. [34] developed an evolutionary CA model to investigate the cell-microenvironmental interactions that mediate somatic evolution of cancer cells. In particular, the authors investigate the sequence of tumor phenotypes that ultimately leads to invasive behavior. The model considers three phenotypes, (i) the hyperplastic phenotype that allows growth away from the basement membrane, (ii) the glycolytic phenotype that allows anaerobic metabolism (the "fuel" is glucose), and (iii) the acid-resistant phenotype that enables the cell to survive in low pH. Cells are allowed to proliferate, die or adapt (change their phenotype). No cell motion is explicitly considered.

The model predicts three phases of somatic evolution: (i) Initially, cell survival and proliferation are dependent on the oxygen concentration. (ii) When the oxygen becomes scarce, the glycolytic phenotype confers a significant proliferative advantage. (iii) The side-products of glycolysis, e. g. galactic acid, increase the microenvironmental pH and promote the selection of acid-resistant phenotypes. The latter cell type is able to invade the neighboring tissue since it takes advantage of the death of host cells, due to acidification, and proliferates using the available free space.

The Game of Invasion Recently, Basanta et al. [35] have developed a game theory inspired CA that addresses the question of how invasive behavior emerges during tumor progression (see also [36]). The authors study the circumstances under which mutations that confer increased motility to cells can spread through a tumor composed of rapidly proliferating cells. The model assumes the existence of only two phenotypes: "proliferative" (high division rate and no motility) and "migratory" (low division rate and high motility). Mutations are allowed for by the random change of phenotypes. Nutrients are assumed to be uniformly distributed over the lattice.

Simulations show that low nutrient conditions confer a reproductive advantage to motile cells over the proliferative ones. The model suggests novel ideas for therapeutic strategies, e. g. by increasing the oxygen supply around the tumor to favor the reproduction of proliferative cells over the migrating ones. This is not necessarily a therapy since there are benign tumors that are life threatening even if they do not become invasive. Despite that, in most cases a growing but non-aggressive tumor will have a much better prognosis than a smaller but invasive one.

Discussion

In this review, we have focused on one of the most important aspects of cancer progression: tumor invasion. The main processes involved in tumor invasion are related to tumor cell migration, cell-ECM interactions, especially ECM degradation/remodeling and tumor cell proliferation. These processes are evolving at different scales, e. g. cell-ECM adhesion is the response of tumor cells to ECM integrins (molecular level) leading to a haptotactical cell motion (cellular level) and influencing the tumor morphology (macroscopic level). Therefore, in order to understand tumor invasion dynamics, it is important to use mathematical tools that allow for modeling sub-cellular or cellular processes and to analyze the emergent macroscopic behavior. Individual-based models, especially CA, are well-suited for this task. Moreover, some types of CA models, such as lattice-gas cellular automata [17,31], facilitate analytical investigations allowing for deeper insight into the modeled phenomena.

In this chapter, we reviewed the existing CA models of tumor invasion. The presented models explore central aspects of tumor invasion. Some of the models are in good agreement with biomedical observations for in-vitro and in-vivo tumors. In the following, we list the most interesting biological insights that can be gained from the reviewed models:

- The significance of hypoxia in the process of tumor progression: Activation of glycolysis and acidification of the host tissue facilitate tumor invasion. Low nutrient conditions, such as hypoxia, may trigger invasive behaviors.
- Cell-cell adhesion: It is evident that intercellular adhesion has a great impact in the early stages of tumor growth. However, in tumor invasion the role of cell-cell adhesion is minor, since mainly the cell-ECM interactions appear to dictate the tumor cell behavior.
- Cell-ECM adhesion: This is an important process for tumor invasion. In particular, the heterogeneous structure of the ECM strongly influences the spatial morphology of invasive tumors.

Mathematical modeling offers potentially significant insight into tumor invasion. Several crucial questions have not been adequately addressed so far by modeling efforts:

Branching morphology: Several mechanisms have been proposed that lead to branching patterns, e. g. diffusion-limited aggregation, the interplay of cell-cell and cell-ECM adhesion, as well as chemotaxis or slime trail following motion. However, biologists and modelers have not yet identified a unique mechanism that drives the branching morphology of invasive tumors.

Go or grow: The mechanisms of invasive tumor cell migration are still not understood. Recently, Fedotov et al. [37] have analyzed the effect of a postulated migration/proliferation dichotomy on cell migration.

Emergence: Concerning the emergence of invasion in tumor progression little is known. Mechanisms related to tumor cell motion and other cell processes, such as proliferation (migration/proliferation dichotomy), may play an important role for the dominance of invasive phenotypes [35,38].

Angiogenesis: Another open issue is the influence of angiogenesis and vasculogenesis on tumor invasion. Despite significant efforts to describe the mechanisms of angio- and vasculogenesis, little is known about the effect of these processes on tumor invasion [32].

Robustness: The identification of cellular mechanisms that are responsible for tumor robustness remains significant challenge.

Finally, for clinical purposes, future models should be able to provide accurate and quantitative predictions. Simplified models considering only the essential ingredients for tumor growth, and especially tumor invasion, but validated with actual clinical data may be helpful in this regard. We sincerely hope that a more profound knowledge of important tumor characteristics, such as tumor invasion, will eventually lead to the design of more effective therapeutic strategies.

Acknowledgments

We are grateful to D. Basanta, L. Brusch, A. Chauviere, E. Flach and F. Peruani for the comments and the fruitful discussions. We acknowledge support from the systems biology network HepatoSys of the German Ministry for Education and Research through grant 0313082J. Andreas Deutsch is a member of the DFG Research Center for Regenerative Therapies Dresden – Cluster of Excellence – and gratefully acknowledges support by the Center. The research was supported in part by funds from the EU Marie Curie Network "Modeling, Mathematical Methods and Computer Simulation of Tumor Growth and Therapy" (EU-RTD IST-2001-38923). Finally, the authors would like to thank for the financial support of the Gottfried Daimler- and Karl Benz foundation through the project "Biologistics: From bio-inspired engineering of complex logistical systems until nanologistics" (25-02/07).

Bibliography

1. Bodmer W (1997) Somatic evolution of cancer cells. J R Coll Physicians Lond 31(1):82–89
2. Nowell PC (1976) The clonal evolution of tumor cell populations. Science 4260 194:23–28
3. Hanahan D, Weinberg R (2000) The hallmarks of cancer. Cell 100:57–70
4. Friedl P (2004) Prespecification and plasticity: shifting mechanisms of cell migration. Curr Opin Cell Biol 16(1):14–23
5. Sanga S, Frieboes H, Zheng X, Gatenby R, Bearer E, Cristini V (2007) Predictive oncology: multidisciplinary, multi-scale in-silico modeling linking phenotype, morphology and growth. Neuroim 37(1):120–134
6. Hatzikirou H, Deutsch A, Schaller C, Simon M, Swanson K (2005) Mathematical modelling of glioblastoma tumour development: a review. Math Mod Meth Appl Sc 15(11):1779–1794
7. Preziozi L (ed) (2003) Cancer modelling and simulation. Chapman & Hall CRC Press
8. Marchant BP, Norbury J, Perumpanani AJ (2000) Traveling shock waves arising in a model of malignant invasion. SIAM J Appl Math 60(2):263–276
9. Perumpanani AJ, Sherratt JA, Norbury J, Byrne HM (1996) Biological inferences from a mathematical model of malignant invasion. Invas Metast 16:209–221
10. Perumpanani AJ, Sherratt JA, Norbury J, Byrne HM (1999) A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cellular invasion. Phys D 126:145–159
11. Sherratt JA, Nowak MA (1992) Oncogenes, anti-oncogenes and the immune response to cancer: a mathematical model. Proc Roy Soc Lond B 248:261–271
12. Sherratt JA, Chaplain MAJ (2001) A new mathematical model for avascular tumour growth. J Math Biol 43:291–312
13. Swanson KR, Alvord EC, Murray J (2002) Quantifying efficacy of chemotherapy of brain tumors (gliomas) with homogeneous and heterogeneous drug delivery. Acta Biotheor 50:223–237
14. Jbabdi S, Mandonnet E, Duffau H, Capelle L, Swanson K, Pelegrini-Issac M, Guillevin R, Benali H (2005) Simulation of anisotropic growth of low-grade gliomas using diffusion tensor imaging. Magn Res Med 54:616–624
15. Anderson A, Weaver A, Cummings P, Quaranta V (2006) Tumor morphology and phenotypics evolution driven by selective pressure from the microenvironment. Cell 127:905–915
16. Frieboes H, Lowengrub J, Wise S, Zheng X, Macklin P, Bearer E, Cristini V (2007) Computer simulation of glioma growth and morphology. Neuroim 37(1):59–70
17. Deutsch A, Dormann S (2005) Cellular Automaton Modeling of Biological Pattern Formation. Birkhauser, Boston
18. Bru A, Albertos S, Subiza JL, Lopez Garcia-Asenjo J, Bru I (2003) The universal dynamics of tumor growth. Bioph J 85:2948–2961
19. Lesne A (2007) Discrete vs continuous controversy in physics. Math Struct Comp Sc 17:185–223
20. Chopard B, Dupuis A, Masselot A, Luthi P (2002) Cellular automata and lattice Boltzmann techniques: an approach to model and simulate complex systems. Adv Compl Syst 5(2):103–246
21. Moreira J, Deutsch A (2002) Cellular automaton models of tumour development: a critical review. Adv Compl Syst 5:1–21
22. Succi S (2001) The lattice Boltzmann equation: for fluid dynamics and beyond. Series Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford
23. Sander LM, Deisboeck TS (2002) Growth patterns of microscopic brain tumours. Phys Rev E 66:051901
24. Wolgemuth CW, Hoiczyk E, Kaiser D, Oster GF (2002) How myxobacteria glide. Curr Biol 12(5):369–377
25. Anderson ARA (2005) A hybrid model of solid tumour invasion: the importance of cell adhesion. Math Med Biol 22:163–186
26. Habib S, Molina-Paris C, Deisboeck TS (2003) Complex dynamics of tumors: modeling an emerging brain tumor system with coupled reaction-diffusion equations. Phys A 327:501–524
27. Turner S, Sherratt JA (2002) Intercellular adhesion and cancer invasion: A discrete simulation using the extended Potts model. J Theor Biol 216:85–100

28. Graner F, Glazier J (1992) Simulation of biological cell sorting using a two-dimensional extended Potts Model. Phys Rev Lett 69:2013–2016
29. Aubert M, Badoual M, Freol S, Christov C, Grammaticos B (2006) A cellular automaton model for the migration of glioma cells. Phys Biol 3:93–100
30. Wurzel M, Schaller C, Simon M, Deutsch A (2005) Cancer cell invasion of normal brain tissue: Guided by Prepattern? J Theor Med 6(1):21–31
31. Hatzikirou H, Deutsch A (2008) Cellular automata as microscopic models of cell migration in heterogeneous environments. Curr Top Dev Biol 81:401–434
32. Patel A, Gawlinski E, Lemieux S, Gatenby R (2001) Cellular automaton model of early tumor growth and invasion: the effects of native tissue vascularity and increased anaerobic tumor metabolism. J Theor Biol 213:315–331
33. Gillies RJ, Gatenby RA (2007) Hypoxia and adaptive landscapes in the evolution of carcinogenesis. Canc Metast Rev 26:311–317
34. Smallbone K, Gatenby R, Gillies R, Maini P, Gavaghan D (2007) Metabolic changes during carcinogenesis: Potential impact on invasiveness. J Theor Biol 244:703–713
35. Basanta D, Hatzikirou H, Deutsch A (2008) The emergence of invasiveness in tumours: a game theoretic approach. Eur Phys J B 63:393–397
36. Basanta D, Simon M, Hatzikirou H, Deutsch A (2009) An evolutionary game theory perspective elucidates the role of glycolysis in tumour invasion. Cell Prolif (to appear)
37. Fedotov S, Iomin A (2007) Migration and proliferation dichotomy in tumor-cell invasion. Phys Rev Let 98:118101–4
38. Hatzikirou H, Basanta B, Simon M, Schaller C, Deutsch A (2009) "Go or Grow": the key to the emergence of invasion in tumor progression? (under submission)

Cellular Computing

Christof Teuscher
Los Alamos National Laboratory, Los Alamos, USA

Article Outline

Glossary

Cell The biological cell is the smallest self-contained, self-maintaining, and self-reproducing unit of all living organisms. Various computing paradigms were inspired by the biological cell.

Molecular computing A subfield of cellular computing, where the molecules instead of the cell play a central functional role.

Computing The science that deals with the manipulation of symbols. Also refers to the processes carried out by real or abstract computers.

Computation Synonymous with information processing or also algorithm. Computations can for example be performed by abstract machines, real computer hardware, or biological systems. The abstract concept of the Turing machine separates the class of computable from the class of non-computable functions.

Parallel computing Parallel computing involves the execution of a task on multiple processors with the goal to speed-up the execution process by dividing up the task into smaller sub-tasks that can be executed simultaneously.

Definition of the Subject

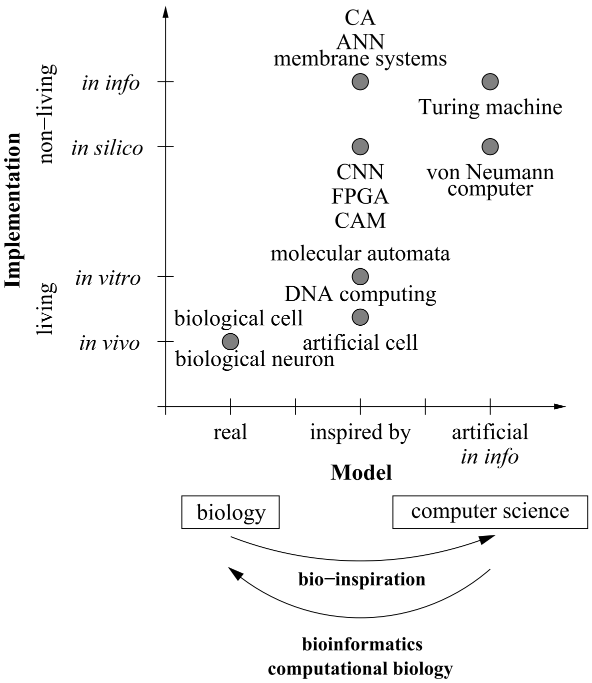
The field of *cellular computing* (abbreviated, CC) defines both a general computing framework and a discipline concerned with the analysis, modeling, and engineering of real cellular processes for the purpose of computation. The biological cell, discovered and coined by R. Hooke in 1665, is the smallest self-contained, self-maintaining, and self-reproducing unit of all living organisms. Its understanding and modeling is crucial both for the understanding of life and for the ability to use, control, and modify its complex bio-chemical processes to perform specific functions for the purpose of *in vivo* or *in vitro* computation. The cellular metaphor has inspired and influenced numerous, both abstract computing models and *in silico* implementations, such as *cellular automata* (abbreviated, CA), *membrane systems* (or P systems), or *Field Programmable Gate Arrays* (abbreviated, FPGAs), with the main purpose to solve algorithmic problems in alternative ways. The cellular computing approach is appealing because the cell provides a convenient level of abstraction and functionality, is reasonably simple with enough abstractions, and can be used as a building block to compose more complex systems. On the other hand, a wide range of computational approaches are used to model and understand the functioning of the inter- and intra-cellular processes of biological cells on its various levels of complexity. The cell's bio-chemical processes can either directly be interpreted as computations or they can be modified for the specific purposes of computations through bio-engineering methods. The hope in using biological cells as computing devices is to (1) ultimately go beyond complexity and fabrication limits currently encountered with traditional silicon circuits, (2) to embed computing capabilities into living organisms for autonomous therapeutic and diagnostic applications, and (3) to realize integrated biological sensing applications that use living organisms, such as for example plants and bacteria, as intelligent sensors.

Introduction

The English naturalist Robert Hooke is commonly accredited for the discovery of the cell [7] and the coining of its term, which he borrowed from the Latin word *cella*, that designates a small chamber. Hooke's work was published in 1665 in his book *Micrographia* [45], which contained observations he made with a very rudimentary microscope of dead cork cells. Since then, tremendous discoveries have been made around the cell, ranging from the pioneering work of German anatomist Walther Flemming [38,75] on cell division (mitosis) to the discovery of the DNA double helix structure by Watson and Crick in 1953 [117]. Nowadays, a sheer endless amount of knowledge and data on cells is available, originating from genetics, biochemistry, and molecular dynamics research. Nevertheless, the computational modeling at the molecular level of complete cells has been notoriously difficult because of the enormous complexity of the real cell, a lack of qualitative understanding, and the large number of unknowns in the models.

With the advent of modern computer science in the 1940s under pioneering fathers such as John von Neumann and Alan M. Turing, biology and computer science started to grow closer together. Biologists increasingly used computers to analyze experimental data and to model real biological organisms, while computer scientists began to draw inspiration from biology with the essential goal to engineer better and more "lifelike" machines. Using computer science, mathematics, statistics, machine learning, and artificial intelligence as tools to solve real problems in biology is commonly known under the term *bioinformatics* [11], while *computational biology* is more concerned with the exploration and discovery of new knowledge, for example by testing specific hypotheses by means of computer models. *Systems biology* [48], on the other hand, focuses on the system level and how the participating biological components interact. This is typically done in a combined approach, which involves theory, computational modeling, and experiments.

Living organisms are complex systems exhibiting a range of desirable characteristics, such as evolution, adaptation, and fault tolerance, that have proved difficult to realize using traditional engineering methodologies. Drawing inspiration from biology to design better and both more "lifelike" machines and algorithms is generally known under the term *biologically-inspired computer science* [39,59,90] (also *nature-inspired*, *bio-mimetic*, or *biologically-motivated*). Typically, this approach involves simulations, algorithms, and classical silicon-based circuits, but not living matter. The goal is not to copy nature, but to solely draw inspiration from it, while the field of *artificial life* [52] is more concerned with understanding life by building it. Figure 1 illustrates the various possibilities to implement cellular models that range from realistic to abstract.



Cellular Computing, Figure 1
Illustration of the various possibilities to implement cellular models that range from realistic to abstract. Drawing inspiration from biology to build better computing machines is known under the term *bio-inspiration*, while using machines to address problems in biology is known as *bioinformatics* and *computational biology*

The broad and very interdisciplinary field of *cellular computing*, in its natural and artificial dimension, defines both a

- General computing framework [89] and a
- Discipline concerned with the analysis, modeling, and engineering of real cellular processes for the purpose of computation [8].

As opposed to pure bio-inspired computing, cellular computing thus also includes the use of real biological cells for information processing. Cellular computing encompasses abstract theory, simulations, models, experiments, and includes both the paradigm of "the cell as a machine" and "the machine as a cell." Sometimes, cellular computing is used interchangeably with *molecular computing* [86], however, while molecular computing is concerned with the information processing in which molecules play a central functional role, cellular computing looks at the cell as the functional building block for more complex systems. Naturally, molecules are involved in cellular information processing as well, so we therefore consider molecular computing as a subset of cellular computing.

The birth hour of *connectionism* in the 1940s can be considered as the first occurrence of looking at cells as automata and modeling them by digital systems. The well known McCulloch–Pitts neuron [60] simulated a wealth of research in neural information processing and modeling that continues with increasing activity until today [9]. In 1948 (but first published in 1969, reprinted in [111]), Alan Turing proposed his own connectionist ideas [103,110], which were highly influenced by the digital systems paradigms. His neurons were simple NAND (not AND) logical gates, which would thus allow in principle to compute any logical function, provided enough neurons are available and that they can be interconnected in arbitrary ways.

The concept of molecular automata and of cells seen as information processing devices appeared around the 1960s. In their studies of the control of gene expression and the synthesis of proteins, Jacob and Monod [46,62], Davis [26], and others emphasized deterministic operations at the level of DNA, RNA, and ribosomes. Early work on protein synthesis, for example by Warner et al. [116], suggested that ribosomes move along the messenger RNA, not unlike the head reading the tape of a Turing machine. In 1961, Pattee [71] used the concept of the Turing machine computation to explain the generation of growing macromolecular sequences and rhetorically asked in the title of a section of his paper "Can molecules compute?" Stahl and Goheen [95] further pioneered the idea of seeing the cell as an information processor and Turing machine. They considered enzymes as computational primitives whose operations are simulated by a Turing machine, and are therefore computable in the formal sense of the term [27]. At about the same time, Sugita considered some aspects of molecular systems by using a logical circuit equivalent [98,99,100]. He states: "The mechanism of life shall be clarified by using electronic analogues, because the chemical system composing the living organism is nothing but a complicated network of rate processes from the chemical or physical point of view. Simulation of such a network may be done either by using an analogue computer or a digital differential analyzer, as has been done at the laboratory[…]. A digital system composed of some chemical reactions may play the role of a controlling computer of the living chemical plant, and the reactions having analogue behavior may be controlled by that computer as well as by analogue control systems which can be seen in primitive control" [98]. Schmitt [83] discussed the biological memory of macromolecular automata in his 1962 book, while Feynman [36] elaborated in his famous talk "There's plenty of room at the bottom" on the possibility to compute with atoms, molecules, and cells. Feynman was more focused on the miniaturization of the computer by storing and processing information at atomic levels, but his ideas were pursued by many others later (e. g. [14,21,86]), and ultimately led to what we know today as *nano-* and *molecular electronics* [24], but also greatly stimulated the earlier cellular computing research.

Michael Conrad, a pioneer in biologically-inspired, molecular, and cellular computing, as well as many others observed that living organisms process information in a very different way than digital computers [20]. Organisms typically exploit the inherent physical properties of the matter they are made of, while digital computers are based on multiple levels of abstractions, which are typically far away from the physical substrate and do not directly exploit any of its specific physical characteristics for the purpose of more efficient computation. Cellular computing, both in its natural and artificial dimension, tries to address these issues by drawing inspiration from real cells and by using them for the purpose of computation.

The principal reason for the increasing interest in cellular computing paradigms seem twofold. Firstly, Moore's Law [63] has dominated the progress of electronic circuits in the past decades, but fundamental limits of silicon-based technology now begin to become serious showstoppers [47]. The hope is that cellular computing will help to go beyond such limits by using alternative computing paradigms inspired by the cell or by using real cells, molecules, and atoms as information processing devices. Secondly, despite lots of progress, Turing's hope that "machines will eventually compete with men in all purely intellectual fields" [108] is far from accomplished. Biological systems outperform their silicon counterparts in various areas, such as for example with their ability to adapt, to recognize objects, or to self-repair. Again, by using cellular computing paradigms, the hope is that such bio-inspired approaches will ultimately allow to go beyond existing limits.

Computation and Computability

Both from the perspective of computer science and biological organisms, a central question is what kind of operations a given system can perform, what kind of problems it can solve, and how efficiently it can do this. Despite their ubiquitousness and seemingly endless power, there are clear and fundamental limitations – that are too often ignored unfortunately – to what computers can do [43,55]. Two fields of computer science are relevant in this context: *computability theory* [27,61] deals with what problems can be solved, while *computational complexity theory* [68] deals with how efficient they can be solved.

Since Alan M. Turing's seminal paper on the abstract concept of the *Turing machine* (abbreviated, TM) [107], *computability* is well defined and we generally know what classes of problems can be solved by which classes of machines. The *Chomsky hierarchy* [94], for example, partitions the formal languages into classes of different expressive power. Each class can be associated a class of abstract automaton, which is able to generate and recognize the corresponding language. At the top of this hierarchy sits the Turing machine, which is able to generate the most expressive class of languages. Along with other abstract formalisms and machines, a Turing machine is able to carry out any effective computation, i. e., it can simulate any other machine capable of performing a well-defined computational procedure. In this context, "effective" is synonymous with "mechanical" or "algorithmic." In its original form, the *Church–Turing thesis* states that there is an equivalence between Church's λ-calculus and Turing machines. Since then, numerous more general and stronger forms of the thesis have emerged. For example, David Deutsch wrote: "Every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means" (p. 99 in [30]). Despite claims from "hypercomputationalists," (e. g., see [28,104] for an overview of this debate) the fact is that no physically realizable device, which thus necessarily operates on finite resources and time, was ever able to compute functions (in the formal sense of the definition) that a Turing machine could not compute, which means that the Church–Turing thesis has not been disproved as of today.

It is obviously an interesting question to ask how and what biological systems "compute," what their limits are, and how efficient they can solve problems. Biological organisms naturally process some information in some way, but how – if at all – can these processes be described in an algorithmic manner? What kind of abstract automata are useful to describe the many computational mechanisms of a biological cell? Paton describes four general types of computations in [70] under which the information processing capabilities of a cell can be seen:

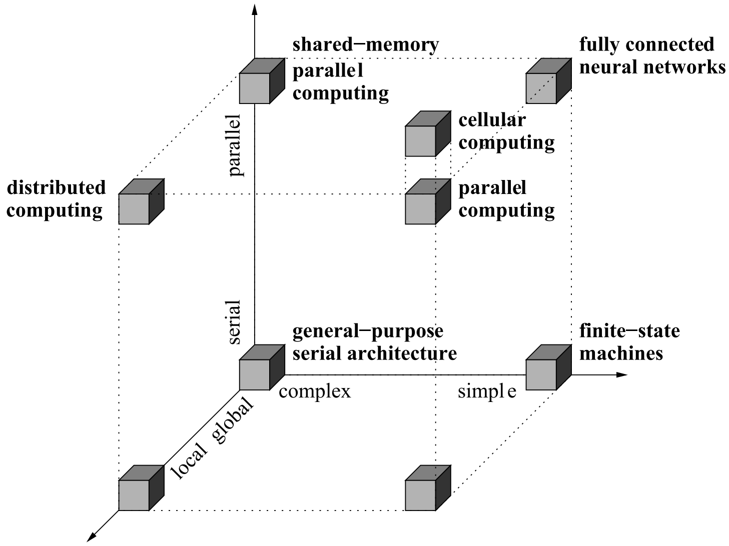
- *Sequential*. The most common form of machine, which usually has a central processor and a global memory. The instructions are carried out serially, such as for example by a Turing machine.
- *Parallel*. Several processors jointly execute a task. A wide range of parallel computing paradigms exist.
- *Distributed*. Similar to parallel computation, but the processors are usually located at larger distances.
- *Emergent*. The global behavior of such a computation emerges from the local interactions between the processors. This form is typical of self-organizing systems, cellular automata, and neural networks.

As he states, "[c]omputational models of the cell can utilize any of these views," however, most processes are best described as a parallel distributed computing paradigm with emergent properties. Paton also points out the risk of trying to force biological systems to "fit" into a certain computational model. Important details might be ignored and one is forced to think within the existing computing paradigms. We can conclude that abstract and formal models of computation have been very useful in describing the information processing capabilities of biological systems, however, not every formalism is well suited for such descriptions, even if his expressiveness matches that of the corresponding biological system. For example, modeling the highly parallel processes of a biological cell by a Turing machine is possible since parallel machines are not more expressive than sequential machines, however, the sequential paradigm does not offer a convenient and useful model in this context.

The Machine as a Cell

In 1999, Moshe Sipper, used the term "cellular computing" in a broad sense to introduce a general computing philosophy and framework. He hypothesized that "[t]his philosophy promises to provide new means for doing computation more efficiently – in terms of speed, cost, power dissipation, information storage, and solution quality. Simultaneously, cellular computing offers the potential to addressing much larger problem instances than previously possible, at least for some application domains" (p. 18 in [89]). Cellular computing consists of three essential and admittedly very general principles:

- Simplicity,



Cellular Computing, Figure 2

The "computing cube." Illustration of the cellular computing paradigm: simple + vastly parallel + local. Redrawn from Sipper [89] and extended. Cellular computing has been placed further along the parallelism axis to emphasize the "vastness" aspect

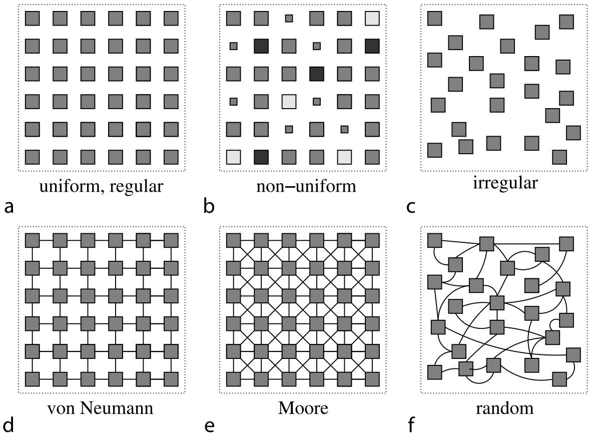
- Vast parallelism, and
- Locality.

Figure 2 illustrates these three principles. While most general-purpose processors are universal machines in the sense of a Turing machine, the concept of cellular computation expects the complexity of the basic unit, the cell, to be significantly lower, which is characterized by *simplicity*. While "simplicity" is not well or formally defined in this context, Sipper provides the example of an AND gate as being simple, while a Pentium processor is not. Most of today's computers are either based on a multi-core processor, which offers several cores that allow to process information in parallel. This trend – generally considered as an elegant way to keep up with Moore's law – is expected to continue in the next few years. As of November 2006, the Top500 supercomputing list [3] is headed by a BlueGene/L server with 131 072 processors. Cellular computing wants to be orders of magnitude above this level of complexity in terms of the number of cells (i. e., the processors) involved. Sipper uses the term *vast parallelism* to characterize systems where the number of cells is measured by the exponential notation 10^x. For example, an *Avogadro-scale system* would involve an Avogadro number of cells, i. e., 10^{23}. As we scale down electronic circuits to molecular and atomic dimensions and make use of self-assembling techniques, assemblies of this complexity become increasingly feasible. It is unlikely that systems of such complexity will ever be built on macroscopic levels. The last principle concerns the interactions among the cells: cellular computing is based on *local interactions* only, where no cell has a global view or control over the entire system. Obviously, the three principles are related and the cell's simplicity, for example, helps to achieve vast parallelism. Similarly, the vast parallelism makes it hard to implement non-local interactions. These interactions are illustrated in Fig. 2 by the "computing cube." Changing a single term in the "equation" simple + vastly parallel + local results in a different computing paradigm.

Compared to traditional parallel computing, which typically makes use of a rather small number of complex nodes, the cellular computing framework is based on a vast number of simple nodes. Also, parallel computers often have non-local interconnect topologies or even special nodes, which have some sort of global control over all other nodes, which cellular computing does not allow. Clearly, while it is already very challenging to find enough parallelism in applications to exploit the parallelism offered by traditional parallel computers, this problem will only be aggravated – and further moved to the programmer and the software – for fine-grained cellular computers. It is commonly accepted that such machines will likely only be more efficient for very specific problem domains which already offer plenty of inherent parallelism, such as for example image processing. Obviously, one can always implement some form of serial computing paradigm on a parallel machine (e. g., implementing a Turing machine on the *Game of Life* [81]), but that completely defeats the purpose of such a machine.

This very general framework naturally allows for considerable flexibility in the models. Besides the three basic principles, a number of other important properties play a role, such as for example the detailed local interconnect topology, the cell's arrangement in space, the mobility, the uniformity, and the cell's behavior [89].

- *Arrangement.* While a majority of abstract cellular models do not consider explicit physical locations and spatial dimensions (e. g., the cell dimension of a cellular automata is not generally considered), more biologically and physically plausible models do. Regular grids in 2D (Fig. 3a) or 3D with rectangular or other geometries are most common, but irregular (Fig. 3c) or even non-rigid grids are possible. In addition, one needs to specify the boundary conditions of the grid (e. g., fixed or wrapped around) and whether the grid has a finite or infinite dimension (open boundary).
- *Interconnect topology.* Cellular computing is based on local interactions only, but there are many possibilities to interconnect neighboring cells in a non-global way. In case of a regular rectangular grid, purely local four- (von Neumann, see Fig. 3d) and eight-cell neighborhoods (Moore, see Fig. 3e) are most common [105]. However, interconnect topologies that are completely random or that have the small-world [118] or power-law [6] property are also possible since no cell has a global view over the whole system.
- *Mobility.* Cells can either be physically mobile (e. g., for ant colony optimization [32]) or immobile. Alternatively, their "program" can also move from one cell to another (e. g., for self-replicating loops [58]) without a physical relocation of the cell.
- *Uniformity.* Both the cell type (i. e, its state and behavior) and it's interconnections with the neighboring cells can be uniform (Fig. 3a) or non-uniform (Fig. 3b) to some degree. Most cellular paradigms use uniform cells and interconnection schemes.
- *State and Behavior.* Cells can have internal states that are either discrete or analog. Based on its internal state and the internal states read from the neighboring cells, the cell changes it own state according to some algorithm. This can be as simple as a look-up-table (abbreviated, LUT) or a mathematical function, or as complicated as a small program. The cells are usually updated in discrete time steps, either in a synchronous or asynchronous way. A continuous behavior is also possible, for example if the cell's behavior is guided by differential equations or analog electronics. Finally, both the behavior and the cell's updating can be deterministic or non-deterministic.



Cellular Computing, Figure 3
Illustration of different cell arrangements and interconnect topologies. a regular and uniform; b regular and non-uniform, different shadings and sizes indicate different cell programs; c irregular and uniform; d regular, uniform, von Neumann neighborhood; e regular, uniform, Moore neighborhood; f irregular, uniform, random neighborhood

Cellular computing does not claim to be a faithful model of biological cells, however, the framework is general and flexible enough to allow for biologically-plausible models as well. As Toffoli and Margulus point out in their book [105] with regards to cellular automata, the generality and flexibility of cellular approaches also comes with a cost. Instead of only a few variables as typical in traditional computing models, the cellular computing paradigm uses at least one variable (or "program") per cell. With the vast number of cells involved, the challenging task is thus to find the right rule(s) which allow to solve a given task. As already stated above, the cellular computing paradigm is ideally suited for tasks that already contain a high degree of parallelism and require local interactions only.

A large number of computing models have drawn inspiration from real cells. While a very large part of the work is concerned with modeling neural cells, a number of researchers were also interested in developmental and other aspects. For example, Astor and Adami [10] used a developmental model for the growth of neural networks. Each neuron is an autonomous entity, which behavior is only determined by its genetic information contained within the cell. An artificial chemistry is also used to model chemicals and substrates. Fleischer and Barr [37] introduced a model of multicellular development that combines elements of the chemical, cell lineage, and mechanical models of morphogenesis pioneered by Turing [109], Lindenmayer [54], and Odell [65]. Cell migration is an important aspect of development, especially of the neural development. Compared to previous models, Fleischer's approach allows cells to move freely within the environment. Other pioneers of morphogenetic, developmental, and growing neural networks models, which necessarily involve cells in various forms, were for example Dellaert and Beer [29], Rust [82], Eggenberger [34], Gruau [42]. More abstract models that deal with the self-maintenance, the self-reproduction, and the self-replication came for example from Ono and Ikegami [66], Varela et al. [113], Langton [51], Mange et al. [58], and various others.

In the following, we will give three more detailed examples of cellular computing models.

Example: Cellular Automata

Cellular automata (abbreviated, CA) were originally conceived by Ulam and von Neumann in the 1940s to provide a formal framework for investigating the behavior of complex, extended systems [115]. CAs are dynamical systems in which space and time are discrete. In its basic version, a cellular automaton [105,120] consists of a regular grid of cells, each of which can be in one of a finite number of *k* possible states, updated synchronously (or asynchronously) in discrete time steps according to a local, usually identical (uniform, otherwise non-uniform) interaction rule. The state of a given cell is determined by the previous states of a specified neighborhood of cells. Figure 3d and e show two typical local CA neighborhoods, the von Neumann and the Moore neighborhood.

In CAs and many other cellular computing systems, there is a confound between the data and the computing devices, i. e., the structure, which process it. The abstract Turing machine and the von Neumann computer architecture make a clear distinction between these two. The Turing machine stores data on the tape, however, a CA cell can both store and process data and the machine can even extend or modify itself. Although it has been shown that elementary CAs are universal [22], this is more of theoretical than practical interest since most CAs that solve some real problem cannot – and do not need to – perform any universal computation.

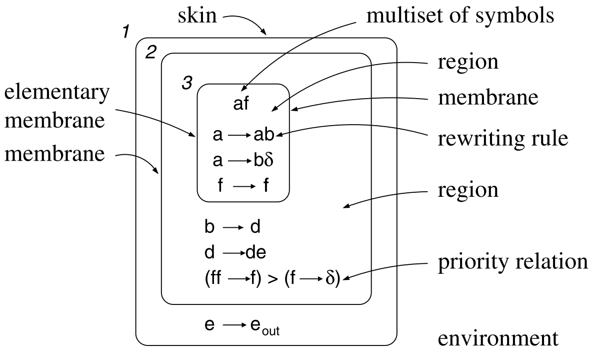
One of the main challenges with cellular automata is to find the cell's rule(s) which allow to obtain a global behavior from the local interactions only. For example, given a specific task, say, detecting the contours of an object in an image, what are the individual cell rules (either uniform or non-uniform) that allow to successfully solve that task for a wide range of inputs? The large – or even vast – number of cells involved and the local interactions make this problem very challenging. Thus, most cellular automata are either programmed by hand – which becomes infeasible for larger systems and complex problems – or the rules are for example found by means of an evolutionary algorithm [18,87,88,92].

Example: Membrane Computing

Artificial chemistries [31] are man-made systems that are a very general formulation of abstract systems of *objects* that follow arbitrary *rules of interaction*. More formally speaking, an artificial chemistry essentially consists of a set of *molecules S*, a set of *rules R*, and a definition of the *reactor algorithm A* that describes how the set of rules is applied to the set of molecules. This very broad paradigm, inspired by bio-chemical systems, allows to describe many complex natural and artificial systems by means of simple rules of decentralized, local, and parallel interactions. The chemical paradigm thus also fits in the bigger framework of cellular computing.

In 1998, George Paun initiated *membrane computing* (also *P systems*) [72,73] as an abstract computational model afar inspired by biochemistry and by some of the basic features of biological membranes. Membrane computing makes use of a hierarchical membrane structure that is similar to the structure used in the chemical abstract machine as proposed by Berry and Boudol [15]. The evolution rules that transform the multisets are inspired by *Gamma systems* as proposed by Banâtre et al. [12].

A typical membrane system consists of cell-like membranes placed inside a unique "skin" membrane. Multisets of *objects* – usually strings of symbols – and a set of *evolution rules* (or *reaction rules*) are then placed inside the regions delimited by the membranes. As an example, a simple membrane system is depicted in Fig. 4. The evolution between system configurations is done nondeterministically but synchronously by applying the rules in parallel for all objects able to evolve. A sequence of transitions in a membrane system is called a *computation*. A computation *halts* when a halting configuration is reached, i. e., when no rule can be applied in any region. In classical membrane systems, a computation is considered successful if and only if it halts, but other interpretations of inputs and outputs are possible. A great feature of membrane systems is their inherent parallelism, which allows to trade space for time, for example to solve computationally hard problems. Using membrane division, it has been shown that NP-complete problems can be solved in poly-



Cellular Computing, Figure 4
An example of a membrane system as given in [73]. The system generates $n^2, n \geq 1$, where *n* is the number of steps *before* the first application of the rule $a \rightarrow b\delta$. The rules are applied synchronously and in a maximum parallel manner, i. e., all rules that can be applied have to be applied. If multiple rules can be applied, one is picked nondeterministically

nomial and even linear time [73]. Most membrane systems are also Turing universal.

A wide variety of different membrane system flavors exist (see [2] for the latest publications), which have been applied to applications ranging from modeling spiking neural systems to economics.

Example: Amorphous Computing

In a 1996 white paper, Abelson et al. first described the *amorphous computing* (abbreviated, AC) framework [4]. Amorphous computing is the development of organizational principles and programming languages for obtaining coherent global behavior from the local cooperation of a myriad of unreliable parts that are interconnected in unknown, irregular, and time-varying ways. In biology, this question has been recognized as fundamental in the context of animals (such as ants and bees) that cooperate and form organizations. Amorphous computing asks this question in the field of computer science and engineering. Using the biological metaphor, the cells cooperate to form a multicellular organism (also called *programmable multitude*) under the direction of a genetic program shared by all members of the colony. Again, the cellular computing philosophy applies particularly well to the amorphous computing framework. The properties of an amorphous computer can be summarized under the following assumptions, where processor is used synonymous with cell:

- individual processors are identical and mass produced;
- processors possess no *a priori* knowledge of their location, orientation, or neighbors' identities;
- processors operate asynchronously although they have similar clock speeds;
- processors are distributed densely and randomly;
- processors are unreliable;
- processors communicate only locally and do not have a precision interconnect;
- the processors are arranged on a 2D surface or in 3D space;
- it is assumed that the number of particles is very large; and
- the communication model assumes that all processors have a circular broadcast of approximately the same fixed radius (large compared to the fixed size of a processor) and share a single channel.

The biggest challenge in amorphous computing – as with the majority of cellular computing paradigms – is how related to how to program the individual cells in order to obtain a consistent and robust global behavior. The amorphous computing research community came up with several automated approaches, which allow to compile a global system description into local interactions. Example are Nagpal's *origami shape language* [64], Coore's *growing point language* [23], and Butera's *process fragments* [16].

Cellular Computing Hardware

Cells have not only inspired abstract computing models but also real hardware, which is commonly called *biologically-inspired hardware* [93]. On the other hand, both specialized and non-specialized computers have been used for the acceleration and the faithful simulation of many biological processes related to the cell, particularly within the field of *bioinformatics* [11]. A biological cell is a massively parallel system, where thousands of highly complex processes run concurrently. Simulating these processes on a sequential von Neumann computer will obviously not be very efficient and will seriously limit the performance. Specialized hardware, which offers a high level of inherent parallelism, has thus attracted considerable attention in this field of cellular computing.

Cellular automata machines (abbreviated, CAMs) are specialized machines with the goal to very efficiently simulate cellular automata. For example, Hillis' connection machine [44] was the first milestone in that endeavor in the early 1980s. The processors were extremely simple and there was a strong emphasis on the interconnectivity, but not particularly on local topologies because the machines were targeted for supercomputer applications. Toffoli and Margolus' CAM machines [105] represent other examples of highly specialized machines. Nowadays, because of their inherent fine-grained parallelism, reconfigurable circuits, such as *Field Programmable Gate Arrays* (abbreviated, FPGAs) [41,114], are ideal candidates to simulate cellular systems. For example, Petreska and Teuscher [77] proposed a very first hardware realization of membrane systems using reconfigurable circuits, which was based on a universal and minimal membrane hardware component that allowed to very efficiently evolve membrane systems in hardware. Another example of cellular hardware are *cellular neural network* (abbreviated, CNN) [19] chips, which are massively parallel analog processor arrays. The CNN approach is extremely powerful for specific applications, such as real-time image and video processing.

In the following, we will provide three more detailed examples of typical and biologically-inspired hardware, which is based on cellular paradigms and metaphors.

Example: Embryonics

Biological systems grow, live, adapt, and reproduce, characteristics that are not truly encompassed by any existing computing system. Sipper et al. [93] proposed the *POE model* of bio-inspired hardware systems, which stands for *phylogeny* (abbreviated, P), *ontogeny* (abbreviated, O), and *epigenesis* (abbreviated, E). This model allows to partition the hardware space along three axes, which, in very simplified terms, correspond to evolution (P), development (O), and learning (E). Each axis thus represents a different form of adaption and organization on a different time scale. The ultimate goal are machines that combine all three forms of adaption, so called *POEtic machines* [112].

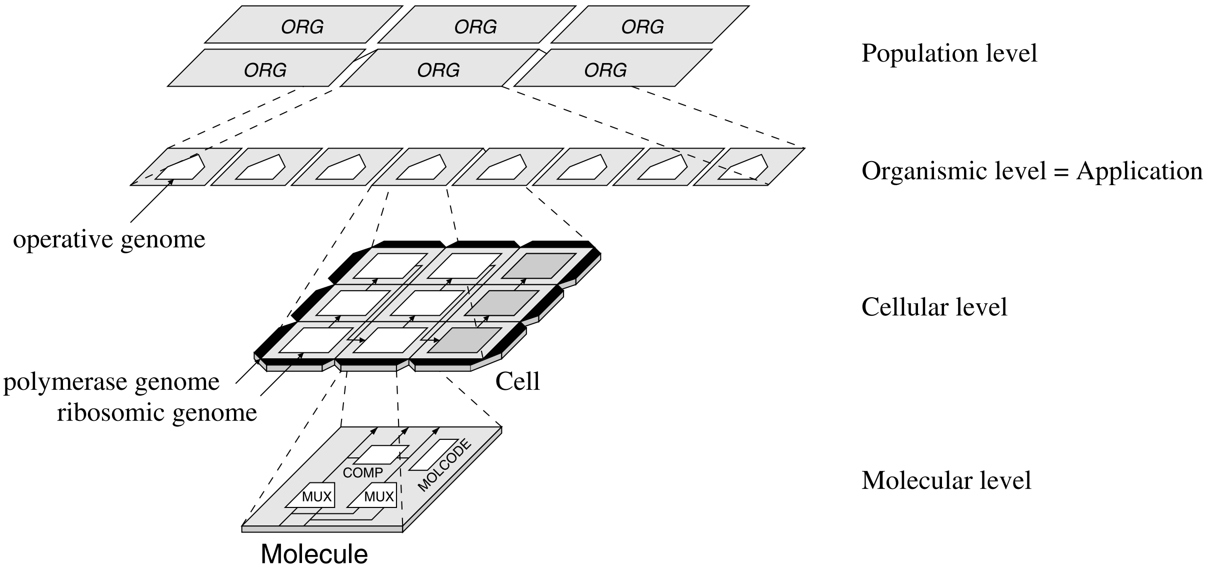
The *embryonics project* [57,59] (embryonics stands for embryonic electronics), is an attempt to design highly-robust integrated circuits, endowed with the properties usually associated with the living world: self-repair (cicatrization) and self-replication. In this context, self-replication allows for a complete reconstruction of an organism by making a copy of itself. The approach draws inspiration from the basic processes of molecular biology and from the embryonic development of living beings, which is represented by the ontogenetic axis in the POE model. An embryonic circuit is based on a finite but arbitrarily large two-dimensional surface of silicon. This surface is divided into rows and columns, whose intersections define the cells. Since such cells (i. e., a small processor and its memory) have an identical physical structure (i. e., an identical set of logic operators and connections), the cellular array is homogeneous. Embryonics largely draws inspiration from the following biological features, which the majority of living beings (at the exception of unicellular organisms) share: (1) multicellular organization, (2) cellular division, and (3) cellular differentiation. Similar to biological cells, each embryonic cell is "universal" in the sense that it contains the whole of the organism's genetic material, the genome. This feature allows to implement self-repair in an elegant way since each cell is potentially able to replace every other cell within an embryonic organism. The final embryonics architecture makes use of four hierarchical levels of organization:

- *Molecule.* The basic primitive of the system, which essentially consists of a multiplexer associated with a programmable connection network. The multiplexer is duplicated to allow the detection of faults. The logic function of each molecule is defined by its molecular code, the *MOLCODE.*
- *Cell.* A finite set of molecules makes up a cell, which is essentially a processor with an associated memory.
- *Organism* A finite set of cells makes up an organism (synonymous with application), an application-specific multiprocessor system.
- *Population.* The organism can itself self-replicate, given rise to a population of identical organisms.

Figure 5 illustrates these four hierarchical levels. As a showcase of bio-inspired systems and hardware, a complete embryonic machine and application has been implemented in real hardware by using *Field Programmable Gate Arrays* (abbreviated, FPGAs) [41] to implement the basic molecule [102]. The main application, which used several thousand molecules, was a fault-tolerant electronic watch, the BioWatch [96]. The project illustrates how biological organisms, and cells in particular, can help to build alternative computing machines with properties not truly encompassed by any existing architecture.

Example: Field Programmable Gate Arrays

A *Field Programmable Gate Array* (abbreviated, FPGA) [41] is a reconfigurable device [114] which is based on a set of *configurable logic blocks* (abbreviated, CLBs) that are interconnected by a programmable interconnect fabric. The logic blocks of most FPGAs are rather simple, consisting typically of some combinational logic (e. g., a 4-input look-up-table) and some sequential logic (e. g., a flip-flop). Some CLBs offer additional memory too. The interconnect fabric is typically organized in a hierarchical manner with mostly local interconnections between CLBs, but with the possibility to establish limited long(er)-distance connections as well. Sophisticated design tools, compilers, and hardware description languages allow to map essentially any digital circuit onto such a reconfigurable chip, provided enough CLBs are available.



Cellular Computing, Figure 5
The four-level hierarchy of the embryonics landscape. Each level is configured by a part of the genome, which in parallel into all cells once their cell membranes have been constructed by a mechanism similar to cellular division. After this process, the cells differentiate according to their spatial location in the organism [57]

In accordance with the cellular computing framework, FPGAs are based on simple basic building blocks (the CLBs), are locally interconnected (in the sense that no cell controls the entire system), and today's high-end circuits contain hundreds of thousands of CLBs. FPGAs tend to be slower than full-custom VLSI circuits, but offer more flexibility at a lower price. The application of FPGA is obviously most interesting where their massive parallelism can be fully exploited, for example for image processing, neural networks [67], DNA sequence matching, or cellular automata. Sipper et al., for example, used FPGAs for their firefly CA machine [91], which were also used to implement on-line evolution algorithms to automatically find solutions to the synchronization task for CAs [25,88].

Example: The Cell Matrix

The *Cell Matrix* [1,33,56] is an extremely fine-grained and universal hardware architecture, not unlike the structure of an FPGA. The reconfigurable device consists of a homogeneous collection of cells that are interconnected in a nearest-neighbor scheme (e. g., von Neumann or other neighborhoods). Each cell performs very basic operations only that are implemented by a simple look-up table. Unlike traditional reconfigurable devices that are controlled by external systems, the Cell Matrix is self-configurable, i. e., each cell within the matrix can reconfigure other cells. In order to accomplish this, each cell can operate independently in one of two modes: (1) the D mode or (2) the C mode. In the D mode (processing mode), incoming data is processed by the cell's internal look-up table and used to generate the output. In the C mode (configuration mode), on the other hand, incoming data is used to re-write the cell's look-up table. A cell can therefore modify one of its neighbors by placing it in the C mode and by writing the data into the look-up table. Since a cell's look-up table governs the ability to control a neighboring cell's mode, a cell X, for example, can configure a neighboring cell Y in such a way that Y will subsequently configure a third cell Z. By combining cells to so-called *supercells*, hierarchical designs of almost any complexity can be realized. The small groups of cells then interact with nearby groups to achieve higher functionalities, which are used to perform more complex functions, and so on. The Cell Matrix architecture was extended in 2004 by Petraglio in his Ph D thesis [76].

The Cell as a Machine

"The brain computes!" This is accepted as a truism by the majority of neuroscientists engaged in discovering the principles employed in the design and operation of nervous systems. What is meant here is that any brain takes the incoming sensory data, encodes them into various biophysical variables, such as the membrane potential or neural firing rates, and subsequently performs a very large number of ill-specified operations, frequently termed computations, on these variables to extract relevant features from the input" (p. 1 in [49]). Further on, Koch continues by elaborating that any physical process, which transforms variables, can in principle be thought of as computation, as long as it can be described by some mathematical formulation. As also noted by Conrad [20], living organisms process information in a very different way than digital computers and typically exploit the inherent physical properties of the matter they are made of. The direct, efficient, and sophisticated ways to process information in cells, molecules, and atoms are the results of billions of years of evolution.

In general, computation in living systems is more a question of what level is being analyzed and through which glasses the analyzer looks. A cell has thousands of information processing mechanisms, some are easy to map to existing computing paradigms, others are far away from any existing model and can therefore only be captured by new paradigms. If we want to use a biological cell as a computing device, we essentially have two possibilities:

- *Interpretation.* Interpret existing bio-chemical cellular mechanisms as computation.
- *Modification.* With the advances in bio-technology and bio-engineering in the last decade, bio-chemical processes at all organization levels of a cell can increasingly be modified for the specific purposes of computation.

For example, DNA can be described in computational terms rather easily and the DNA strand interpreted as data, which is being read by ribosomes is a fairly straightforward natural analog to the Turing machine [116]. But this undertaking becomes more interesting if the analogy can be reversed and the DNA is used to realize Boolean *in vivo* or *in vitro* logic gates in a well controlled manner.

As stated in the introduction, the concept of molecular automata and cells seen as information processing devices appeared around the 1960s and the field has seen impressive developments since then. The field is vast and we shall only give a very shallow taste here. The books by Amos [8], Calude and Paun [17], Paton et al. [69], and Sienko et al. [86] provide excellent overviews from the "cell as a machine" perspective on more recent work, while Paton [70], for example, deals with some of the earlier concepts.

Since Adelman's seminal work on *in vitro* DNA computation [5], this unconventional computing paradigm has played a very important role in cellular computation in both experiment and theory [74]. DNA computation clearly outperforms traditional silicon-based electronics in terms of speed, energy efficiency, and information density. However, while it is rather easy to explore huge search spaces, it is often non-trivial to find the correct answer within all the generated solutions, which has somehow limited the practical applications. Other work has focused on RNA editing as a computational process, which offers another interesting cellular paradigm for "biological software" [50]. Later, Winfree, Weiss, and various others, pioneered DNA-based and other *in vivo* logical circuits [13,84,85,119]. Based on deoxyribozymes, Stefanovic's group has developed molecular automata and logic [53,97]. Their circuits include an full adder, three-input logic gates, and an automaton that plays a version of the game tic-tac-toe. While such circuits serve as examples and proof-of-concepts, the group's long-term goal is to use engineered molecules to control autonomous therapeutic and diagnostic devices. These research areas are fast moving and lots of novel results are to be expected in the next few years.

Future Directions

"The simplest living cell is so complex that supercomputer models may never simulate its behavior perfectly. But even imperfect models could shake the foundations of biology" [40]. Gibbs further emphasizes that most attempts to create artificial life or to faithfully model biological systems suffered from a tremendous number of degrees of freedom. The high number of unconstrained system parameters can then lead to almost any desired behavior. Also, the models are often so complicated that they have a very limited ability to predict anything at all. Tomita believes that the study of the cell will never be complete unless its dynamic behavior on all levels is fully understood [106]. He suggests that the complex behavior can only be understood by means of sophisticated computer models that allow to undertake whole cell simulations. Endy and Brent [35] emphasize that "[p]ast efforts to model behavior of molecular and cellular systems over absolute time typically were qualitatively incomplete or oversimplified compared to available knowledge, and qualitatively incomplete in the sense that key numbers were unknown." It is estimated that complete E. coli simulations could run, perhaps, on a single processor system by 2020.

But the future not only lies in more faithful cellular models and in the ability to tweak cells for the purpose of computation. The construction of realistic living cells *in silico*, thought intractable for a long time, has now become within reach, and so does the possibility to build artificial living cells. Several efforts with this specific goal in mind are currently under way in both Europe and the United

States (see for example [78,79,101]). The question is, how and under which conditions simple life forms, which are able to self-maintain and self-replicate, can be synthesized artificially? Steen Rasmussen's team has taken one of the simpler and more artificial routes [79,80], but there is still a long way to go towards artificial cells that will be able to truly self-replicate and to hopefully perform some specific user-defined functions.

Cellular computing is still a very young field and lots of progress is to be expected in the next few decades, particularly in the areas of bio-technology, nano-technology, modeling, and computing. The main challenges in this interdisciplinary field are mostly related to the remaining wide-open gap between biological systems and machines [20]. On one hand, a much better understanding of cellular systems is required, on the other hand, we need to be able to apply this vast and growing body of knowledge to build better and more "lifelike" computing machines by either drawing inspiration from them or by using real cells and organisms for the purpose of computation.

The following research directions and challenges need to be addressed in the next few years:

1. Faithful qualitative and quantitative cellular models require a better understanding of the biological cell and its bio-chemical mechanism. For example, better non-invasive brain imaging techniques are required for a better understanding of neurons, populations of neurons, and brain regions.
2. Large-scale molecular simulations of cellular processes and large populations of cells. For example, by using specialized supercomputers, large-scale brain simulations only become possible now.
3. Novel computing paradigms are required to model and to understand biological systems. Boolean logic is appropriate for digital systems, where 0s and 1s can easily be represented by the presence and absence of electrical current, but alternative representations may be more appropriate for biological systems.
4. Novel programming paradigms for cellular systems made up from a vast number of cells. Traditional top-down design approaches do not typically scale up to such complexities. Nature-inspired bottom-up design approaches are required.

Bibliography

Primary Literature

1. Cell matrix corporation. http://www.cellmatrix.com. Accessed 12 Jul 2008
2. The P systems web page. http://ppage.psystems.eu. Accessed 12 Jul 2008
3. Top500 supercomputing sites. http://www.top500.org. Accessed 12 Jul 2008
4. Abelson H, Allen D, Coore D, Hanson C, Rauch E, Sussman GJ, Weiss R (2000) Amorphous computing. Commun ACM 43(5):74–82
5. Adelman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266:1021–1024
6. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
7. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (eds) (1994) Molecular Biology of the Cell, 3rd edn. Garland Publishing, New York
8. Amos M (ed) (2004) Cellular Computing. Oxford University Press, New York
9. Arbib MA (ed) (1995) The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge
10. Astor JC, Adami C (2001) A developmental model for the evolution of artificial neural networks. Artif life 6(3):189–218
11. Baldi P, Brunak S (2001) Bioinformatics: The Machine Learning Approach, 2nd edn. MIT Press, Cambridge
12. Banâtre JP, Coutant A, Le Metayer D (1988) A parallel machine for multiset transformation and its programming style. Future Gener Comput Syst 4:133–144
13. Benenson Y, Paz-Elizur T, Adar R, Keinan E, Leivneh Z, Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. Nature 414(6862):430–434
14. Bennett CH (1982) The thermodynamics of computation a review. Int J Theor Phys 21(12):905–940
15. Berry G, Boudol G (1992) The chemical abstract machine. Theor Comput Sci 96:217–248
16. Butera WJ (2002) Programming a Paintable Computer. Ph D thesis, MIT Media Lab, Cambridge
17. Calude CS, Paun G (2000) Computing with Cells and Atoms: An Introduction to Quantum and Membrane Computing. Taylor & Francis, New York
18. Capcarrere MS (2002) Cellular Automata and Other Cellular Systems: Design & Evolution. Ph D thesis, Swiss Federal Institute of Technology Lausanne
19. Chua LO, Roska T (2002) Cellular Neural Networks & Visual Computing. Cambridge University Press, Cambridge
20. Conrad M (1989) The brain-machine disanalogy. BioSystems 22(3):197–213
21. Conrad M, Liberman EA (1982) Molecular computing as a link between biological and physical theory. J Theor Biol 98(2):239–252
22. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15(1):1–40
23. Coore D (1999) Botanical Computing: A Developmental Approach to Generating Interconnect Topologies on an Amorphous Computer. Ph D thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge
24. Cuniberti G, Fagas G, Richter K (eds) (2005) Introducing Molecular Electronics. In: Lecture Notes in Physics, vol 680. Springer, Berlin
25. Das R, Crutchfield JP, Mitchell M, Hanson JE (1995) Evolving globally synchronized cellular automata. In: Eshelman LJ (ed) Proceedings of the Sixth International Conference on Genetic Algorithms. Morgan Kaufmann, San Francisco, pp 336–343
26. Davis BD (1961) The teleonomic significance of biosynthetic control mechanisms. Cold Spring Harbor Symp Quant Biol 26:1–10

27. Davis M (1958) Computability and Unsolvability. McGraw-Hill, New York
28. Davis M (2004) The myth of hypercomputation. In: Teuscher C (ed) Alan Turing: Life and Legacy of a Great Thinker. Springer, Berlin, pp 195–212 (reprinted with color images in 2005)
29. Dellaert F, Beer RD (1994) Toward an evolvable model of development for autonomous agent synthesis. In: Brooks RA, Maes P (eds) Artificial Life IV. Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems. A Bradford Book, MIT Press, Cambridge, pp 246–257
30. Deutsch D (1985) Quantum theory, the Church–Turing principle of the Universal Quantum Computer. Proc Royal Soc London A400:97–117
31. Dittrich P, Ziegler J, Banzhaf W (2001) Artificial chemistries a review. Artif Life 7(3):225–275
32. Dorigo M, Stützle T (2004) Ant Colony Optimization. MIT Press, Cambridge
33. Durbeck LJK, Macias NJ (2001) The Cell Matrix: An architecture for nanocomputing. Nanotechnology 12:217–230
34. Eggenberger P (1997) Creation of neural networks based on developmental and evolutionary principles. In: Gerstner W, Germond A, Hasler M, Nicod JD (eds) Proceedings of the International Conference on Artificial Neural Networks. ICANN'97, Lausanne, Switzerland. Lecture Notes in Computer Science, vol 1327. Springer, Berlin, pp 337–342
35. Endy D, Brent R (2001) Modelling cellular behavior. Nature 409(6818):391–395
36. Feynman RP (1960) There's plenty of room at the bottom: An invitation to enter a new field of physics. Caltech's Eng Sci (Feb 1960):22–36
37. Fleischer KW, Barr AH (1994) A simulation testbed for the study of multicellular development: The multiple mechanisms of morphogenesis. In: Langton CG (ed) Artificial Life III. SFI Studies in the Science of Complexity, vol XVII. Addison–Wesley, Redwood City, pp 389–416
38. Flemming W (1880) Beiträge zur Kenntnis der Zelle und ihrer Lebenserscheinungen. Arch Mikrosk Anat 18:151–289
39. Forbes N (2004) Imitation of Life: How Biology is Inspiring Computing. MIT Press, Cambridge
40. Gibbs WW (2001) Cybernetic cells. Sci Am 285(2):43–47
41. Gokhale M, Graham PS (2005) Reconfigurable Computing: Accelerating Computation with Field Programmable Gate Arrays. Springer, Berlin
42. Gruau F (1994) Neural Network Synthesis Using Cellular Encoding and the Genetic Algorithm. Ph D thesis, Ecole Normale Supérieure de Lyon
43. Harel D (2003) Computers Ltd.: What They Really Can't Do. Oxford University Press, New York
44. Hillis DW (1985) The Connection Machine. MIT Press, Cambridge
45. Hooke R (1665) Micrographia: Or, some physiological descriptions of minute bodies made by magnifying glasses. J. Martyn and J. Allestry, London
46. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356
47. Kish LB (2002) End of Moore's law: Thermal (noise) death of integration in micro and nano electronics. Phys Lett A 305: 144–149
48. Kitano H (ed) (2001) Foundations of Systems Biology. MIT Press, Cambridge
49. Koch C (1999) Biophysics of Computation. Oxford University Press, New York
50. Landweber LF, Kari L (1999) The evolution of cellular computing: Nature's solution of a computational problem. BioSystems 52:3–13
51. Langton CG (1984) Self-reproduction in cellular automata. Physica D 10(1–2):135–144
52. Langton CG (ed) (1995) Artificial Life: An Overview. MIT Press, Cambridge
53. Lederman H, Macdonald J, Stefanovic D, Stojanovic MN (2006) Deoxyribozyme-based three-input logic gates and construction of a molecular full adder. Biochemistry 45(4): 1194–1199
54. Lindenmayer A (1968) Mathematical models for cellular interaction in development, parts I and II. J Theor Biol 18:280–315
55. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054
56. Macias NJ (1999) The PIG paradigm: the design and use of a massively parallel fine grained self-reconfigurable infinitely scalable architecture. In: Stoica A, Keymeulen D, Lohn J (eds) Proceedings of the First NASA/DOD Workshop on Evolvable Hardware. IEEE Computer Society, Los Alamitos, pp 175–180
57. Mange D, Sipper M, Stauffer A, Tempesti G 2000 Toward robust integrated circuits: The embryonics approach. Proc IEEE 88(4):516–540
58. Mange D, Stauffer A, Peparolo L, Tempesti G (2004) A macroscopic view of self-replication. Proc IEEE 92(12):1929–1945
59. Mange D, Tomassini M (eds) (1998) Bio-Inspired Computing Machines: Towards Novel Computational Architectures. Presses Polytechniques et Universitaires Romandes, Lausanne
60. McCulloch WS, Pitts WH (1943) A logical calculus of the ideas immanent in neural nets. Bull Math Biophys 5:115–133
61. Minsky ML 1967 Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs
62. Monod J, Jacob F (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harbor Symp Quantit Biol 26:389–401
63. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8):114–117
64. Nagpal R (2001) Programmable Self-Assembly: Constructing Global Shape using Biologically-inspired Local Interactions and Origami Mathematics. Ph D thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge
65. Odell GM, Oster G, Albrech P, Burnside B (1981) The mechanical basis of morphogenesis. In: Epithelial folding and invagination. Dev Biol 85:446–462
66. Ono N, Ikegami T (2000) Self-maintenance and self-reproduction in an abstract cell model. J Theor Biol 206(2):243–253
67. Ormondi AR, Rajapakse JC (eds) (2006) FPGA Implementations of Neural Networks. Springer, Dordrecht
68. Papadimitrou CH (1994) Computational Complexity. Addison–Wesley, Reading
69. Paton R, Bolouri H, Holcombe M, Parish JH, Tateson R (eds) (2004) Computation in Cells and Tissues. Springer, Berlin
70. Paton RC (1993) Some computational models at the cellular level. Biosystems 29:63–75
71. Pattee HH (1961) On the origin of macromolecular sequences. Biophys J 1(8):683–710
72. Paun G (2000) Computing with membranes. J Comput Syst

Sci 61(1):108–143 (first published in a TUCS Research Report, No 208, Nov 1998, http://www.tucs.fi)

73. Paun G (2002) Membrane Computing. Springer, Berlin
74. Paun G, Rozenberg G, Salomaa A (1998) DNA Computing: New Computing Paradigms. Springer, Heidelberg
75. Paweletz N (2001) Walter flemming: Pioneer of mitosis research. Nat Rev Mol Cell Biol 2:72–75
76. Petraglio E (2004) Fault tolerant self-replicating systems. Ph D thesis, Swiss Federal Institute of Technology (EPFL), Lausanne
77. Petreska B, Teuscher C (2004) A reconfigurable hardware membrane system. In: Martin-Vide C, Mauri G, Paun G, Rozenberg G, Salomaa A (eds) Membrane Computing. Lecture Notes in Computer Science, vol 2933. Springer, Berlin, pp 269–285
78. Pohorille A, Deamer D (2002) Artificial cells: Prospects for biotechnology. Trends Biotechnol 20(3):123–128
79. Rasmussen S, Chen L, Deamer D, Krakauer DC, Packard NH, Stadler PF, Bedau MA (2004) Transitions from nonliving to living matter. Science 303:963–965
80. Rasmussen S, Chen L, Nilsson M, Abe S (2003) Bridging nonliving and living matter. Artif Life 9(3):269–316
81. Rendell P (2002) Turing universality of the Game of Life. In: Adamatzky A (ed) Collision-Based Computing. Springer, London, pp 513–539
82. Rust AG (1998) Developmental Self-Organisation in Artificial Neural Networks. Ph D thesis, University of Hertfordshire
83. Schmitt FO (1962) Macromolecular Specificity and Biological Memory. MIT Press, Cambridge
84. Seelig G, Soloveichik D, Zhang DY, Winfree E (2006) Enzyme-free nucleic acid logic circuits. Science 314:1585–1588
85. Shapiro E, Gil B (2007) Logic goes in vitro. Nat Nanotechnol 2:84–85
86. Sienko T, Adamatzky A, Rambidi NG, Conrad M (eds) (2003) Molecular Computing. MIT Press, Cambridge
87. Sipper M (1996) Co-evolving non-uniform cellular automata to perform computations. Physica D 92:193–208
88. Sipper M (1997) Evolution of Parallel Cellular Machines: The Cellular Programming Approach. Springer, Heidelberg
89. Sipper M (1999) The emergence of cellular computing. IEEE Comput 32(7):18–26
90. Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York
91. Sipper M, Goeke M, Mange D, Stauffer A, Sanchez E, Tomassini M (1997) The firefly machine: Online evolware. In: Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC'97), Piscataway. IEEE Press, Piscataway, pp 181–186
92. Sipper M, Ruppin E (1997) Co-evolving architectures for cellular machines. Physica D 99:428–441
93. Sipper M, Sanchez E, Mange D, Tomassini M, Pérez-Uribe A, Stauffer A (1997) A phylogenetic, ontogenetic, and epigenetic view of bio-inspired hardware systems. IEEE Trans Evol Comput 1(1):83–97
94. Sipser M (2006) Introduction to the Theory of Computation, 2nd edn. Thomson, Boston
95. Stahl WR, Goheen HE (1963) Molecular algorithms. J Theor Biol 5(2):266–287
96. Stauffer A, Mange D, Tempesti G, Teuscher C (2001) Bio Watch: A giant electronic bio-inspired watch. In: Keymeulen D, Stoica A, Lohn J, Zebulum RS (eds) Proceedings of the Third NASA/DoD Workshop on Evolvable Hardware, EH-2001. IEEE Computer Society, Los Alamitos, pp 185–192
97. Stojanovic MN, Stefanovic D (2003) A deoxyribozyme-based molecular automaton. Nat Biotechnol 21:1069–1074
98. Sugita M (1961) Functional analysis of chemical systems in vivo using a logical circuit equivalent. J Theor Biol 1(2):415–430
99. Sugita M (1963) Functional analysis of chemical systems in vivo using a logical circuit equivalent. In: The idea of a molecular automaton. J Theor Biol 4(2):179–192
100. Sugita M, Fukuda N (1963) Functional analysis of chemical systems in vivo using a logical circuit equivalent. III: Analysis using a digital circuit combined with an analogue computer. J Theor Biol 5(3):412–425
101. Szostak JW, Bartel DP, Luisi PL (2001) Synthesizing life. Nature 409:387–390
102. Tempesti G, Mange D, Stauffer A, Teuscher C (2002) The Biowall: An electronic tissue for prototyping bio-inspired systems. In: Stoica A, Lohn J, Katz R, Keymeulen D, Zebulum RS (eds) Proceedings of the 2002 NASA/DoD Conference on Evolvable Hardware. IEEE Computer Society, Los Alamitos, pp 221–230
103. Teuscher C (2002) Turing's Connectionism. An Investigation of Neural Network Architectures. Springer, London
104. Teuscher C, Sipper M (2002) Hypercomputation: Hype or computation? Commun ACM 45(8):23–24
105. Toffoli T, Margolus N (1987) Cellular Automata Machines. MIT Press, Cambridge
106. Tomita M (2001) Whole-cell simulation: A grand challenge of the 21st century. Trends Biotechnol 19(6):205–210
107. Turing AM (1937) On computable numbers, with an application to the Entscheidungsproblem. Proc London Math Soc 42:230–265; Corrections. Proc London Math Soc 43:544–546
108. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460
109. Turing AM (1952) The chemical basis of morphogenesis. Philos Trans Royal Soc London B 237:37–72
110. Turing AM (1969) Intelligent machinery. In: Meltzer B, Michie D (eds) Machine Intelligence, vol 5. Edinburgh University Press, Edinburgh, pp 3–23
111. Turing AM (1992) Intelligent machinery. In: Ince DC (ed) Collected Works of A.M. Turing: Mechanical Intelligence. North-Holland, Amsterdam, pp 107–127
112. Tyrrell A, Sanchez E, Floreano D, Tempesti G, Mange D, Moreno JM, Rosenberg J, Alessandro Villa EP (2003) Poetic tissue: An integrated architecture for bio-inspired hardware. In: Tyrrell AM, Haddow PC, Torresen J (eds) Evolvable Systems: From Biology to Hardware. Proceedings of the 5th International Conference (ICES2003). Lecture Notes in Computer Science, vol 2606. Springer, Berlin, pp 129–140
113. Varela F, Maturana H, Uribe R (1974) Autopoiesis: The organization of living systems, its characterization and a model. BioSystems 5:187–196
114. Villasenor J, Mangione-Smith WH (1997) Configurable computing. Sci Am 276(6):54–59
115. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Urbana
116. Warner JR, Rich A, Hall CE (1962) Electron microscope studies of ribosomal clusters synthesizing hemoglobin. Science 138(3548):1299–1403

117. Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738

118. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442

119. Weiss R (2001) Cellular Computation and Communications using Engineered Genetic Regulatory Networks. Ph D thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge

120. Wolfram S (1984) Cellular automata as models of complexity. Nature 311:419–424

Books and Reviews

Adamatzky A (ed) (2002) Collision-Based Computing. Springer, London

Adami C (1998) Introduction to Artificial Life. Springer, New York

Amos M (2005) Theoretical and Experimental DNA Computation. Springer, Berlin

Bentley BJ (2001) Digital Biology: Now Nature is Transforming our Technology. Headline Book Publishing, London

Ehrenfeucht A, Harju T, Petre I, Prescott DM, Rozenberg G (2004) Computation in Living Cells: Gene Assembly in Ciliates. Springer, Berlin

Hey AJG (ed) (1998) Feynman and Computation: Exploring the Limits of Computers. Westview, Boulder

Landweber LF, Winfree E (eds) (2002) Evolution as Computation. Springer, Berlin

Paun G, Rozenberg G (2002) A guide to membrane computing. J Theor Comput Sci 287(1):73–100

Petty MC, Bryce MR, Bloor D (eds) (1995) An Introduction to Molecular Electronics. Oxford University Press, New York

Pozrikidis C (ed) (2003) Modeling and Simulation of Capsules and Biological Cells. Chapman and Hall/CRC, Boca Raton

Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York

Trimberger SM (1994) Field-Programmable Gate Array Technology. Kluwer, Boston

Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign

Zauner KP (2005) Molecular information technology. Crit Rev Solid State Mater Sci 30(1):33–69

Center Manifolds

GEORGE OSIPENKO

State Polytechnic University, St. Petersburg, Russia

Article Outline

Glossary

Bifurcation Bifurcation is a qualitative change of the phase portrait. The term “bifurcation” was introduced by H. Poincaré.

Continuous and discrete dynamical systems A dynamical system is a mapping $X(t,x)$, $t \in R$ or $t \in Z$, $x \in E$ which satisfies the group property $X(t+s,x) = X(t, X(s,x))$. The dynamical system is continuous or discrete when t takes real or integer values, respectively. The continuous system is generated by an autonomous system of ordinary differential equations

$$\dot{x} \equiv \frac{\mathrm{d}x}{\mathrm{d}t} = F(x) \tag{1}$$

as the solution $X(t,x)$ with the initial condition $X(0,x) = x$. The discrete system generated by a system of difference equations

$$x_{m+1} = G(x_m) \tag{2}$$

as $X(n,x) = G^n(x)$. The phase space E is the Euclidean or Banach.

Critical part of the spectrum Critical part of the spectrum for a differential equation $\dot{x} = Ax$ is σ_c = {eigenvalues of A with zero real part}. Critical part of the spectrum for a diffeomorphism $x \to Ax$ is σ_c = {eigenvalues of A with modulus equal to 1}.

Eigenvalue and spectrum If for a matrix (linear mapping) A the equality $Av = \lambda v, v \neq 0$ holds then v and λ are called eigenvector and eigenvalue of A. The set of the eigenvalues is the spectrum of A. If there exists k such that $(A - \lambda I)^k v = 0$, v is said to be generating vector.

Equivalence of dynamical systems Two dynamical systems f and g are topologically equivalent if there is a continuous one-to-one correspondence (homeomorphism) that maps trajectories (orbits) of f on trajectories of g. It should be emphasized that the homeomorphism need not be differentiable.

Invariant manifold In applications an invariant manifold arises as a surface such that the trajectories starting on the surface remain on it under the system evolution.

Local properties If $F(p) = 0$, the point p is an equilibrium of (1). If $G(q) = q$, q is a fixed point of (2). We study dynamics near an equilibrium or a fixed point of the system. Thus we consider the system in a neighborhood of the origin which is supposed to be an equilibrium or a fixed point. In this connection we use terminology local invariant manifold or local topological equivalence.

Reduction principle In accordance with this principle a locally invariant (center) manifold corresponds to the critical part of the spectrum of the linearized system. The behavior of orbits on a center manifold determines the dynamics of the system in a neighborhood of its equilibrium or fixed point. The term "reduction principle" was introduced by V. Pliss [59].

Definition of the Subject

Let M be a subset of the Euclidean space R^n.

Definition 1 A set M is said to be smooth manifold of dimension $d \le n$ if for each point $p \in M$ there exists a neighborhood $U \subset M$ and a smooth mapping $g\colon U \to U_0 \subset R^d$ such that there is the inverse mapping g^{-1} and its differential Dg^{-1} (matrix of partial derivatives) is an injection i. e. its maximal rang is d.

For sphere in R^3 which is a two-dimensional smooth manifold the introduction of the described neighborhoods is demonstrated in [77]. A manifold M is C^k-smooth, $k \ge 1$ if the mappings g and g^{-1} have k continuous derivatives. Moreover, chosing the mappings g and g^{-1} as C^∞-smooth or analytical, we obtain the manifold with the same smoothness.

Consider a continuous dynamical system

$$\dot{x} = F(x) , \quad x \in R^n , \tag{3}$$

where $F\colon R^n \to R^n$ is a C^k-smooth vector field, $k \ge 1$, i. e. the mapping F has k continuous derivatives. Let us denote the solution of (3) passing through the point $p \in R^n$ at $t = 0$ by $X(t, p)$. Suppose that the solution is determined for all $t \in R$. By the fundamental theorems of the differential equations theory, the hypotheses imposed on F guarantee the existence, uniqueness, and smoothness of $X(t, p)$ for all $p \in R^n$. In this case the mapping $X_t(p) = X(t, p)$ under fixed t is a diffeomorphism of R^n. Thus, F generates the smooth flow $X_t\colon R^n \to R^n$, $X_t(p) = X(t, p)$. The differential $DX_t(0)$ is a fundamental matrix of the linearized system $\dot{v} = DF(0)v$. A trajectory (an orbit) of the system (3) through x_0 is the set $T(x_0) = \{x = X(t, x_0),\ t \in R\}$.

Definition 2 A manifold M is said to be invariant under (3) if for any $p \in M$ the trajectory through p lies in M. A manifold M is said to be locally invariant under (3) if for every $p \in M$ there exists, depending on p, an interval $T_1 < 0 < T_2$, such that $X_t(p) \in M$ as $t \in (T_1,\ T_2)$. It means that the invariant manifold is formed by trajectories of the system and the locally invariant manifold consists of arcs of trajectories. An equilibrium is 0-dimensional invariant manifold and a periodic orbit is 1-dimensional one. The concept of invariant manifold is a useful tool for simplification of dynamical systems.

Definition 3 A manifold M is said to be invariant in a neighborhood U if for any $p \in M \bigcap U$ the moving point $X_t(p)$ remains on M as long as $X_t(p) \in U$. In this case the manifold M is locally invariant.

Near an equilibrium point O the system (3) can be rewritten in the form

$$\dot{x} = Ax + f(x) , \tag{4}$$

where $O = \{x = 0\}$, $A = DF(O)$, f is second-order at the origin, i. e. $f(0) = 0$ and $Df(0) = 0$. Let us show that the system (4) near O can be considered as a perturbation of the linearized system

$$\dot{x} = Ax . \tag{5}$$

For this we construct a C^k-smooth mapping g which coincides with f in a sufficiently small neighborhood of the origin and is C^1-close to zero.

To construct the mapping g one uses the C^∞-smooth cut-off function $\alpha\colon R^+ \to R^+$

$$\alpha(r) = \begin{cases} 1, & r < 1/2 , \\ > 0, & 1/2 \le r \le 1 , \\ 0, & r > 1 . \end{cases}$$

Then set $g(x) = \alpha(|x|/\epsilon) f(x)$ where ε is a parameter. If $|x| > \epsilon$, $g(x) = 0$ and if $|x| < \epsilon/2$, $g(x) = f(x)$ and g is C^1-close to zero. Consider the system

$$\dot{x} = Ax + g(x) . \tag{6}$$

It is evident that the dynamics of (4) and (6) is the same on $U = \{|x| < \epsilon/2\}$. If the constructed system (6) has an invariant manifold M then $M \bigcap U$ is locally invariant manifold for the initial system (4) or, more precisely, the manifold M is invariant in U. It should be noted that for the case of infinite-dimensional phase space the described construction is more delicate (see details below).

Let us consider a C^k-smooth mapping $(k \ge 1)$ $G\colon R^n \to R^n$ which has the inverse C^k-smooth mapping G^{-1}. The mapping G generates the discrete dynamical system of the form

$$x_{m+1} = G(x_m) , \quad m = \ldots, -2, -1, 0, 1, 2, \ldots . \tag{7}$$

Each continuous system $X(t, p)$ gives rise to the discrete system $G(x) = X(1, x)$ which is the shift operator on unit time. Such a system preserves the orientation of the phase space,where as an arbitrary discrete system may change it.

Hence, the space of discrete systems is more rich than the space of continuous ones. Moreover, usually investigation of a discrete system is, as a rule, simpler than study of a differential equations one. In many instances the investigation of a differential equation may be reduced to study of a discrete system by Poincaré (first return) mapping. An orbit of (7) is the set $T = \{x = x_m,\ m \in \mathbb{Z}\}$, where x_m satisfies (7).

Near a fixed point O the mapping $x \to G(x)$ can be rewritten in the form

$$x \to Ax + f(x)\,, \tag{8}$$

where $O = \{x = 0\}$, $A = DG(O)$, f is second-order at the origin. It is shown above that near O (8) may be considered as a perturbation of the linearized system

$$x \to Ax\,. \tag{9}$$

Motivational Example

Consider the discrete dynamical system

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} X(x,y) = x + xy \\ Y(x,y) = \frac{1}{2}y + \frac{1}{2}x^2 + 2x^2y + y^3 \end{pmatrix}. \tag{10}$$

The origin $(0,0)$ is fixed point. Our task is to examine the stability of the fixed point. The linearized system at 0 is defined by the matrix

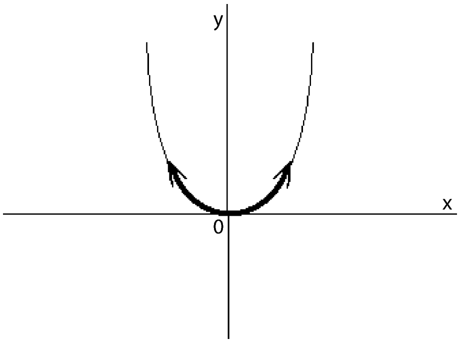
$$\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$

which has two eigenvalues 1 and 1/2. Hence, the first approximation system contracts along y-axis, whereas its action along x-axis is neutral. So the stability of the fixed point depends on nonlinear terms. Let us show that the curve $y = x^2$ is invariant for the mapping (10). It is enough to check that if a point (x, y) is on the curve then the image (X, Y) is on the curve, i. e. $Y = X^2$ as $y = x^2$. We have

$$\begin{aligned} Y|_{y=x^2} &= \tfrac{1}{2}y + \tfrac{1}{2}x^2 + 2x^2y + y^3|_{y=x^2} \\ &= x^2 + 2x^4 + x^6\,, \\ X^2|_{y=x^2} &= (x + xy)^2|_{y=x^2} \\ &= x^2 + 2x^4 + x^6\,. \end{aligned} \tag{11}$$

Thus, the curve $y = x^2$ is an invariant one-dimensional manifold which is a center manifold W^c for the system (10), see Fig. 1. The fixed point O is on W^c. The restriction of the system on the manifold is

$$x \to x + xy|_{y=x^2} = x(1 + x^2)\,. \tag{12}$$



Center Manifolds, Figure 1
Dynamics on the invariant curve $y = x^2$

It follows that the fixed point 0 is unstable and $x_m \to \infty$ as $m \to \infty$, see Fig. 1. Hence, the origin is unstable fixed point for the discrete system (10). It turns out the system (10) near O is topologically equivalent to the system

$$\begin{aligned} x &\to x(1 + x^2)\,, \\ y &\to \tfrac{1}{2}y \end{aligned} \tag{13}$$

which is simpler than (10). The center manifold $y = x^2$ is tangent to the x-axis at the origin and W^c near 0 can be considered as a perturbed manifold of $\{y = 0\}$.

Introduction

In his famous dissertation “General Problem on Stability of Motion” [41] published at 1892 in Khar’kov (Ukraine) A.M. Lyapunov proved that the equilibrium O of the system (4) is stable if all eigenvalues of matrix A have negative real parts and O is unstable if there exists an eigenvalue with positive real part. He studied the case when some eigenvalues of A have negative real part and the rest of them have zero real one. Lyapunov proved that if the matrix A has a pair of pure imaginary eigenvalues and the other eigenvalues have negative real parts then there exists a two-dimensional invariant surface M through O, and the equilibrium O is stable if O is stable for the system restricted on M. Speaking in modern terms, Lyapunov proved the existence of “center manifold” and formulated “reduction principle” under the described conditions. Moreover, he found the stability condition by using power series expansions, which is an extension of his first method to evaluate stability of systems whose eigenvalues have zero real part. Now we use the same method to check stability [10,25,31,36,43,76]. The general reduction principle in stability theory was established by V. Pliss [59] in 1964, the term “center manifold” was introduced by A. Kelley at 1967 in the paper [38] where the existence of a family of invariant manifolds through equilibrium was proved in general case.

As we see the center manifold can be considered as a perturbation of the center subspace of linearized system. H. Poincaré [61] was probably the first who perceived the importance of the perturbation problem and began to study conditions ensuring the preservation of the equilibriums and periodic orbits under a perturbation of differential equations. Hadamard [26] and Perron [57] proved the existence of stable and unstable invariant manifolds for a hyperbolic equilibrium point and, in fact, showed their preservation. The conditions necessary and sufficient for the preservation of locally invariant manifolds passing through an equilibrium have been obtained in [51]. Many results on the center manifold follow from theory of normal hyperbolicity [77] which studies dynamics of a system near a compact invariant manifold. They will be considered below. The books by Carr [10], Guckenheimer and Holmes [25], Marsden and McCracken [43], Iooss [36], Hassard, Kazarinoff and Wan [31], and Wiggins [76] are popular sources of the information about center manifolds.

Center Manifold in Ordinary Differential Equations

Consider a linear system of differential equations

$$\dot{v} = Lv\,, \tag{14}$$

where $v \in R^n$ and L is a matrix. Divide the eigenvalues of the matrix L into three parts: stable $\sigma_s = \{$eigenvalues with negative real part$\}$, unstable $\sigma_u = \{$eigenvalues with positive real part$\}$ and central (neutral, critical) $\sigma_c = \{$eigenvalues with zero real part$\}$. If the matrix does not have the eigenvalues with zero real part, the system is called hyperbolic. The matrix L has three eigenspaces: the stable subspace E^s, the unstable subspace E^u, and the central subspace E^c correspond to these parts of spectrum. The subspaces E^s, E^u, and E^c are invariant under the system (14), that is, the solution starting on the subspace stays on it [33]. The solutions on E^s have exponential tending to 0, the solutions on E^u have exponential growth. The subspaces E^s, E^c, E^u meet in pairs only at the origin and $E^s + E^c + E^u = R^n$, i. e. we have the invariant decomposition $R^n = E^s \oplus E^c \oplus E^u$. There exists a linear change of coordinates (see [33]) transforming (14) to the form

$$\begin{aligned} \dot{x} &= Ax\,,\\ \dot{y} &= By\,,\\ \dot{z} &= Cz\,, \end{aligned} \tag{15}$$

where the matrix A has eigenvalues with zero real part, the matrix B has eigenvalues with negative real part, and the matrix C has eigenvalues with positive real part. So, (14) decomposes into three independent systems of differential equations. It is known [1,23,24] that the system (15) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= Ax\,,\\ \dot{y} &= -y\,,\\ \dot{z} &= z\,. \end{aligned} \tag{16}$$

Thus, the dynamics of the system (14) is determined by the system $\dot{x} = Ax$ which is the restriction of (14) on the center subspace E^c. Our goal is to justify a similar "reduction principle" for nonlinear systems of differential equations.

Summarizing the results of V. Pliss [59,60], A. Kelley [38], N. Fenichel [19], and M. Hirsch, C. Pugh, M. Shub [34] we obtain the following theorem.

Theorem 1 (Existence Theorem) *Consider a C^k-smooth ($1 \le k < \infty$) system of differential equations $\dot{v} = F(v)$, $F(O) = 0$ and $L = DF(O)$. Let E^s, E^u and E^c be the stable, unstable, and central eigenspaces of L.*

Then near the equilibrium O there exist the following five C^k-smooth invariant manifolds (see Fig. 2):

- *The center manifold W^c, tangent to E^c at O;*
- *The stable manifold W^s, tangent to E^s at O;*
- *The unstable manifold W^u, tangent to E^u at O;*
- *The center-stable manifold W^{cs}, tangent to $E^c + E^s$ at O;*
- *The center-unstable manifold W^{cu}, tangent to $E^c + E^u$ at O;*

The stable and unstable manifolds are unique, but center, center-stable, and center-unstable manifolds are not necessarily unique.

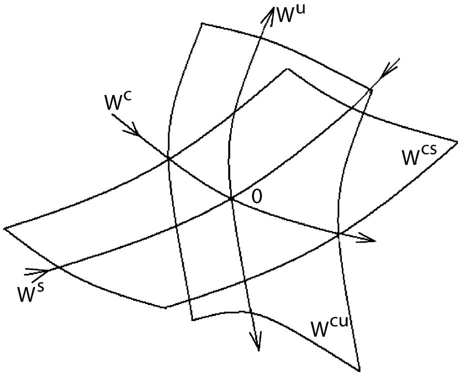
Solutions on W^s have exponential tending to O as $t \to \infty$. Solutions on W^u have exponential tending to O as $t \to -\infty$.

Remark 1 The expression "near the equilibrium O" means that there exists a neighborhood $U(O)$, depending on F and k, where the statements of the Existence Theorem hold.

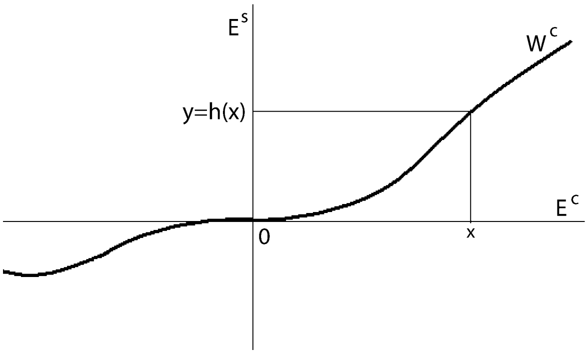
Remark 2 The linearized system $\dot{v} = DF(O)v$ has the invariant subspace $E^s + E^u$ whereas the complete system $\dot{v} = F(v)$ may not have any smooth invariant manifolds with the tangent space $E^s + E^u$ at O, see [5,30].

Representation of the Center Manifold

At first we suppose that our system does not have the unstable eigenspace and is transformed by the linear change



Center Manifolds, Figure 2
The classical five invariant manifolds



Center Manifolds, Figure 3
Representation of a center manifold

of coordinates mentioned above to the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y)\,, \\ \dot{y} &= By + g(x, y)\,, \end{aligned} \tag{17}$$

where $E^c = \{(x, 0)\}$, $E^s = \{(0, y)\}$. Since the center manifold W^c is tangent to $E^c = \{(x, 0)\}$ at the origin $O = (0, 0)$, it can be represented near O in the form

$$W^c = \{(x, y) \mid |x| < \varepsilon,\ y = h(x)\}\,.$$

In other words the center manifold is represented as the graph of a smooth mapping $h\colon V \subset R^c \to R^s$, where c and s are dimensions of the center and stable subspaces, see Fig. 3. Since W^c goes through the origin and is tangent to E^s, we have the equalities

$$h(0) = 0\,, \quad Dh(0) = 0\,. \tag{18}$$

The invariance of W^c means that if an initial point (x, y) is in W^c, i. e. $y = h(x)$ then the solution $(X(t, x, y), Y(t, x, y))$ of (17) is in W^c, i. e. $Y(t, x, y)) = h(X(t, x, y))$. Thus, we get the invariance condition

$$Y(t, x, h(x))) = h(X(t, x, h(x)))\,. \tag{19}$$

Differentiating (19) with respect to t and putting $t = 0$ we get

$$Bh(x) + g(x, h(x)) = Dh(x)(Ax + f(x, h(x)))\,. \tag{20}$$

Thus, the mapping h have to be a solution of the partial differential equation (20) and satisfy (18).

Suppose that the system has the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y, z)\,, \\ \dot{y} &= By + g_1(x, y, z)\,, \\ \dot{z} &= Cz + g_2(x, y, z)\,, \end{aligned} \tag{21}$$

where A has the eigenvalues with zero real part, B has the eigenvalues with negative real part, C has the eigenvalues with positive real part. Analogously to the previous case we can show that the center manifold is represented in the form

$$W^c = \{(x, y, z) \mid |x| < \epsilon,\ y = h_1(x),\ z = h_2(x)\}$$

where the mappings h_1 and h_2 satisfy the equations

$$\begin{aligned} &Bh_1(x) + g_1(x, h_1(x), h_2(x)) \\ &\qquad = Dh_1(x)(Ax + f(x, h_1(x), h_2(x))) \\ &Ch_2(x) + g_2(x, h_1(x), h_2(x)) \\ &\qquad = Dh_2(x)(Ax + f(x, h_1(x), h_2(x))) \end{aligned} \tag{22}$$

and the conditions

$$\begin{aligned} h_1(0) &= 0\,, \quad Dh_1(0) = 0 \\ h_2(0) &= 0\,, \quad Dh_2(0) = 0\,. \end{aligned} \tag{23}$$

Uniqueness and Nonuniqueness of the Center Manifold

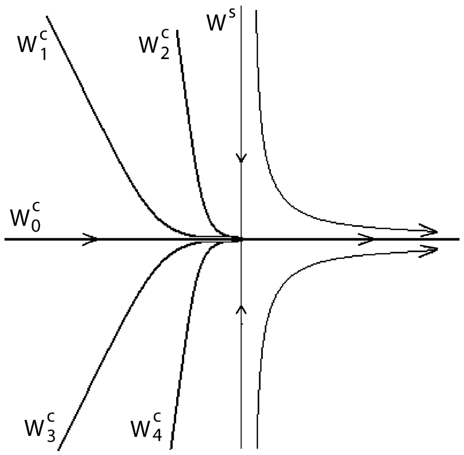
The center manifold is nonunique in general.

Let us consider the illustrating example from [38]. The system has the form

$$\begin{aligned} \dot{x} &= x^2\,, \\ \dot{y} &= -y\,, \end{aligned} \tag{24}$$

where $(x, y) \in R^2$. It is obviously that the origin $(0, 0)$ is the equilibrium with single stable manifold $W^s = \{(0, y)\}$. There is a center manifold of the form $W^c = \{(x, 0)\}$. Moreover, there are the center manifolds that can be obtained when solving the equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{-y}{x^2}\,. \tag{25}$$



Center Manifolds, Figure 4
Nonuniqueness of the center manifold

The solution of (24) has the form

$$y(x) = a \exp\left(\frac{1}{x}\right) \tag{26}$$

for $x \neq 0$ and any constant $a \in R$. It follows that

$$W^c(a) = \{(x, y) | y = a \exp\left(\frac{1}{x}\right) \\ \text{as } x < 0 \text{ and } y = 0 \text{ as } x \geq 0\}$$

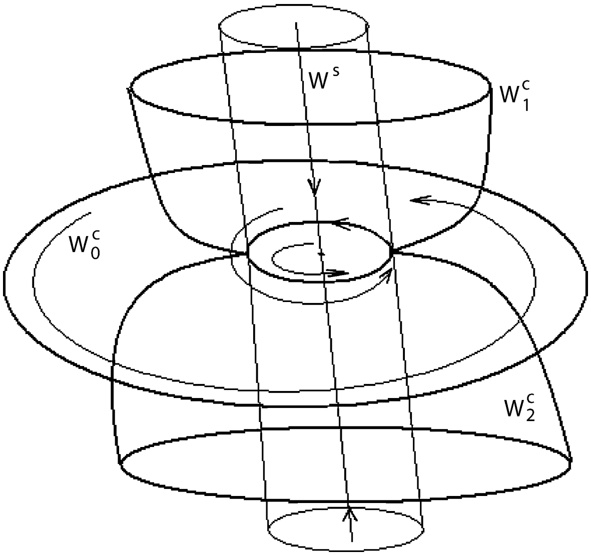
is center manifold for each a, being C^∞-smooth and pasting together the curve $\{y = a \exp(1/x)\}$ as $x < 0$ and the straight line $\{y = 0\}$ as $x \geq 0$., see Fig. 4.

However, the center manifolds possess the following weak uniqueness property. Let U be a neighborhood of the W^c. It turns out that the maximal invariant set I in U must be in W^c, that follows from the Reduction Theorem, see below. Consequently, all center manifolds contact on I. In this case the center manifold is called locally maximal. Thus, the center manifolds may differ on the trajectories leaving the neighborhood as $t \to \infty$ or $-\infty$.

For example, suppose that a center manifold is two-dimensional and a limit cycle is generated from the equilibrium through the Hopf bifurcation [31,35]. In this case the invariant set I is a disk bounded by the limit cycle and all center manifolds will contain this disk, see Fig. 5.

Smoothness

M. Hirsch, C. Pugh, M. Shub [34] proved that the center manifold is C^k-smooth if the system is C^k-smooth and $k < \infty$. Moreover if the kth derivative of the vector field is α-Hölder or Lipschitz mapping, the center manifold is the same.



Center Manifolds, Figure 5
Weak uniqueness of the center manifold

However, if the system is C^∞ or analytic then the center manifold is not necessary the same. First consider the analytic case. It is clear that if the analytic center manifold exists then it is uniquely determined by applying the Taylor power series expansion. Consider the illustrating example [31]

$$\begin{aligned} \dot{x} &= -x^2 , \\ \dot{y} &= -y + x^2 \end{aligned} \tag{27}$$

which does not have an analytic center manifold. In fact, applying the Taylor power series expansions we obtain that the center manifold has to be given by the series $y = \sum_{n=2}^{\infty}(n-1)!x^n$ which diverges when $x \neq 0$.

If the system is C^∞-smooth then the Existence Theorem guaranties the existence of a C^k-smooth center manifold for any $k < \infty$. However, the center manifold may not be C^∞-smooth. Van Strien S.J. [71] showed that the system

$$\begin{aligned} \dot{x} &= -x^2 + \mu^2 , \\ \dot{y} &= -y - (x^2 - \mu^2) , \\ \dot{\mu} &= 0 . \end{aligned} \tag{28}$$

does not have a C^∞-smooth center manifold. In fact if the system is C^∞-smooth then for each k there is a neighborhood $U_k(O)$ of a C^k-smooth central manifold W^c. There are the systems for which the sequence $\{U_k\}$ can shrink to O as $k \to \infty$ [25,71]. The results of the papers [19,34] show that the smoothness of the invariant manifold depends on the relation between Lyapunov exponents on the center manifold and on the normal subspace (stable plus unstable subspaces). This relation σ is included in the concept of the normal hyperbolicity. At an equilibrium $\sigma = \min(\mathrm{Re}\lambda_n/\mathrm{Re}\lambda_c)$, where $\mathrm{Re}\lambda_n/\mathrm{Re}\lambda_c > 0$, λ_c is the eigenvalue on the center subspace and λ_n is the eigenvalue on the normal subspace. The condition $\sigma > 1$ is necessary for a persistence of smooth invariant manifold. One can guarantee degree of smoothness k of the manifold provided $k < \sigma$. Moreover, there exists examples showing that the condition $k < \sigma$ is essential. It means that the system has to contract (expand) along $E^s(E^u)k$-times stronger than along the center manifold.

If a center manifold to the equilibrium O exists then $\mathrm{Re}\,\lambda_c = 0$ and $\sigma = \infty$ at O, but near O may be other equilibrium (or other orbits) on the center manifold where $\sigma < \infty$, and as a consequence, the center manifold may not be C^∞-smooth. Let us consider the illustrating example [25]

$$\begin{aligned} \dot{x} &= \mu x - x^3\,, \\ \dot{y} &= y + x^4\,, \\ \dot{\mu} &= 0\,. \end{aligned} \tag{29}$$

The point $O = (0, 0, 0)$ is equilibrium, the system linearized at O has the form

$$\begin{aligned} \dot{x} &= 0\,, \\ \dot{y} &= y\,, \\ \dot{\mu} &= 0\,. \end{aligned} \tag{30}$$

The system (30) has the following invariant subspace: $E^s = \{0, 0, 0\}$, $E^c = \{x, 0, \mu\}$, $E^u = \{0, y, 0\}$. The μ-axis consists of the equilibriums $(0, 0, \mu_0)$ for the system (29). The system linearized at $(0, \mu_0, 0)$ is the following

$$\begin{aligned} \dot{x} &= \mu_0 x\,, \\ \dot{y} &= y\,, \\ \dot{\mu} &= 0\,. \end{aligned} \tag{31}$$

The eigenvalues on the center subspace are 0 and μ_0, the eigenvalue on the normal (unstable) subspace is 1. Therefore the degree of smoothness is bounded by $\sigma = 1/\mu_0$. Detailed information and examples are given in [10,25,31,36,43,76].

Reduction Principle

As we saw above, a smooth system of differential equations near an equilibrium point O can be written in the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y, z)\,, \\ \dot{y} &= By + g(x, y, z)\,, \\ \dot{z} &= Cz + q(x, y, z)\,, \end{aligned} \tag{32}$$

where $O = (0, 0, 0)$, A has eigenvalues with zero real part, B has eigenvalues with negative real part and C has eigenvalues with positive real part; $f(0, 0, 0) = g(0, 0, 0) = q(0, 0, 0) = 0$ and $Df(0, 0, 0) = Dg(0, 0, 0) = Dq(0, 0, 0) = 0$. In this case the invariant subspaces at O are $E^c = \{(x, 0, 0)\}$, $E^s = \{(0, y, 0)\}$, and $E^u = \{(0, 0, z)\}$. The center manifold has the form

$$W^c = \{(x, y, z) | x \in V \subset R^c,\ y = h_1(x),\ z = h_2(x)\}\,, \tag{33}$$

where c is the dimension of the center subspace E^c, the mappings $h_1(x)$ and $h_2(x)$ are C^k-smooth, $k \geq 1$, $h_1(0) = h_2(0) = 0$ and $Dh_1(0) = Dh_2(0) = 0$. The last equalities mean that the manifold W^c goes through O and is tangent to E^c at O.

Summarizing the results of V. Pliss [59,60], A. Shoshitaishvili [68,69], A. Reinfelds [63], K. Palmer [55], Pugh C.C., Shub M. [62], Carr J. [10], Grobman D.M. [24], and F. Hartman [29] we obtain the following theorem.

Theorem 2 (Reduction Theorem) *The system of differential Equations (32) near the origin is topologically equivalent to the system*

$$\begin{aligned} \dot{x} &= Ax + f(x, h_1(x), h_2(x))\,, \\ \dot{y} &= -y\,, \\ \dot{z} &= z\,. \end{aligned} \tag{34}$$

The first equation is the restriction of the system on the center manifold. The theorem allows to reduce the investigation of the dynamics near a nonhyperbolic fixed point to the study of the system on the center manifold.

Construction of the Center Manifold

The center manifold may be calculated by using simple iterations [10]. However, this method does not have wide application and we consider a method based on the Taylor power series expansion. First it was proposed by A. Lyapunov in [41]. Suppose that the unstable subspace is 0, that is the last equation of the system (32) is absent. Such systems have a lot of applications in stability theory. So we consider a system

$$\begin{aligned} \dot{x} &= Ax + f(x, y)\,, \\ \dot{y} &= By + g(x, y)\,, \end{aligned} \tag{35}$$

The center manifold is the graph of the mapping $y = h(x)$ which satisfies the equation

$$Dh(x)(Ax + f(x, h(x))) - Bh(x) - g(x, h(x)) = 0 \quad (36)$$

and the conditions $h(0) = 0$ and $Dh(0) = 0$. Let us try to solve the equation by applying the Taylor power series expansion. Denote the left part of (36) by N(h(x)). J. Carr [10] and D. Henry [32] proved the following theorem.

Theorem 3 *Let $\phi\colon V(0) \subset R^c \to R^s$ be a smooth mapping, $\phi(0)=0$ and $D\phi(0)=0$ such that $N(\phi(x))=o(|x|^m)$ for some $m > 0$ as $|x| \to 0$ then*

$$h(x) - \phi(x) = o(|x|^m) \text{ as } |x| \to 0$$

where $r(x)=o(|x|^m)$ means that $r(x)/|x|^m \to 0$ as $|x| \to 0$.

Thus, if we solve the Eq. (36) with a desired accuracy we construct h with the same accuracy. Theorem 3 substantiates the application of the described method. Let us consider a simple example from [76]

$$\begin{aligned} \dot{x} &= x^2 y - x^5 , \\ \dot{y} &= -y + x^2 , \end{aligned} \quad (37)$$

$(x, y) \in R^2$, the equilibrium is at the origin $(0, 0)$. The eigenvalues of the linearized system are 0 and -1. According to the Existence Theorem the center manifold is locally represented in the form

$$W^c = \{(x, y) \mid y = h(x),\ |x| < \delta, \\ h(0) = 0, Dh(0) = 0)\} ,$$

where δ is sufficiently small. It follows that h has the form

$$h = ax^2 + bx^3 + \cdots . \quad (38)$$

The equation for the center manifold is given by

$$Dh(x)(Ax + f(x, h(x))) - Bh(x) - g(x, h(x)) = 0 , \quad (39)$$

where $A = 0$, $B = -1$, $f(x, y) = x^2 y - x^5$, $g(x, y) = x^2$. Substituting (38) in (39) we obtain the equality

$$(2ax + 3bx^2 + \cdots)(ax^4 + bx^5 - x^5 + \cdots) \\ + ax^2 + bx^3 - x^2 + \cdots = 0 . \quad (40)$$

Equating coefficients on each power of x to zero we obtain

$$\begin{aligned} x^2 &\ :\ a - 1 = 0,\ \Rightarrow\ a = 1 \\ x^3 &\ :\ b = 0, \\ \vdots &\quad\ \ \vdots . \end{aligned} \quad (41)$$

Therefore we have

$$h(x) = x^2 + 0x^3 + \cdots . \quad (42)$$

In this connection we have to decide how many terms (powers of x) have be computed? The solution depends on our goal. For example, suppose that we study the Lyapunov stability of the equilibrium $(0, 0)$ for (37). According to the Reduction Theorem the system (37) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= x^2 h(x) - x^5 , \\ \dot{y} &= -y , \end{aligned} \quad (43)$$

where $h(x) = x^2 + 0x^3 + \cdots$. Substituting h we obtain the equation

$$\dot{x} = x^4 + \dots . \quad (44)$$

Hence the equilibrium is unstable. It should be noted that the calculation of bx^3 is unnecessary, since substituting $h(x) = x^2 + bx^3 + \cdots$ in the first equation of the system (43), we also obtain (44). Thus, it is enough to compute the first term ax^2 of the mapping h.

This example brings up the following question. The system (37) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= x^4 + \cdots , \\ \dot{y} &= -y , \end{aligned} \quad (45)$$

which has many center manifolds. From this it follows that the initial system has a lot of center manifolds. Which of center manifolds is actually being found when approximating the center manifold via power series expansion?

It turns out [10,70,74] that any two center manifolds differ by transcendentally small terms, i. e. the terms of their Taylor expansions coincide up to all existing orders. For example, the system (24) has the center manifolds of the form

$$W^c(a) = \left\{(x, y) | y = a \exp\left(\tfrac{1}{x}\right) \\ \text{as } x < 0 \text{ and } y = 0 \text{ as } x \geq 0\right\}$$

for any a. However, each manifold has null-expansion at 0.

Center Manifold in Discrete Dynamical Systems

Consider a dynamical system generated by the mapping

$$v \to Lv + G(v) , \quad (46)$$

where $v \in R^n$, L is a matrix and G is second-order at the origin. Without loss of generality we consider the system (46) as a perturbation of the linear system

$$x \to Lx . \quad (47)$$

Divide the eigenvalues of L into three parts:
stable $\sigma_s = \{$eigenvalues with modulus less than $1\}$,
unstable $\sigma_u = \{$eigenvalues with modulus greater than $1\}$,
critical $\sigma_c = \{$eigenvalues with modulus equal to $1\}$.
The matrix L has three eigenspaces: the stable subspace E^s, the unstable subspace E^u, and the central subspace E^c, which correspond to the mentioned spectrum parts respectively, being $E^s \oplus E^u \oplus E^c = R^n$. The next theorem follows from the theorem on perturbation of ρ-hyperbolic endomorphism on a Banach space [34].

Theorem 4 (Existence Theorem) *Consider a C^k-smooth $(1 \leq k < \infty)$ discrete system*

$$v \to Lv + G(v)\,,$$

$G(0) = 0$ and G is C^1-close to zero. Let E^s, E^u and E^c be the stable, unstable, and central eigenspaces of the linear mapping L.

Then near the fixed point O there exist the following C^k-smooth invariant manifolds:

- *The center manifold W^c tangent to E^c at O,*
- *The stable manifold W^s tangent to E^s at O,*
- *The unstable manifold W^u tangent to E^u at O,*
- *The center-stable manifold W^{cs} tangent to $E^c + E^s$ at O,*
- *The center-unstable manifold W^{cu} tangent to $E^c + E^u$ at O.*

The stable and unstable manifolds are unique, but center, center-stable, and center-unstable manifolds may not be unique.

The orbits on W^s have exponential tending to O as $m \to \infty$. The orbits on W^u have exponential tending to O as $m \to -\infty$.

There exists a linear change of coordinates transforming (46) to the form

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \to \begin{pmatrix} Ax + f(x,y,z) \\ By + g_1(x,y,z) \\ Cz + g_2(x,y,z) \end{pmatrix}\,, \tag{48}$$

where A has eigenvalues with modulus equal 1, B has eigenvalues with modulus less than 1, C has eigenvalues with modulus greater than 1. The system (48) at the origin has the following invariant eigenspaces: $E^c = \{(x,0,0)\}$, $E^s = \{(0,y,0)\}$, and $E^u = \{(0,0,z)\}$. Since the center manifold W^c is tangent to E^c at the origin $O = (0,0)$, it can be represented near O in the form

$$W^c = \{(x,y,z) \,|\, |x| < \epsilon,\; y = h_1(x), z = h_2(x)\}\,,$$

where h_1 and h_2 are second-order at the origin, i. e. $h_{1,2}(0) = 0$, $Dh_{1,2}(0) = 0$. The invariance of W^c means that if a point $(x,y,z) \in W^c$, i. e. $y = h_1(x)$, $z = h_2(x)$, then $(Ax + f(x,y,z), By + g_1(x,y,z), Cz + g_2(x,y,z)) \in W^c$, i. e. $By + g_1(x,y,z) = h_1(Ax + f(x,y,z))$, $Cz + g_2(x,y,z) = h_2(Ax + f(x,y,z))$. Thus, we get the invariance property

$$\begin{aligned} &Bh_1(x) + g_1(x,h_1(x),h_2(x)) \\ &\qquad = h_1(Ax + f(x,h_1(x),h_2(x)))\,, \\ &Ch_2(x) + g_2(x,h_1(x),h_2(x)) \\ &\qquad = h_2(Ax + f(x,h_1(x),h_2(x)))\,. \end{aligned} \tag{49}$$

Results on smoothness and uniqueness of center manifold for discrete systems are the same as for continuous systems.

Theorem 5 (Reduction Theorem [55,62,63,64]) *The discrete system (48) near the origin is topologically equivalent to the system*

$$\begin{aligned} x &\to Ax + f(x,h_1(x),h_2(x))\,, \\ y &\to By\,, \\ z &\to Cz\,. \end{aligned} \tag{50}$$

The first mapping is the restriction of the system on the center manifold. The theorem reduces the investigation of dynamics near a nonhyperbolic fixed point to the study of the system on the center manifold.

Normally Hyperbolic Invariant Manifolds

As it is indicated above, a center manifold can be considered as a perturbed invariant manifold for the center subspace of the linearized system. The normal hyperbolicity conception arises as a natural condition for the persistence of invariant manifold under a system perturbation. Informally speaking, an f-invariant manifold M, where f-diffeomorphism, is normally hyperbolic if the action of Df along the normal space of M is hyperbolic and dominates over its action along the tangent one. (the rigorous definition is below). Many of the results concerning center manifold follows from the theory of normal hyperbolic invariant manifolds. In particular, the results related to center eigenvalues of Df with $|\lambda| \neq 1$ may be obtained using this theory.

The problem of existence and preservation for invariant manifolds has a long history. Initial results on invariant (integral) manifolds of differential equations were obtaned by Hadamard [26] and Perron [57], Bogoliubov, Mitropolskii [8] and Lykova [48], Pliss [60], Hale [28], Diliberto [16] and other authors (for references see [39,77]). In the late 1960s and throughout 1970s the theory of perturbations of invariant manifolds began to assume a very general and well-developed form. Sacker [66] and Neimark [50] proved (independently and by different methods) that a normally hyperbolic compact invariant manifold is preserved under C^1 perturbations. Final results on preservation of invariant manifolds were obtained by Fenichel [19,20] and Hirsch, Pugh, Shub [34]. The linearization theorem for a normally hyperbolic manifold was proved in [62].

Let $f\colon R^n \to R^n$ be a C^r-diffeomorphism, $1 \le r < \infty$, and a compact manifold M is f-invariant.

Definition 4 An invariant manifold M is called r-normally hyperbolic if there exists a continuous invariant decomposition

$$TR^n|_M = TM \oplus E^{\mathrm{s}} \oplus E^{\mathrm{u}} \tag{51}$$

of the tangent bundle TR^n into a direct sum of subbundles TM, E^{s}, E^{u} and constants a, $\lambda > 0$ such that for $0 \le k \le r$ and all $p \in M$,

$$\begin{aligned} |D^s f^n(p)|\ |(D^0 f^n(p))^{-1}|^k &\le a\exp(-\lambda n) \quad \text{for } n \ge 0\,, \\ |D^u f^n(p)|\ |(D^0 f^n(p))^{-1}|^k &\le a\exp(\lambda n) \quad \text{for } n \le 0\,. \end{aligned} \tag{52}$$

Here $D^0 f$, $D^{\mathrm{s}} f$ and $D^{\mathrm{u}} f$ are restrictions of Df on TM, E^{s} and E^{u}, respectively.

The invariance of the bundle E^* means that for every $p \in M_0$, the image of the fiber $E^*(p)$ at p under $Df(p)$ is the fiber $E^*(q)$ at $q = f(p)$. The bundles E^{s} and E^{u} are called stable and unstable, respectively. In other words, M is normally hyperbolic if the differential Df contracts (expands) along the normal (to M) direction and this contraction (expansion) is stronger in r times than a conceivable contraction (expansion) along M. Summarizing the results of [19,20,34,50,62,66] we obtain the following theorem.

Theorem 6 *Let a C^r diffeomorphism f be r-normally hyperbolic on a compact invariant manifold M with the decomposition $TR^n|_M = TM \oplus E^{\mathrm{s}} \oplus E^{\mathrm{u}}$. Then*

- *there exist invariant manifolds W^{s} and W^{u} near M, which are tangent at M to $TM \oplus E^{\mathrm{s}}$ and $TM \oplus E^{\mathrm{u}}$;*
- *the manifolds M, W^{s} and W^{u} are C^r-smooth;*
- *if g is another C^r-diffeomorphism C^r-close to f, then there exists the unique manifold M_g which is invariant and r-normally hyperbolic for g;*
- *near M f is topologically equivalent to $Df|_{E^{\mathrm{s}}\oplus E^{\mathrm{u}}}$, which in local coordinates has the form*

$$\begin{aligned} &(x, y, z) \to (f(x), Df(x)y, Df(x)z)\,, \\ &\qquad x \in M\,, \quad y \in E^{\mathrm{s}}(x)\,, \quad z \in E^{\mathrm{u}}(x)\,. \end{aligned} \tag{53}$$

Similarly result takes place for flows. Mañé [44] showed that a locally unique, preserved, compact invariant manifold is necessarily normally hyperbolic. Local uniqueness means that near M, an invariant set I of the perturbation g is in M_g. Conditions for the preservation of locally nonunique compact invariant manifolds and linearization were found by G. Osipenko [52,53].

Applications

Stability

As it was mentioned above, it was Lyapunov who pioneered in applying the concept of center manifold to establish the stability of equilibrium. V. Pliss [59] proved the general reduction principle in stability theory. According to this principle the equilibrium is stable if and only if it is stable on the center manifold. We have considered the motivating example where the center manifold concept was used. The system

$$\dot{x} = xy\,, \quad \dot{y} = -y + ax^2\,, \tag{54}$$

has the center manifold $y = h(x)$. Applying Theorem 3, we obtain

$$h(x) = ax^2 + \cdots\,.$$

The reduced system is of the form

$$\dot{x} = ax^3 + \ldots\,.$$

Thus, if $a < 0$, the origin is stable. The book [25] contains other examples of investigation of stability.

Bifurcations

Consider a parametrized system of differential equations

$$\dot{x} = A(\mu)x + f(x, \mu)\,, \tag{55}$$

where $\mu \in R^k$, f is C^1-small when μ is small. To study the bifurcation problem near $\mu = 0$ it is useful to deal with the extended system of the form

$$\begin{aligned} \dot{x} &= A(\mu)x + f(x, \mu)\,, \\ \dot{\mu} &= 0\,. \end{aligned} \tag{56}$$

Suppose that (55) has a n-dimensional center manifold for $\mu = 0$. Then the extended system (56) has a $n + k$-dimensional center manifold. The reduction principle guarantees that all bifurcations lie on the center manifold. Moreover, the center manifold has the form $y = h(x, \mu_0)$ for every μ_0 which is a solution of the equation $\dot{\mu} = 0$ and the mapping h may be presented as a power series in μ. Further interesting information regarding to recent advancements in the approximation and computation of center manifolds is in [37]. The book [31] deals with Hopf bifurcation and contains a lot of examples and applications. Invariant manifolds and foliations for nonautonomous systems are considered in [3,13]. Partial linearization for noninvertible mappings is studied in [4].

Center Manifold in Infinite-Dimensional Space

Center manifold theory is a standard tool to study the dynamics of discrete and continuous dynamical systems in infinite dimensional phase space. We start with the discrete dynamical system generated by a mapping of a Banach space. The existence of the center manifold of the mappings in Banach space was proved by Hirsch, Pugh and Shub [34]. Let $T\colon E \to E$ be a linear endomorphism of a Banach space E.

Definition 5 A linear endomorphism T is said to be ρ-hyperbolic, $\rho > 0$, if no eigenvalue of T has modulus ρ, i. e., its spectrum Spect(T) has no points on the circle of radius ρ in the complex plane C. A linear 1-hyperbolic isomorphism T is called hyperbolic.

For a ρ-hyperbolic linear endomorphism there exists a T-invariant splitting of $E = E_1 \oplus E_2$ such that the spectrum of the restriction $T_1 = T|_{E_1}$ lies outside of the disk of the radius ρ, where as the spectrum of $T_2 = T|_{E_2}$ lies inside it. So T_1 is automorphism and T_1^{-1} exists. It is known (for details see [34]) that one can define norms in E_1 and E_2 such that associated norms of T_1^{-1} and T_2 may be estimated in the following manner

$$|T_1^{-1}| < \tfrac{1}{\rho}\,, \quad |T_2| < \rho\,. \tag{57}$$

Conversely, if T admits an invariant splitting $T = T_1 \oplus T_2$ with $|T_1^{-1}| \equiv a$, $|T_2| \equiv b$, $ab < 1$ then Spect(T_1) lies outside the disk $\{\lambda \in C\colon |\lambda| < 1/a\}$, and Spect($T_2$) lies in the disk $\{\lambda \in C\colon |\lambda| \le b\}$. Thus, T is ρ-hyperbolic with $b < \rho < 1/a$.

Theorem 7 *Let T be a ρ-hyperbolic linear endomorphism, $\rho \le 1$. Assume that $f\colon E \to E$ is C^r, $r \ge 1$, $f = T + g$, $f(0) = 0$ and there is $\delta > 0$ such that if g is a Lipschitz mapping and*

$$\mathrm{Lip}(g) = \mathrm{Lip}(f - T) \le \delta\,, \tag{58}$$

then there exists a manifold W_f, which is a graph of a C^1 map $\varphi\colon E_1 \to E_2$, i. e.

$$W_f = \{x + y \mid y = \varphi(x),\ x \in E_1,\ y \in E_2\}\,,$$

with the following properties:

- *W_f is f-invariant;*
- *if $\|T_1^{-1}\|^j \|T_2\| < 1$, $j = 1, \dots, r$ then W_f is C^r and depends continuously on f in the C^r topology;*
- *$W_T = E_1$;*
- *for $x \in W_f$*

$$|f^{-1}(x)| \le (a + 2\varepsilon)\,|x|\,, \tag{59}$$

where $a < 1/\rho$, and ε is small provided δ is small;
- *if $Df(0) = T$ then W_f is tangent to E_1 at 0.*

Suppose that the spectrum of T is contained in $A_1 \bigcup A_2$ where

$$A_1 = \{z \in C,\ |z| \ge 1\}\,, \quad A_2 = \{z \in C,\ |z| \le a < 1\}\,.$$

Corollary 1 *If $f\colon E \to E$ is C^r, $1 \le r < \infty$, $f(0) = 0$ and $\mathrm{Lip}(f - T) \le \epsilon$ is small, then there exists a center-unstable manifold $W^{cu} = W_f$ which is the graph of a C^r function $E_1 \to E_2$. The center-unstable manifold is attractor, i. e. for any $x \in E$ the distance between $f^n(x)$ and W^{cu} tends to zero as $n \to \infty$. The manifold $W_f = W^c$ is center manifold when $A_1 = \{z \in C,\ |z| = 1\}$.*

Apply this theorem to a mapping of the general form

$$u \to Au + f(u)\,, \tag{60}$$

where E is a Banach space, $u \in E$, $A\colon E \to E$ is a linear operator, f is second-order at the origin. The question arises of whether there exists a smooth mapping g such that g coincides with f in a sufficiently small neighborhood of the origin and g is C^1-close to zero? The answer is positive if E admits C^∞-norm, i. e. the mapping $u \to \|u\|$ is C^∞-smooth for $u \neq 0$. It is performed for Hilbert space. The desired mapping g may be constructed with the use of a cut-off function. To apply Theorem 7 to (60) we have to suppose *Hypothesis on C^∞-norm*: the Banach space E admits C^∞-norm.

Thus, Theorem 7 together with Hypothesis on C^∞-norm guarantee an existence of center manifold for the system (60) in a Banach space. A. Reinfelds [63,64] prove a reduction theorem for homeomorphism in a metric space from which an analog of Reduction Theorem 5 for Banach space follows.

Flows in Banach Space

Fundamentals of center manifold theory for flows and differential equations in Banach spaces can be found in [10,21,43,73]. In Euclidean space a flow is generated by a solution of a differential equation and the theory of ordinary differential equations guarantees the clear connection between flows and differential equations. In infinite dimensional (Banach) space this connection is more delicate and at first, we consider infinite dimensional flows and then (partial) differential equations.

Definition 6 A flow (semiflow) on a domain D is a continuous family of mappings $F_t\colon D \to D$ where $t \in R\,(t \geq 0)$ such that $F_0 = I$ and $F_{t+s} = F_t F_s$ for $t, s \in R\,(t, s \geq 0)$.

Theorem 8 (Center Manifold for Infinite Dimensional Flows [43]) *Let the Hypothesis on C^{∞}-norm be fulfilled and $F_t\colon E \to E$ be a semiflow defined near the origin for $0 \leq t \leq \tau$, $F_t(0) = 0$. Suppose that the mapping $F_t(x)$ is C^0 in t and C^{k+1} in x and the spectrum of the linear semigroup $DF_t(0)$ is of the form $\mathrm{e}^{t(\sigma_1 \cup \sigma_2)}$, where $\mathrm{e}^{t(\sigma_1)}$ is on the unit circle (i. e. σ_1 is on the imaginary axis) and $\mathrm{e}^{t(\sigma_2)}$ is inside the unit circle with a positive distance from one as $t > 0$ (i. e. σ_2 is in the left half-plane). Let $E = E_1 \oplus E_2$ be the $DF_t(0)$-invariant decomposition corresponding to the decomposition of the spectrum $\mathrm{e}^{t(\sigma_1 \cup \sigma_2)}$ and $dimE_1 = d < \infty$.*

Then there exists a neighborhood U of the origin and C^k-submanifold $W^c \subset U$ of the dimension d such that

- *W^c is invariant for F_t in U, $0 \in W^c$ and W^c is tangent to E_1 at 0;*
- *W^c is attractor in U, i. e. if $F_t^n(x) \in U$, $n = 0, 1, 2, \ldots$ then $F_t^n(x) \to W^c$ as $n \to \infty$.*

In infinite dimensional dynamics it is also popular notion of "inertial manifold" which is an invariant finite dimensional attracting manifold corresponding to W^c of the Theorem 8.

Partial Differential Equations

At the present the center manifold method takes root in theory of partial differential equations. Applying center manifold to partial differential equations leads to a number of new problems. Consider an evolution equation

$$\dot{u} = Au + f(u)\,, \tag{61}$$

where $u \in E$, E is a Banach space, A is a linear operator, f is second-order at the origin. Usually A is a differential operator defined on a domain $D \subset E$, $D \neq E$. To apply Theorem 8 we have to construct a semiflow F_t defined near the origin of E. As a rule, an infinite dimensional phase space E is a functional space where the topology may be chosen by the investigator. A proper choice of functional space and its topology may essentially facilitate the study. As an example we consider a nonlinear heat equation

$$u'_t = \Delta u + f(u)\,, \tag{62}$$

where u is defined on a bounded domain $\Omega \subset R^n$ with Dirichlet condition on the boundary $u|_{\partial\Omega} = 0$, $\Delta u = (\partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_n^2)u$ is the Laplace operator defined on the set

$$C_0^2 = \{u \in C^2(\overline{\Omega}),\ u = 0 \text{ on } \partial\Omega\}\,.$$

According to [49] we consider a Hilbert space $E = L^2(\Omega)$ with the inner product $< u, v >= \int_\Omega uvdx$. The operator Δ may be extended to a self-conjugate operator $A\colon D_A \to L^2(\Omega)$ with the domain D_A that is the closure of C_0^2 in $L^2(\Omega)$. Consider the linear heat equation $u'_t = \Delta u$ and its extension

$$u'_t = Au, \quad u \in D_A\,. \tag{63}$$

What do we mean by a flow generated by (63)?

Definition 7 A linear operator A is the infinitesimal generator of a continuous semigroup $U(t)$, $t \geq 0$ if:
$U(t)$ is a linear mapping for each t;
$U(t)$ is a semigroup and $\|U(t)u - u\| \to 0$ as $t \to +0$;

$$Au = \lim_{t \to +0} \frac{U(t)u - u}{t} \tag{64}$$

whenever the limit exists.

For Laplace operator we have $< \Delta u, u >\leq 0$ for $u \in C_0^2$, hence the operator A is dissipative. It guarantees [6] the existence of the semigroup $U(t)$ with the infinitesimal generator A. The mapping $U(t)$ is the semigroup $DF_t(0)$ in Theorem 8.

The next problem is the relation between the linearized system $u' = Au$ and full nonlinear Equation (61). To consider the mapping f in (62) as C^1-small perturbation, the Hypothesis on C^{∞}-norm has to be fulfilled. To prove the existence of the solution of the nonlinear equation with $u(0) = u_0$ one uses the Duhamel formula

$$u(t) = U_t u_0 + \int_0^t U_{t-s} f(u(s))\mathrm{d}s$$

and the iterative Picard method. By this way we construct the semiflow required to apply Theorem 8. The relation between the spectrum of a semigroup $U(t)$ and the spectrum of its infinitisemal generator gives rise to an essential problem. All known existence theorems are formulated in terms of the semigroup as in Theorem 8. The existence of the spectrum decompositions of a semigroup $U(t)$ and the corresponding estimations on the appropriate projections are supposed. This framework is inconvenient for many applications, especially in partial differential equations. In finite dimensions, a linear system $\dot{x} = Ax$ has a solution of the form $U(t)x = \mathrm{e}^{At}x$ and λ is an eigenvalue of A if and only if $\mathrm{e}^{\lambda t}$ is an eigenvalue of $U(t)$. We have the spectral equality

$$\mathrm{Spect}(U(t)) = \mathrm{e}^{\mathrm{Spect}(A)t} .$$

In infinite dimensions, relating the spectrum of the infinitesimal generator A to that of the semigroup $U(t)$ is a spectral mapping problem which is often nontrivial. The spectral inclusion

$$\mathrm{Spect}(U(t)) \subset \mathrm{e}^{\mathrm{Spect}(A)t}$$

always holds and the inverse inclusion is a problem that is solved by the spectral mapping theorems [12,18,22,65].

Often an application of center manifold theory needs to prove an appropriate version of the reduction theorem [7,11,12,14,15,18]. Pages 1–5 of the book [40] give an extensive list of the applications of center manifold theory to infinite dimensional problems. Here we consider a few of typical applications. Reaction-diffusion equations are typical models in chemical reaction (Belousov–Zhabotinsky reaction), biological systems, population dynamics and nuclear reaction physics. They have the form

$$u'_t = (K(\lambda) + D\Delta)u + f(u, \lambda) , \tag{65}$$

where K is a matrix depending on a parameter, D is a symmetric, positive semi-definite, often diagonal matrix, Δ is the Laplace operator, λ is a control parameter. The papers [2,27,31,67,78] applied center manifold theory to the study of the reaction-diffusion equation. Invariant manifolds for nonlinear Schrodinger equations are studied in [22,40]. In the elliptic quasilinear equations the spectrum of the linear operator is unbounded in unstable direction and in this case the solution does not generate even a semiflow. Nevertheless, center manifold technique is successfully applied in this case as well [45]. Henry [32] proved the persistence of normally hyperbolic closed linear subspace for semilinear parabolic equations. It follows the existence of the center manifold which is a perturbed manifold for the center subspace. I. Chueshov [14,15] considered a reduction principle for coupled nonlinear parabolic-hyperbolic partial differential equations that has applications in thermoelastisity. Mielke [47] has developed the center manifold theory for elliptic partial differential equations and has applied it to problems of elasticity and hydrodynamics. The results for non-autonomous system in Banach space can be found in [46]. The reduction principle for stochastic partial differential equations is considered in [9,17,75].

Future Directions

Now we should realize that principal problems in the center manifolds theory of finite dimensional dynamics have been solved and we may expect new results in applications. However the analogous theory for infinite dimensional dynamics is far from its completion. Here we have few principal problems such as the reduction principle (an analogue of Theorem 2 or 5) and the construction of the center manifold. The application of the center manifold theory to partial differential equations is one of promising and developing direction. In particular, it is very important to investigate the behavior of center manifolds when Galerkin method is applied due to a transition from finite dimension to infinite one. As it was mentioned above each application of center manifold methods to nonlinear infinite dimensional dynamics is nontrivial and gives rise to new directions of researches.

Bibliography

Primary Literature

1. Arnold VI (1973) Ordinary Differential Equations. MIT, Cambridge
2. Auchmuty J, Nicolis G (1976) Bifurcation analysis of reaction-diffusion equation (III). Chemical Oscilations. Bull Math Biol 38:325–350
3. Aulbach B (1982) A reduction principle for nonautonomous differential equations. Arch Math 39:217–232
4. Aulbach B, Garay B (1994) Partial linearization for noninvertible mappings. J Appl Math Phys (ZAMP) 45:505–542
5. Aulbach B, Colonius F (eds) (1996) Six Lectures on Dynamical Systems. World Scientific, New York
6. Balakrishnan AV (1976) Applied Functional Analysis. Springer, New York, Heidelberg
7. Bates P, Jones C (1989) Invariant manifolds for semilinear partial differential equations. In: Dynamics Reported 2. Wiley, Chichester, pp 1–38
8. Bogoliubov NN, Mitropolsky YUA (1963) The method of integral manifolds in non-linear mechanics. In: Contributions Differential Equations 2. Wiley, New York, pp 123–196
9. Caraballo T, Chueshov I, Landa J (2005) Existence of invariant manifolds for coupled parabolic and hyperbolic stochastic partial differential equations. Nonlinearity 18:747–767
10. Carr J (1981) Applications of Center Manifold Theory. In: Applied Mathematical Sciences, vol 35. Springer, New York

11. Carr J, Muncaster RG (1983) Applications of center manifold theory to amplitude expansions. J Diff Equ 59:260–288
12. Chicone C, Latushkin Yu (1999) Evolution Semigroups in Dynamical Systems and Differential Equations. Math Surv Monogr 70. Amer Math Soc, Providene
13. Chow SN, Lu K (1995) Invariant manifolds and foliations for quasiperiodic systems. J Diff Equ 117:1–27
14. Chueshov I (2007) Invariant manifolds and nonlinear master-slave synchronization in coupled systems. In: Applicable Analysis, vol 86, 3rd edn. Taylor and Francis, London, pp 269–286
15. Chueshov I (2004) A reduction principle for coupled nonlinesr parabolic-hyperbolic PDE. J Evol Equ 4:591–612
16. Diliberto S (1960) Perturbation theorem for periodic surfaces I, II. Rend Cir Mat Palermo 9:256–299; 10:111–161
17. Du A, Duan J (2006) Invariant manifold reduction for stochastic dynamical systems. http://arXiv:math.DS/0607366
18. Engel K, Nagel R (2000) One-parameter Semigroups for Linear Evolution Equations. Springer, New York
19. Fenichel N (1971) Persistence and smoothness of invariant manifolds for flows. Ind Univ Math 21:193–226
20. Fenichel N (1974) Asymptotic stability with rate conditions. Ind Univ Math 23:1109–1137
21. Gallay T (1993) A center-stable manifold theorem for differential equations in Banach space. Commun Math Phys 152:249–268
22. Gesztesy F, Jones C, Latushkin YU, Stanislavova M (2000) A spectral mapping theorem and invariant manifolds for nonlinear Schrodinger equations. Ind Univ Math 49(1):221–243
23. Grobman D (1959) Homeomorphism of system of differential equations. Dokl Akad Nauk SSSR 128:880 (in Russian)
24. Grobman D (1962) The topological classification of the vicinity of a singular point in *n*-dimensional space. Math USSR Sbornik 56:77–94; in Russian
25. Guckenheimer J, Holmes P (1993) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vectors Fields. Springer, New York
26. Hadamard J (1901) Sur l'etaration et les solution asymptotiques des equations differentielles. Bull Soc Math France 29:224–228
27. Haken H (2004) Synergetics: Introduction and Advanced topics. Springer, Berlin
28. Hale J (1961) Integral manifolds of perturbed differential systems. Ann Math 73(2):496–531
29. Hartman P (1960) On local homeomorphisms of Euclidean spaces. Bol Soc Mat Mex 5:220
30. Hartman P (1964) Ordinary Differential Equations. Wiley, New York
31. Hassard BD, Kazarinoff ND, Wan YH (1981) Theory and Applications of Hopf Bifurcation. Cambridge University Press, Cambridge
32. Henry D (1981) Geometric theory of semilinear parabolic equations. Lect Notes Math 840:348
33. Hirsch M, Smale S (1974) Differential Equations, Dynamical Systems and Linear Algebra. Academic Press, Orlando
34. Hirsch M, Pugh C, Shub M (1977) Invariant manifolds. Lect Notes Math 583:149
35. Hopf E (1942) Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems. Ber Verh Sachs Akad Wiss Leipzig Math-Nat 94:3–22
36. Iooss G (1979) Bifurcation of Maps and Application. N-Holl Math Stud 36:105
37. Jolly MS, Rosa R (2005) Computation of non-smooth local centre manifolds. IMA J Numer Anal 25(4):698–725
38. Kelley A (1967) The stable, center stable, center, center unstable and unstable manifolds. J Diff Equ 3:546–570
39. Kirchgraber U, Palmer KJ (1990) Geometry in the Neighborhood of Invariant Manifolds of the Maps and Flows and Linearization. In: Pitman Research Notes in Math, vol 233. Wiley, New York
40. Li C, Wiggins S (1997) Invariant Manifolds and Fibrations for Perturbed Nonlinear Schrodinger Equations. Springer, New York
41. Lyapunov AM (1892) Problémé Générale de la Stabilité du Mouvement, original was published in Russia 1892, transtalted by Princeton Univ. Press, Princeton, 1947
42. Ma T, Wang S (2005) Dynamic bifurcation of nonlinear evolution equations and applications. Chin Ann Math 26(2): 185–206
43. Marsden J, McCracken M (1976) Hopf bifurcation and Its Applications. Appl Math Sci 19:410
44. Mañé R (1978) Persistent manifolds are normally hyperbolic. Trans Amer Math Soc 246:261–284
45. Mielke A (1988) Reduction of quasilinear elliptic equations in cylindrical domains with application. Math Meth App Sci 10:51–66
46. Mielke A (1991) Locally invariant manifolds for quasilinear parabolic equations. Rocky Mt Math 21:707–714
47. Mielke A (1996) Dynamics of nonlinear waves in dissipative systems: reduction, bifurcation and stability. In: Pitman Research Notes in Mathematics Series, vol 352. Longman, Harlow, pp 277
48. Mitropolskii YU, Lykova O (1973) Integral Manifolds in Nonlinear Mechanics. Nauka, Moscow
49. Mizohata S (1973) The Theory of Partial Differential Equations. Cambridge University Press, Cambridge
50. Neimark Y (1967) Integral manifolds of differential equations. Izv Vuzov, Radiophys 10:321–334 (in Russian)
51. Osipenko G, Ershov E (1993) The necessary conditions of the preservation of an invariant manifold of an autonomous system near an equilibrium point. J Appl Math Phys (ZAMP) 44:451–468
52. Osipenko G (1996) Indestructibility of invariant non-unique manifolds. Discret Contin Dyn Syst 2(2):203–219
53. Osipenko G (1997) Linearization near a locally non-unique invariant manifold. Discret Contin Dyn Syst 3(2):189–205
54. Palis J, Takens F (1977) Topological equivalence of normally hyperbolic dynamical systems. Topology 16(4):336–346
55. Palmer K (1975) Linearization near an integral manifold. Math Anal Appl 51:243–255
56. Palmer K (1987) On the stability of center manifold. J Appl Math Phys (ZAMP) 38:273–278
57. Perron O (1928) Über Stabilität und asymptotisches Verhalten der Integrale von Differentialgleichungssystem. Math Z 29:129–160
58. Pillet CA, Wayne CE (1997) Invariant manifolds for a class of dispersive. Hamiltonian, partial differential equations. J Diff Equ 141:310–326
59. Pliss VA (1964) The reduction principle in the theory of stability of motion. Izv Acad Nauk SSSR Ser Mat 28:1297–1324; translated (1964) In: Soviet Math 5:247–250
60. Pliss VA (1966) On the theory of invariant surfaces. In: Differential Equations, vol 2. Nauka, Moscow pp 1139–1150

61. Poincaré H (1885) Sur les courbes definies par une equation differentielle. J Math Pure Appl 4(1):167–244
62. Pugh C, Shub M (1970) Linearization of normally hyperbolic diffeomorphisms and flows. Invent Math 10:187–198
63. Reinfelds A (1974) A reduction theorem. J Diff Equ 10:645–649
64. Reinfelds A (1994) The reduction principle for discrete dynamical and semidynamical systems in metric spaces. J Appl Math Phys (ZAMP) 45:933–955
65. Renardy M (1994) On the linear stability of hyperbolic PDEs and viscoelastic flows. J Appl Math Phys (ZAMP) 45:854–865
66. Sacker RJ (1967) Hale J, LaSalle J (eds) A perturbation theorem for invariant Riemannian manifolds. Proc Symp Diff Equ Dyn Syst Univ Puerto Rico. Academic Press, New York, pp 43–54
67. Sandstede B, Scheel A, Wulff C (1999) Bifurcations and dynamics of spiral waves. J Nonlinear Sci 9(4):439–478
68. Shoshitaishvili AN (1972) Bifurcations of topological type at singular points of parameterized vector fields. Func Anal Appl 6:169–170
69. Shoshitaishvili AN (1975) Bifurcations of topological type of a vector field near a singular point. Trudy Petrovsky seminar, vol 1. Moscow University Press, Moscow, pp 279–309
70. Sijbrand J (1985) Properties of center manifolds. Trans Amer Math Soc 289:431–469
71. Van Strien SJ (1979) Center manifolds are not C^{∞}. Math Z 166:143–145
72. Vanderbauwhede A (1989) Center Manifolds, Normal Forms and Elementary Bifurcations. In: Dynamics Reported, vol 2. Springer, Berlin, pp 89–169
73. Vanderbauwhede A, Iooss G (1992) Center manifold theory in infinite dimensions. In: Dynamics Reported, vol 1. Springer, Berlin, pp 125–163
74. Wan YH (1977) On the uniqueness of invariant manifolds. J Diff Equ 24:268–273
75. Wang W, Duan J (2006) Invariant manifold reduction and bifurcation for stochastic partial differential equations. http://arXiv:math.DS/0607050
76. Wiggins S (1992) Introduction to Applied Nonlinear Dynamical Systems and Chaos. Springer, New York
77. Wiggins S (1994) Normally Hyperbolic Invariant Manifolds of Dynamical Systems. Springer, New York
78. Wulff C (2000) Translation from relative equilibria to relative periodic orbits. Doc Mat 5:227–274

Books and Reviews

Bates PW, Lu K, Zeng C (1998) Existence and persistence of invariant manifolds for semiflows in Banach spaces. Mem Amer Math Soc 135:129

Bates PW, Lu K, Zeng C (1999) Persistence of overflowing manifolds for semiflow. Comm Pure Appl Math 52(8):983–1046

Bates PW, Lu K, Zeng C (2000) Invariant filiations near normally hyperbolic invariant manifolds for semiflows. Trans Amer Math Soc 352:4641–4676

Babin AV, Vishik MI (1989) Attractors for Evolution Equations. Nauka. Moscow; English translation (1992). Elsevier Science, Amsterdam

Bylov VF, Vinograd RE, Grobman DM, Nemyskiy VV (1966) The Theory of Lyapunov Exponents. Nauka, Moscow (in Russian)

Chen X-Y, Hale J, Tan B (1997) Invariant foliations for C^1 semigroups in Banach spaces. J Diff Equ 139:283–318

Chepyzhov VV, Goritsky AYU, Vishik MI (2005) Integral manifolds and attractors with exponential rate for nonautonomous hyperbolic equations with dissipation. Russ J Math Phys 12(1): 17–39

Chicone C, Latushkin YU (1997) Center manifolds for infinite dimensional nonautonomous differential equations. J Diff Equ 141:356–399

Chow SN, Lu K (1988) Invariant manifolds for flows in Banach spaces. J Diff Equ 74:285–317

Chow SN, Lin XB, Lu K (1991) Smooth invariant foliations in infinite dimensional spaces. J Diff Equ 94:266–291

Chueshov I (1993) Global attractors for non-linear problems of mathematical physics. Uspekhi Mat Nauk 48(3):135–162; English translation in: Russ Math Surv 48:3

Chueshov I (1999) Introduction to the Theory of Infinite-Dimensional Dissipative Systems. Acta, Kharkov (in Russian); English translation (2002) http://www.emis.de/monographs/Chueshov/. Acta, Kharkov

Constantin P, Foias C, Nicolaenko B, Temam R (1989) Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations. Appl Math Sci, vol 70. Springer, New York

Gonçalves JB (1993) Invariant manifolds of a differentiable vector field. Port Math 50(4):497–505

Goritskii AYU, Chepyzhov VV (2005) Dichotomy property of solutions of quasilinear equations in problems on inertial manifolds. SB Math 196(4):485–511

Hassard B, Wan Y (1978) Bifurcation formulae derived from center manifold theory. J Math Anal Appl 63:297–312

Hsia C, Ma T, Wang S (2006) Attractor bifurcation of three-dimensional double-diffusive convection. http://arXiv:nlin.PS/0611024

Knobloch HW (1990) Construction of center manifolds. J Appl Math Phys (ZAMP) 70(7):215–233

Latushkin Y, Li Y, Stanislavova M (2004) The spectrum of a linearized 2D Euler operator. Stud Appl Math 112:259

Leen TK (1993) A coordinate independent center manifold reduction. Phys Lett A 174:89–93

Li Y (2005) Invariant manifolds and their zero-viscosity limits for Navier–Stokes equations. http://arXiv:math.AP/0505390

Osipenko G (1989) Examples of perturbations of invariant manifolds. Diff Equ 25:675–681

Osipenko G (1985, 1987, 1988) Perturbation of invariant manifolds I, II, III, IV. Diff Equ 21:406–412, 21:908–914, 23:556–561, 24:647–652

Podvigina OM (2006) The center manifold theorem for center eigenvalues with non-zero real parts. http://arXiv:physics/0601074

Sacker RJ, Sell GR (1974, 1976, 1978) Existence of dichotomies and invariant splitting for linear differential systems. J Diff Equ 15:429-458, 22:478–522, 27:106–137

Scarpellini B (1991) Center manifolds of infinite dimensional. Main results and applications. J Appl Math Phys (ZAMP) 43:1–32

Sell GR (1983) Vector fields on the vicinity of a compact invariant manifold. Lect Notes Math 1017:568–574

Swanson R (1983) The spectral characterization of normal hyperbolicity. Proc Am Math Soc 89(3):503–508

Temam R (1988) Infinite Dimensional Dynamical Systems in Mechanics and Physics. Springer, Berlin

Zhenquan Li, Roberts AJ (2000) A flexible error estimate for the application of center manifold theory. http://arXiv.org/abs/math.DS/0002138

Zhu H, Campbell SA, Wolkowicz (2002) Bifurcation analysis of a predator-prey system with nonmonotonic functional response. SIAM J Appl Math 63:636–682

Chaos and Complexity in Astrophysics, Introduction to

Steven N. Shore[1,2]
[1] Dipartmento di Fisica Enrico Fermi, Università di Pisa, Pisa, Italy
[2] Istituto Nazionale di Fiscia Nucleare, Sezione di Pisa, Pisa, Italy

We live in a hierarchically structured, complex Universe and the articles in this section demonstrate the breadth of vision required by astrophysicists addressing a vast spectrum of problems from dynamical systems to biological and galactic evolution to cosmology. Fundamentally, astronomical problems present a huge variety of scales of time, length, mass, and energy. For example, to understand the transport of energy in hot, opaque bodies without which we cannot understand the structure of stars requires kinetic theory of gases and atomic and molecular processes (the equation of state for stellar matter, and opacities), hydrodynamics, (especially turbulence for convective energy transport and circulation), magnetohydrodynamics and plasma physics (since stars generate magnetic fields by dynamo processes), nuclear and particle physics (for thermonuclear processing and energy loss), and even general relativity (for the late stages of evolution). But stars are formed and die within galaxies, self-gravitating ensembles of many billions of individual stars and cold clouds of millions of solar masses that are actually analogies of ecosystems. Thus the tools developed to address biological systems can be carried over into cosmic problems.

At the close of the 1920s, Arthur Eddington quipped that "someday, perhaps, we shall understand something as simple as a star". How wrong he was. Stars are self-gravitating nuclear engines and it is the mass of a gaseous body that determines whether it will or won't ignite its central thermonuclear energy source. Below a few percent of a solar mass, the stellar core never reaches the required temperature even if the densities are high enough to initiate nuclear reactions. Such a low mass object is called a brown dwarf if, during its initial contraction from an interstellar dark cloud, it can briefly burn deuterium. If even that source remains inaccessible, it is called a planet, regardless of whether it is bound in an orbit to another, more massive companion. Although various claims had been made throughout the 20th century for the presence of planetary systems, especially for the nearby very low mass object called Barnard's Star, the first direct detection of a planetary mass body orbiting another star came only in 1995 with the discovery of a massive Jupiter-like companion to the solar type star 51 Peg. Since then over 300 planets have been discovered orbiting stars other than the Sun, now including multiple systems and even several in other galaxies discovered by gravitational lensing. This sample is, however, currently still biased toward masses much greater than the Earth and systems that are still rather different from our solar system. But with the passage of years and the accumulation of velocity and photometric measurements, systems progressively more similar to ours have begun to emerge.

The first discoveries already, however, presented a paradox: It has generally been thought that the structure of our system is the product of thermal evaporation and restructuring the orbiting bodies by the early Sun, the evaporation of nearby gaseous masses and leaving behind only their rocky cores.

The detection of planetary mass bodies began, however, with the discovery of precisely such planets within a few hundred stellar radii of the central star. This has stimulated a renaissance in dynamical astronomy of few bodies and the re-discovery of effects previously thought to be of only minor importance in the present solar system. This too is reflected in the articles in this section. And because it is clear now that there must also be systems like our own, the study of extraterrestrial environments and the search for extrasolar Earths and the origin of life has emerged from the back rooms to take a central stage as a new sub-discipline: astrobiology (see ► Exobiology (theoretical), Complexity in, by Brandenburg and ► Exobiology and Complexity by Chaisson). It is particularly appropriate that this encyclopedia recognizes the immense promise of this field.

The interplanetary medium, the solar wind, is the topic of several articles. This is the supersonic mass outflow from the only star for which we have virtually all possible information. The wind, synonymous with the interplanetary medium, is the expanding coronal layer of the Sun. Heated by nonthemal mechanisms seated deep within he interior dynamo and convective envelope, magnetohydrodynamic waves and topological relaxation of the outer magnetic field (see ► Topological Magnetohydrodynamics and Astrophysics by Berger).

This is why several of the articles here deal with dynamo processes. This plasma is an ideal laboratory in which the Sun performs the experiments and we capture the results with in situ satellite measurements (see ► Self-

-Organization in Magnetohydrodynamic Turbulence by Veltri et al., and ► Space Plasmas, Dynamical Complexity in, by Chang).

Wave phenomena that are inaccessible under terrestrial laboratory conditions can be studied in detail with spacecraft, MHD turbulence can be observed on a vast range of length scales (from meters to astronomical units) and timescales (from milliseconds to days). We are in a new age of spacecraft measurement with SOHO and STEREO providing continuously monitored dynamical three dimensional imaging of the Sun.

And on the solar surface, in deep in its interior, we see the cyclic generation and decay of a dynamo whose superficial manifestations are the corona, magnetic loop structures of sizes comparable to the stellar radius, and dissipative processes that release high energy particles and locally heat the outer solar atmosphere. All of these topics show the complex interplay of many physical processes and are discussed in many of the articles in this section.

Cosmic rays were discovered at the start of the 20th century by Hess and others making balloon flights to high altitudes carrying devices for measuring atmospheric electricity. They discovered that the ionization increases with height, which led to the hypothesis of some form of extraterrestrial energetic particles impinging on the atmosphere. Further study showed these to be charged, high energy particles, protons, alpha particles, and electrons, whose source was a mystery. For some time, these were the only particles with which fundamental physics could be studied. Somehow, the universe is capable of producing relativistic ions and electrons that traverse the space between the galaxies, reaching energies well in excess of TeV. The Sun and solar type stars, during flares, are capable of accelerating electrons to GeV energies and beyond in milliseconds or shorter on scales of kilometers, but the process is still not understood. And on larger scales, the detection of TeV mission from supernova remnants, the supersonically ejected debris of stellar explosions, shows that these shocks can also accelerate particles on length scales of the size of the solar system throughout the life of the galaxy. Hyper-relativistic phenomena connected with stellar collapse and explosion, such as Gamma Ray Bursts, also demonstrate that complex acceleration processes happen on many different length and time scales in cosmic objects (see ► Acceleration Mechanisms by Melrose).

Historically, it was in problems arising within dynamical astronomy that chaos was first recognized – in the orbital dynamics of the solar system during the 19th century – and the richness of this field has extended far beyond the confines of the realm of the planets. It was the source of the analogies used by Boltzmann and Gibbs in the early stages of the development of statistical mechanics (see ► Astrophysics, Chaos and Complexity in, by Regev).

Beginning with the two body problem, for point masses the physical solution is almost trivial. The motion is conservative, the total energy is constant and, except for effect of general relativity for the most extreme systems, the problem can be solved in closed orbits knowing only the total energy and angular momentum. If one or both objects are extended, or if they are internally complicated, the problem ceases to be tractable in closed form. The orbits evolve in time, either periodically or secularly depending on whether or not there are internal dissipative processes in the constituent bodies. This feature, which complicates the dynamics of binary star systems and even planetary moons requires a sophisticated physical treatment of microphysical processes from gas laws to fluid mechanics.

Planetary ring systems are even more complex, showing a wealth of dynamical structuring that, from particle trajectories imaged in natural settings, produce images like those resulting from numerical models of nonlinear systems. At the next level, the apparently simple problem of the lunar orbit becomes almost impossible once tidal interactions and the effect of the Sun are included. This was the principal discovery of Poincare, Lyapunov, and the dynamicists of the late 19th century, that the perturbing influences on the motion of so small a body a the Moon, whose mass is negligible with respect to the two principals, has a complex temporal spectrum of resonances that produce, even without dissipation, extreme sensitivity of the predictions to the accuracy of the initial conditions for any dynamical calculation (see ► Orbital Dynamics, Chaos in, by Hadjidemetriou).

The extension of this complexity to N-body calculations (see ► Stellar Dynamics, *N*-body Methods for, by Makino) and the evolution of far more populous systems – clusters and galaxies (see ► Astrophysics: Dynamical Systems by Contopoulos) – are addressed in articles in this section.

Moving to the largest scale of all, the Universe as a whole, we reach the limits of current astronomical observations. In the last decade, gravitational wave observatories have opened a new type of astronomical observation, one connected with neither particles nor light and at the very limits of detection. During the formation of black holes and neutron stars, at the moment of collapse of tars either by intrinsic or extrinsic triggering, gravitational waves certainly carry away a significant portion of the binding energy of the initial system. Such emission is required by the theory of collapse and the detection of the secular evolution of binary pulsar orbits assures the physical correctness of the fundamentally relativistic phenomenon. However, the Universe is a very big place and in all galaxies, whatever their evolutionary history and distance, there must be sources for such waves.

These produce an incoherent signal, unlike the strong, identifiable signature of a single collapse event, that should form a stochastic background from the integration of all such sources. It is also clear that at some stage in the earliest stages of the cosmic expansion there were fluctuations in the matter and fields that were strongly coupled with the spacetime structure. These were not dissipatively damped as would happen in a normal fluid.

Instead, their spectrum was systematically altered by the cosmic expansion to the point of visibility in the present universe. How this might be found and what it tells us about the first moments of the Big Bang and the subsequent evolution of cosmic sources during the Hubble time is also treated in this section (see ▶ Cosmic Gravitational Background, Stochastic by Ungarelli and ▶ Cosmic Strings by Achúcarro and Martins).

Finally, because of this huge range of timescales and multiplicity of processes involved in even some of the apparently simplest phenomena, astronomical observations present particularly challenging problems of data analysis (see ▶ Astronomical Time Series, Complexity in, by Scargle).

From dynamical systems to stochastic fluctuations of spacetime, the problem of finding periods and spectra in complex, not spatially resolved and partially sampled data takes us to the limit of current statistical methodology.

In fact, this is a fundamental feature of all astrophysical research, to stretch all of our human capacities, intellectual and technological, in an ever-widening panorama.

Chaos and Ergodic Theory

JÉRÔME BUZZI
C.N.R.S. and Université Paris-Sud, Orsay, France

Article Outline

Glossary

For simplicity, definitions are given for a continuous or smooth self-map or diffeomorphism T of a compact manifold M.

Entropy, measure-theoretic (or: metric entropy) For an ergodic invariant probability measure μ, it is the smallest exponential growth rates of the number of orbit segments of given length, with respect to that length, after restriction to a set of positive measure. We denote it by $h(T, \mu)$. See ▶ Entropy in Ergodic Theory and Subsect. "Local Complexity" below.

Entropy, topological It is the exponential growth rates of the number of orbit segments of given length, with respect to that length. We denote it by $h_{\text{top}}(f)$. See ▶ Entropy in Ergodic Theory and Subsect. "Local Complexity" below.

Ergodicity A measure is ergodic with respect to a map T if given any measurable subset S which is invariant, i. e., such that $T^{-1}S = S$, either S or its complement has zero measure.

Hyperbolicity A measure is hyperbolic in the sense of Pesin if at almost every point no Lyapunov exponent is zero. See ▶ Smooth Ergodic Theory.

Kolmogorov typicality A property is *typical in the sense of Kolmogorov* for a topological space $\mathcal{F}$ of parametrized families $f = (f_t)_{t \in U}$, U being an open subset of $\mathbb{R}^d$ for some $d \geq 1$, if it holds for f_t for Lebesgue almost every t and topologically generic $f \in \mathcal{F}$.

Lyapunov exponents
The Lyapunov exponents (▶ Smooth Ergodic Theory) are the limits, when they exist, $\lim_{n\to\infty} \frac{1}{n} \log \|(T^n)'(x).v\|$ where $x \in M$ and v is a non zero tangent vector to M at x. The Lyapunov exponents of an ergodic measure is the set of Lyapunov exponents obtained at almost every point with respect to that measure for all non-zero tangent vectors.

Markov shift (topological, countable state) It is the set of all infinite or bi-infinite paths on some countable directed graph endowed with the left-shift, which just translates these sequences.

Maximum entropy measure It is a measure μ which maximizes the measured entropy and, by the variational principle, realized the topological entropy.

Physical measure It is a measure μ whose basin, $\{x \in M \colon \forall \phi \colon M \to \mathbb{R}$ continuous $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1}$

$\phi(f^k x) = \int \phi \, d\mu\}$ has nonzero volume.

Prevalence A property is *prevalent* in some complete metric, separable vector space X if it holds outside of a set N such that, for some Borel probability measure μ on X: $\mu(A+v) = 0$ for all $v \in X$. See [76,141,239].

Sensitivity on initial conditions T has *sensitivity to initial conditions* on $X' \subset X$ if there exists a constant $\rho > 0$ such that for every $x \in X'$, there exists $y \in X$, arbitrarily close to x, and $n \geq 0$ such that $d(T^n y, T^n x) > \rho$.

Sinai–Ruelle–Bowen measures It is an invariant probability measure which is absolutely continuous along the unstable foliation (defined using the unstable manifolds of almost every $x \in M$, which are the sets $W^u(x)$, of points y such that $\lim_{n\to\infty} \frac{1}{n} \log d(T^{-n} y, T^{-n} x) < 0$).

Statistical stability T is statistically stable if the physical measures of nearby deterministic systems are arbitrarily close to the convex hull of the physical measures of T.

Stochastic stability T is stochastically stable if the invariant measures of the Markov chains obtained from T by adding a suitable, smooth noise with size $\epsilon \to 0$ are arbitrarily close to the convex hull of the physical measures of T.

Structural stability T is structurally stable if any S close enough to T is topologically the same as T: there exists a homeomorphism $h\colon M \to M$ such that $h \circ T = S \circ h$ (orbits are sent to orbits).

Subshift of finite type It is a closed subset Σ_F of $\Sigma = \mathcal{A}^{\mathbb{Z}}$ or $\Sigma = \mathcal{A}^{\mathbb{N}}$ where $\mathcal{A}$ is a finite set satisfying: $\Sigma_F = \{x \in \Sigma\colon \forall k < \ell\colon x_k x_{k+1} \dots x_\ell \notin F\}$ for some finite set F.

Topological genericity Let X be a Baire space, e. g., a complete metric space. A property is *(topologically) generic* in a space X (or holds for the (topologically) generic element of X) if it holds on a nonmeager set (or set of second Baire category), i. e., on a dense G_δ subset.

Definition of the Subject

Chaotic dynamical systems are those which present unpredictable and/or complex behaviors. The existence and importance of such systems has been known at least since Hadamard [126] and Poincaré [208], however it became well-known only in the sixties. We refer to [36,128,226,236] and [80,107,120,125,192,213] for the relevance of such dynamics in other fields, mathematical or not (see also ▶ Ergodic Theory: Interactions with Combinatorics and Number Theory, ▶ Ergodic Theory: Fractal Geometry). The numerical simulations of chaotic dynamics can be difficult to interpret and to plan, even misleading, and the tools and ideas of mathematical dynamical system theory are indispensable. Arguably the most powerful set of such tools is ergodic theory which provides a statistical description of the dynamics by attaching relevant probability measures. In opposition to single orbits, the statistical properties of chaotic systems often have good stability properties. In many cases, this allows an understanding of the complexity of the dynamical system and even precise and quantitative statistical predictions of its behavior. In fact, chaotic behavior of single orbits often yields global stability properties.

Introduction

The word *chaos*, from the ancient Greek $\chi\alpha o\sigma$, "shapeless void" [131] and "raw confused mass" [199], has been used [*$\chi\alpha o\sigma$ also inspired Van Helmont to create the word "gas" in the seventeenth century and this other thread leads to the molecular chaos of Boltzmann in the nineteenth century and therefore to ergodic theory itself*.] since a celebrated paper of Li and Yorke [169] to describe evolutions which however deterministic and defined by rather simple rules, exhibit unpredictable or complex behavior.

Attempts at Definition

We note that, like many ideas [237], this is not captured by a single mathematical definition, despite several attempts (see, e. g., [39,112,158,225] for some discussions as well as the monographs on chaotic dynamics [15,16,58,60,78,104,121,127,215,250,261]). Let us give some of the most well-known definitions, which have been given mostly from the topological point of view, i. e., in the setting of a self-map $T\colon X \to X$ on a compact metric space whose distance is denoted by d:

T has *sensitivity to initial conditions* on $X' \subset X$ if there exists a constant $\rho > 0$ such that for every $x \in X'$, there exists $y \in X$, arbitrarily close to x with a finite separating time:

$$\exists n \geq 0 \quad \text{such that} \quad d(T^n y, T^n x) > \rho\,.$$

In other words, any uncertainty on the exact value of the initial condition x makes $T^n(x)$ completely unknown for n large enough. If X is a manifold, then *sensitivity to initial conditions in the sense of Guckenheimer* [120] means that the previous phenomenon occurs for a set X' with nonzero volume.

T is *chaotic in the sense of Devaney* [94] if it admits a dense orbit and if the periodic points are dense in X. It implies sensitivity to initial conditions on X.

T is *chaotic in the sense of Li and Yorke* [169] if there exists an uncountable subset $X' \subset X$ of points, such that, for all $x \neq y \in X'$,

$$\liminf_{n\to\infty} d(T^n x, T^n y) = 0 \quad \text{and}$$
$$\limsup_{n\to\infty} d(T^n x, T^n y) > 0\,.$$

T has *generic chaos* in the sense of Lasota [206] if the set

$$\{(x,y) \in X \times X\colon \liminf_{n\to\infty} d(T^n x, T^n y) = 0 < \limsup_{n\to\infty} d(T^n x, T^n y)\}$$

is topologically generic (*see glossary.*) in $X \times X$.

Topological chaos is also sometimes characterized by *nonzero topological entropy* (▶ Entropy in Ergodic Theory): there exist exponentially many orbit segments of a given length. This implies chaos in the sense of Li and Yorke by [39].

As we shall see ergodic theory describes a number of chaotic properties, many of them implying some or all of the above topological ones. The main such property for a smooth dynamical system, say a $C^{1+\alpha}$-diffeomorphism of a compact manifold, is the existence of an invariant probability measure which is:

1. Ergodic (cannot be split) and aperiodic (not carried by a periodic orbit);
2. Hyperbolic (nearby orbits converge or diverge at a definite exponential rate);
3. Sinai–Ruelle–Bowen (as smooth as it is possible).

(For precise definitions we refer to ▶ Smooth Ergodic Theory or to the discussions below.) In particular such a situation implies nonzero entropy and sensitivity to initial condition of a set of nonzero Lebesgue measure (i. e., positive volume).

Before starting our survey in earnest, we shall describe an elementary and classical example, the full tent map, on which the basic phenomena can be analyzed in a very elementary way. Then, in Sect. "Picking an Invariant Probability Measure", we shall give some motivations for introducing probability theory in the description of chaotic but deterministic systems, in particular the unpredictability of their individual orbits. We define two of the most relevant classes of invariant measures: the physical measures and those maximizing entropy. It is unknown in which generality these measures exist and can be analyzed but we describe in Sect. "Tractable Chaotic Dynamics" the major classes of dynamics for which this has been done. In Sect. "Statistical Properties" we describe some of the finer statistical properties that have been obtained for such good chaotic systems: sums of observables along orbits are statistically undistinguishable from sums of independent and identically distributed random variables. Sect. "Orbit Complexity" is devoted to the other side of chaos: the complexity of these dynamics and how, again, this complexity can be analyzed, and sometimes classified, using ergodic theory. Sect "Stability" describes perhaps the most striking aspect of chaotic dynamics: the unstability of individual orbit is linked to various forms of stability of the global dynamics.

Finally we conclude by mentioning some of the most important topics that we could not address and we list some possible future directions.

Caveat. The subject-matter of this article is somewhat fuzzy and we have taken advantage of this to steer our path towards some of our favorite theorems and to avoid the parts we know less (some of which are listed below). We make no pretense at exhaustivity neither in the topics nor in the selected results and we hope that our colleagues will excuse our shortcomings.

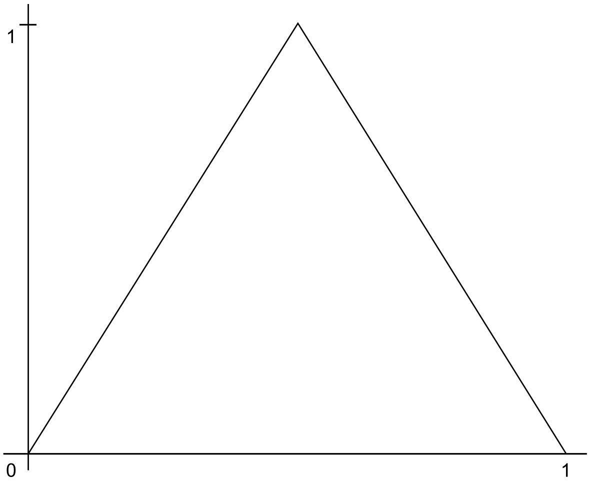
Remark 1 In this article we only consider *compact, smooth and finite-dimensional* dynamical systems *in discrete time*, i. e., defined by self-maps. In particular, we have omitted the natural and important variants applying to flows, e. g., evolutions defined by ordinary differential equations but we refer to the textbooks (see, e. g.,[15,128,148]) for these.

Elementary Chaos: A Simple Example

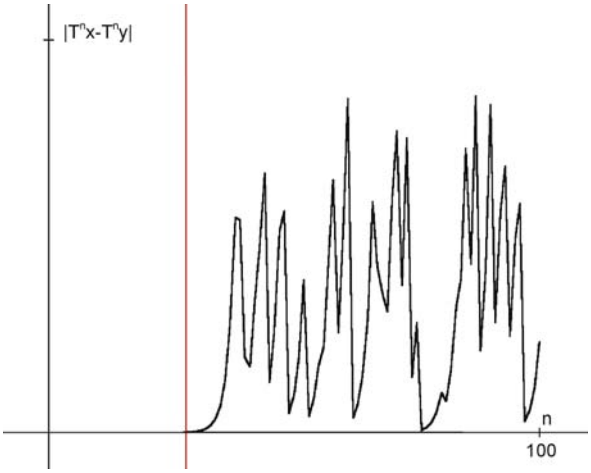
We start with a toy model: the full tent map T of Fig. 1. Observe that for any point $x \in [0,1]$, $T^{-n}(x) = \{(\sigma(k,n)x + k)\cdot 2^{-n}\colon k = 0,1,\dots,2^n - 1\}$, where $\sigma(k,n) = \pm 1$. Hence $\bigcup_{n\geq 0} T^{-n}(x)$ is dense in $[0,1]$. It easily follows that T exhibits *sensitive dependence to initial conditions*. Even worse in this example, the qualitative asymptotic behavior can be completely changed by this arbitrarily small perturbation: x may have a dense orbit whereas y is eventually mapped to a fixed point! This is Devaney chaos [94].

This kind of unstability was first discovered by J. Hadamard [126] in his study of the geodesic flow (i. e., the frictionless movement of a point mass constrained to remain on a surface). At that time, such an unpredictability was considered a purely mathematical pathology, necessarily devoid of any physical meaning [*Duhem qualified Hadamard's result as "an example of a mathematical deduction which can never be used by physics" (see pp. 206–211 in* [103]*)!*].

Returning to out tent map, we can be more quantitative. At any point $x \in [0,1]$ whose orbit never visits 1/2, the *Lyapunov exponent* $\lim_{n\to\infty} \frac{1}{n} \log |(T^n)'(x)|$



Chaos and Ergodic Theory, Figure 1
The graph of the full tent map $f(x) = 1 - |1 - 2x|$ over $[0, 1]$



Chaos and Ergodic Theory, Figure 2
$|T^n(x) - T^n(y)|$ for $T(x) = 1 - |1 - 2x|$, $|x - y| = 10^{-12}$ and $0 \le n \le 100$. The vertical red line is at $n = 28$ and shows when $|T^n x - T^n y| \ge 0.5 \cdot 10^{-4}$ for the first time

is $\log 2$. (*See the glossary*.) Such a positive Lyapunov exponent corresponds to infinitesimally close orbits getting separated exponentially fast. This can be observed in Fig. 2. Note how this exponential speed creates a rather sharp transition.

It follows in particular that experimental or numerical errors can grow very quickly to size 1, [*For simple precision arithmetic the uncertainty is* 10^{-16} *which grows to size 1 in 38 iterations of* T.] i. e., the approximate orbit may contain after a while *no information about the true orbit*. This casts a doubt on the reliability of simulations. Indeed, a simulation of T on most computers will suggest that all orbits quickly converge to 0, which is completely false [*Such a collapse to 0 does really occurs but only for a countable subset of initial conditions in* $[0, 1]$ *whereas the points with dense orbit make a subset of* $[0, 1]$ *with full Lebesgue measure (see below). This artefact comes from the way numbers are represented – and approximated – on the computer: multiplication by even integers tends to "simplify" binary representations. Thus the computations involved in drawing Fig.* 2 *cannot be performed too naively*.]. Though somewhat atypical in its dramatic character, this failure illustrates the *unpredictability* and *unstability* of individual orbits in chaotic systems.

Does this mean that all quantitative predictions about orbits of T are to be forfeited? Not at all, *if we are ready to change our point of view* and look beyond a single orbit. This can be seen easily in this case. Let us start with such a global analysis from the topological point of view. Associate to $x \in [0, 1]$, a sequence $i(x) = i = i_0 i_1 i_2 \dots$ of 0s and 1s according to $i_k = 0$ if $T^k x \le 1/2$, $i_k = 1$ otherwise. One can check that [*Up to a countable set of exceptions*.] $\{i(x) : x \in [0, 1]\}$ is the set $\Sigma_2 := \{0, 1\}^{\mathbb{N}}$ of *all* infinite sequences of 0s and 1s and that at most one $x \in [0, 1]$ can realize a given sequence as $i(x)$.

Notice how the transformation f becomes trivial in this representation:

$$i(f(x)) = i_1 i_2 i_3 \dots \quad \text{if } i(x) = i_0 i_1 i_2 i_3 \dots$$

Thus f is represented by the simple and universal "left-shift" on sequences, which is denoted by σ. This representation of a rather general dynamical system by the left-shift on a space of sequences is called *symbolic dynamics* (▶ Symbolic Dynamics), [171].

This can be a very powerful tool. Observe for instance how here it makes obvious that we have complete *combinatorial freedom* over the orbits of T: one can easily build orbits with various asymptotic behaviors: if a sequence of Σ_2 contains all the finite sequences of 0s and 1s, then the corresponding point has a dense orbit; if the sequence is periodic, then the corresponding point is itself periodic, to give two examples of the *richness* of the dynamics.

More quantitatively, the number of distinct subsequences of length n appearing in sequences $i(x)$, $x \in [0, 1]$, is 2^n. It follows that the *topological entropy* (▶ Entropy in Ergodic Theory) of T is $h_{\text{top}}(T) = \log 2$. [*For the coincidence of the entropy and the Lyapunov exponent see below*.] The positivity of the topological entropy can be considered as the signature of the complexity of the dynamics and considered as the definition, or at least the stamp, of a topologically chaotic dynamics.

Let us move on to a probabilistic point of view. Pick $x \in [0, 1]$ randomly according to, say, the uniform law in $[0, 1]$. It is then routine to check that $i(x)$ follows the (1/2, 1/2)-Bernoulli law: the probability that, for any given k, $i_k(x) = 0$ is 1/2 and the i_ks are independent. Thus $i(x)$, seen as a sequence of random 0 and 1 when x is subject to the uniform law on $[0, 1]$, is *statistically undistinguishable from coin tossing!* This important remark leads to quantitative predictions. For instance, the *strong law of large numbers* implies that, for Lebesgue-almost every $x \in [0, 1]$ (i. e., for all $x \in [0, 1]$ except in a set of Lebesgue measure zero), the fraction of the time spent in any dyadic interval $I = [k \cdot 2^{-N}, \ell \cdot 2^{-N}] \subset [0, 1]$, $k, \ell, N \in \mathbb{N}$, by the orbit of x,

$$\lim_{n\to\infty} \frac{1}{n} \#\{0 \le k < n \colon T^k x \in I\} \tag{1}$$

exists and is equal to the length, 2^{-N}, of that interval. [*Eq.* (1) *in fact holds for any interval I. This implies that the orbit of almost every* $x \in [0, 1]$ *visits all subintervals of* $[0, 1]$, *i. e., the orbit is dense: in complete contradiction with the above mentioned numerical simulation!*] More generally, we shall see that, if $\phi \colon [0, 1] \to \mathbb{R}$ is any continuous function, then, for Lebesgue almost every-x,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(T^k x) \quad \text{exists and is equal to} \quad \int \phi(x)\, \mathrm{d}x\,. \tag{2}$$

Using strong *mixing properties* (▶ Ergodicity and Mixing Properties) of the Lebesgue measure under T, one can prove further properties, e. g., sensitivity on initial conditions in the sense of Guckenheimer [*The Lebesgue measure is* weak-mixing: *Lebesgue-almost all couples of points* $(x, y) \in [0, 1]^2$ *get separated. Note that it is not true of every couple off the diagonal: Counter-examples can be found among couples* (x, y) *with* $T^n x = T^n y$ *arbitrarily close to 1.*] and study the fluctuations of the averages $\frac{1}{n} \sum_{k=0}^{n-1} \phi(Tx)$ by the way of limit theorems.

The above analysis relied on the very special structure of T but, as we shall explain, the ergodic theory of differentiable dynamical systems shows that all of the above (and much more) holds in some form for rather general classes of chaotic systems. The different chaotic properties are independent in general (e. g., one may have topological chaos whereas the asymptotic behavior of almost all orbits is periodic) and the proofs can become much more difficult. We shall nonetheless be rewarded for our efforts by the discovery of unexpected links between chaos and stability, complexity and simplicity as we shall see.

Picking an Invariant Probability Measure

One could think that dynamical systems, such as those defined by self-maps of manifolds, being *completely deterministic*, have nothing to do with probability theory. There are in fact several motivations for introducing various invariant probability measures.

Statistical Descriptions

An abstract goal might be to enrich the structure: a smooth self-map is a particular case of a Borel self-map, hence one can canonically attach to this map its set of all invariant Borel probability measures [*From now on all measures will be Borel probability measures except if it is explicitly stated otherwise.*], or just the set of *ergodic* [*A measure is* ergodic *if all measurable invariant subsets have measure 0 or 1. Note that arbitrary invariant measures are averages of ergodic ones, so many questions about invariant measures can be reduced to ergodic ones.*] ones. By the Krylov–Bogoliubov theorem (see, e. g., [148]), this set is non-empty for any continuous self-map of a compact space.

By the following fundamental theorem ▶ Ergodic Theorems, each such measure is the statistical description of some orbit:

Birkhoff Pointwise Ergodic Theorem *Let* $(X, \mathcal{F}, \mu)$ *be a space with a* σ*-field and a probability measure. Let* $f \colon X \to X$ *be a measure-preserving map, i. e.,* $f^{-1}(\mathcal{F}) \subset \mathcal{F}$ *and* $\mu \circ f^{-1} = \mu$. *Assume ergodicity of* μ *(See the glossary.) and (absolute) integrability of* $\phi \colon X \to \mathbb{R}$ *with respect to* μ. *Then for* μ*-almost every* $x \in X$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(f^k x) \quad \textit{exists and is} \quad \int \phi \, \mathrm{d}\mu\,.$$

This theorem can be interpreted as saying that "time averages" coincide almost surely with "ensemble averages" (or "phase space average"), i. e., that Boltzmann's Ergodic Hypothesis of statistical mechanics [110] holds for dynamical systems that cannot be split in a measurable and non trivial way. [*This indecomposability is however often difficult to establish. For instance, for the hard ball model of a gas it is known only under some generic assumption (see* [238] *and the references therein).*] We refer to [161] for background.

Remark 2 One should observe that the existence of the above limit is not at all obvious. In fact it often fails from other points of view. One can show that for the full tent map $T(x) = 1 - |1 - 2x|$ analyzed above and many functions ϕ, the set of points for which it fails is large both *from the topological point of view* (it contains a dense G_δ set) and *from the dimension point of view* (it has Hausdorff dimension 1 [28]). This is an important point: the introduction of invariant measures allows one to avoid some of the wilder pathologies.

To illustrate this let us consider the full tent map $T(x) = 1 - |1 - 2x|$ again and the two ergodic invariant measures: δ_0 (the Dirac measure concentrated at the fixed point 0) and the Lebesgue measure $\mathrm{d}x$. In the first case, we obtain a complex proof of the obvious fact that the time average at $x = 0$ (some set of full measure!) and the ensemble average with respect to δ_0 are both equal to $f(0)$. In the second case, we obtain a very general proof of the above Eq. (2).

Another type of example is provided by the contracting map $S\colon [0,1] \to [0,1]$, $S(x) = x/2$. S has a unique invariant probability measure, δ_0. For Birkhoff theorem the situation is the same as that of T and δ_0: it asserts only that the orbit of 0 is described by δ_0.

One can understand Birkhoff theorem as a (first and rather weak) *stability result*: the time averages are independent of the initial condition, *almost surely with respect to* μ.

Physical Measures

In the above silly example S, much more is true than the conclusion of Birkhoff Theorem: *all* points of $[0,1]$ are described by δ_0. This leads to the definition of the *basin* of a probability measure μ for a self-map f of a space M:

$$\mathcal{B}(\mu) := \Bigg\{ x \in M\colon \forall \phi\colon \\ M \to \mathbb{R} \text{ continuous } \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(f^k x) = \int \phi \, \mathrm{d}\mu \Bigg\} .$$

If M is a manifold, then there is a notion of volume and one can make the following definition. A *physical measure* is a probability measure whose basin has nonzero volume in M. Say that a dynamical system $f\colon M \to M$ on a manifold has a *finite statistical description* if there exists finitely many invariant probability measures $\mu_1, \ldots, \mu_n$ the union of whose basins is the whole of M, up to a set of zero Lebesgue measure.

Physical measures are among the main subject of interest as they are expected to be exactly those that are "experimentally visible". Indeed, if $x_0 \in \mathcal{B}(\mu)$ and $\epsilon_0 > 0$ is small enough, then, by Lebesgue density theorem, a point x picked according to, say, the uniform law in the ball $B(x_0, \epsilon_0)$ of center x_0 and radius ϵ_0, will be in $\mathcal{B}(\mu)$ with probability almost 1 and therefore its ergodic averages will be described by μ. Hence "experiments" can be expected to follow the physical measures and this is what is numerically observed in most of the situations (see however the caveat in the discussion of the full tent map).

The existence of a finite statistical description (or even of a physical measure) is, as we shall see, not automatic nor routine to prove. Attracting periodic points as in the above silly example provide a first type of physical measures. Birkhoff ergodic theorem asserts that absolutely continuous ergodic invariant measures, usually obtained from some expansion property, give another class of physical measures. These contracting and expanding types can be combined in the class of *Sinai–Ruelle–Bowen measures* [166] which are the invariant measures absolutely continuous "along expanding directions" (see for the precise but technical definition ▶ Smooth Ergodic Theory). Any ergodic Sinai–Ruelle–Bowen measure which is ergodic and without zero Lyapunov exponent [*That is, the set of points* $x \in M$ *such that* $\lim_{n\to\infty} \frac{1}{n} \log \|(f^n)'(x).v\| = 0$ *for some* $v \in T_x M$ *has zero measure.*] is a physical measure. Conversely, "most" physical measures [*For counter-examples see* [136].] are of this type [243,247].

Measures of Maximum Entropy

For all parameters $t \in [3.96, 4]$, the quadratic maps $Q_t(x) = tx(1-x)$, $Q_t\colon [0,1] \to [0,1]$, have nonzero topological entropy [91] and exponentially many periodic points [134]:

$$\lim_{n\to\infty} \frac{\#\{x \in [0,1]\colon Q_t^n(x) = x\}}{e^{n h_{\mathrm{top}}(Q_t)}} = 1 .$$

On the other hand, by a deep theorem [117,178] there is an open and dense subset of $t \in [0,4]$, such that Q_t has a unique physical measure *concentrated on a periodic orbit*! Thus the physical measures can completely miss the topological complexity (and in particular the distribution of the periodic points). Hence one must look at other measures to get a statistical description of the complexity of such Q_t. Such a description is often provided by *measures of maximum entropy* μ_M whose measured entropy [*The usual phrases are "measure-theoretic entropy", "metric entropy".*] (▶ Entropy in Ergodic Theory) satisfies:

$$h(f, \mu_M) = \sup_{\mu \in M(f)} h(f, \mu) =^1 h_{\mathrm{top}}(f) .$$

$M(f)$ is the set of all invariant measures. [*One can restrict this to the ergodic invariant measures without changing the value of the supremum* (▶ Entropy in Ergodic Theory).] Equality 1 above is the *variational principle*: it holds for all continuous self-maps of compact metric spaces. One can say that the ergodic complexity (the complexity of f as seen by its *invariant measures*) captures the full topological complexity (defined by counting *all* orbits).

Remark 3 The variational principle implies the existence of "complicated invariant measures" as soon as the topological entropy is nonzero (see [47] for a setting in which this is of interest).

Maximum entropy measures do not always exist. However, if f is C^∞ smooth, then maximum entropy measures exist by a theorem of Newhouse [195] and they indeed describe the topological complexity in the following sense. Consider the probability measures:

$$\mu_{n,\epsilon} := \frac{1}{n}\sum_{k=0}^{n-1}\sum_{x\in E(n,\epsilon)}\delta_{f^k x}$$

where $E(n,\epsilon)$ is an arbitrary (ϵ, n)-separated subset [*See* ▶ Entropy in Ergodic Theory: $\forall x, y \in E(n,\epsilon)\ x \neq y \implies \exists 0 \leq k < n\ d(T^k x, T^k y) \geq \epsilon$.] of M with maximum cardinality. Then accumulation points for the weak star topology on the space of probability measures on M of $\mu_{n,\epsilon}$ when $n \to \infty$ and then $\epsilon \to 0$ are maximum entropy measures [185].

Let us quote two important additional properties, discovered by Margulis [179], that often hold for the maximum entropy measures:

- The *equidistribution of periodic points* with respect to some maximum entropy measure μ_M:

$$\mu_M = \lim_{n\to\infty}\frac{1}{\#\{x \in X : x = f^n x\}}\sum_{x=f^n x}\delta_x\,.$$

- The *holonomy invariance* which can be loosely interpreted by saying that the past and the future are independent conditionally on the present.

Other Points of View

Many other invariant measures are of interest in various contexts and we have made no attempt at completeness: for instance, invariant measures maximizing dimension [111,204], or pressure in the sense of the thermodynamical formalism [148,222], or some energy [9,81,146], or quasi-physical measures describing the dynamics around saddle-type invariant sets [104] or in systems with holes [75].

Tractable Chaotic Dynamics

The Palis Conjecture

There is, at this point, no general theory allowing the analysis of all dynamical systems or even of most of them despite many recent and exciting developments in the theory of generic C^1-diffeomorphisms [51,84]. In particular, the question of the generality in which physical measures exist remains open. One would like generic systems to have a *finite statistical description* (see Subsect. "Physical Measures"). This fails in some examples but these look exceptional and the following question is asked by Palis [200]:

> *Is it true that any dynamical system defined by a C^r-diffeomorphism on a compact manifold can be transformed by an arbitrarily small C^r-perturbation to another dynamical system having a finite statistical description?*

This is completely open though widely believed [*Observe, however, that such a statement is false for conservative diffeomorphisms with high order smoothness as KAM theory implies stable existence of invariant tori foliating a subset of positive volume.*]. Note that such a good description is not possible for all systems (see, e. g., [136,194]). Note that one would really like to ask about unperturbed "typical" [*The choice of the notion of typicality is a delicate issue. The Newhouse phenomenon shows that among C^2-diffeomorphisms of multidimensional compact manifolds, one cannot use topological genericity and get a positive answer. Popular notions are* prevalence *and* Kolmogorov genericity *– see the glossary.*] dynamical systems in a suitable sense, but of course this is even harder.

One is therefore led to make simplifying assumptions: typically of small dimension, uniform expansion/contraction or geometry.

Uniformly Expanding/Hyperbolic Systems

The most easily analyzed systems are those with uniform expansion and/or contraction, namely the uniformly expanding maps and uniformly hyperbolic diffeomorphisms, see ▶ Smooth Ergodic Theory. [*We require uniform hyperbolicity on the so-called chain recurrent set. This is equivalent to the usual Axiom-A and no-cycle condition.*] An important class of example is obtained as follows. Consider $A : \mathbb{R}^d \to \mathbb{R}^d$, a linear map preserving $\mathbb{Z}^d$ (i. e., A is a matrix with integer coefficients in the canonical basis) so that it defines a map $\bar{A} : \mathbb{T}^d \to \mathbb{T}^d$ on the torus. If there is a constant $\Lambda > 1$ such that for all $v \in \mathbb{R}^d$, $\|A.v\| \geq \Lambda\|v\|$ then $\bar{A}$ is a *uniformly expanding map*. If A has determinant ± 1 and no eigenvalue on the unit circle, then $\bar{A}$ is a *uniformly hyperbolic diffeomorphism* (▶ Smooth Ergodic Theory) (see also [60,148,215,233]). Moreover all C^1-perturbations of the previous examples are again uniformly expanding or uniformly hyperbolic. [*One can define uniform hyperbolicity for flows and an important class of examples is provided by the geodesic flow on compact* *manifolds with negative sectional curvature* [148].] These uniform systems are sometimes called "strongly chaotic".

Remark 4 Mañé Stability Theorem (see below) shows that uniform hyperbolicity is a very natural notion. One can also understand on a more technical level uniform hyperbolicity as what is needed to apply an implicit function theorem in some functional space (see, e. g., [233]).

The existence of a finite statistical description for such systems has been proved since the 1970s by Bowen, Ruelle and Sinai [54,218,235] (the expanding case is much simpler [162]).

Theorem 1 *Let $f\colon M \to M$ be a $C^{1+\alpha}$ map of a compact manifold. Assume f to be (i) a uniformly expanding map on M or (ii) a uniformly hyperbolic diffeomorphism.*

- *f admits a finite statistical description by ergodic and hyperbolic Sinai–Ruelle–Bowen measures (absolutely continuous in case (i)).*
- *f has finitely many ergodic maximum entropy measures, each of which makes f isomorphic to a finite state Markov chain. The periodic points are uniformly distributed according to some canonical average of these ergodic maximum entropy measures.*
- *f is topologically conjugate [Up to some negligible subset.] to a subshift of finite type (See the glossary.)*

The construction of absolutely continuous invariant measures for a uniformly expanding map f can be done in a rather direct way by considering the pushed forward measures $\frac{1}{n}\sum_{k=0}^{n-1} f_*^k$ Leb and taking weak star limits while preventing the appearance of singularities, by, e. g., bounding some Hölder norm of the density using expansion and distortion of f.

The classical approach to the uniformly hyperbolic dynamics [52,222,233] is through symbolic dynamics and coding. Under the above hypothesis one can build a finite partition of M which is tailored to the dynamics (a *Markov partition*) so that the corresponding symbolic dynamics has a very simple structure: it is a full shift $\{1,\ldots,d\}^{\mathbb{Z}}$, like in the example of the full tent map, or a subshifts of finite type. The above problems can then be solved using the thermodynamical formalism inspired from the statistical mechanics of one-dimensional ferromagnets [217]: ergodic properties are obtained through the spectral properties of a suitable transfer operator acting on some space of regular functions, e. g., the Hölder-continuous functions defined over the symbolic dynamics with respect to the distance $d(x,y) := \sum_{n\in\mathbb{Z}} 2^{-n} 1_{x_n \neq y_n}$ where $1_{s\neq t}$ is 1 if $s \neq t$, 0 otherwise.

A recent development [24,43,116] has been to find suitable Banach spaces to apply the transfer operator technique directly in the smooth setting, which not only avoids the complication of coding (or rather replace them with functional analytic preliminaries) but allows the use of the smoothness beyond Hölder-continuity which is important for finer ergodic properties.

Uniform expansion or hyperbolicity can easily be obstructed in a given system: a "bad" point (a critical point or a periodic point with multiplier with an eigenvalue on the unit circle) is enough. This leads to the study of other systems and has motivated many works devoted to relaxing the uniform hyperbolicity assumptions [51].

Pesin Theory

The most general such approach is *Pesin theory*. Let f be a $C^{1+\alpha}$-diffeomorphism [*It is an important open problem to determine to which extent Pesin theory could be generalized to the C^1 setting.*] f with an ergodic invariant measure μ. By Oseledets Theorem ► Smooth Ergodic Theory, for almost every x with respect to any invariant measure, the behavior of the differential $T_x f^n$ for n large is described by the Lyapunov exponents $\lambda_1,\ldots,\lambda_d$, at x. Pesin is able to build charts around almost every orbit in which this asymptotic linear behavior describes that of f at the first iteration. That is, there are diffeomorphisms $\Phi_x\colon U_x \subset M \to V_x \subset \mathbb{R}^d$ with a "reasonable dependence on x" such that the differential of $\Phi_{f^n x} \circ f^n \circ \Phi_x^{-1}$ at any point where it is defined is close to a diagonal matrix with entries $(e^{(\lambda_1 \pm \epsilon)n}, e^{(\lambda_2 \pm \epsilon)n}, \ldots, e^{(\lambda_d \pm \epsilon)n})$.

In this full generality, one already obtains significant results:

- The entropy is bounded by the expansion: $h(f,\mu) \leq \sum_{i=1}^d \lambda_i^+(\mu)$ [219]
- At almost every point x, there are strong stable resp. unstable manifolds $W^{ss}(x)$, resp. $W^{uu}(x)$, coinciding with the sets of points y such that $d(T^n x, T^n y) \to 0$ exponentially fast when $n \to \infty$, resp. $n \to -\infty$. The corresponding holonomies are *absolutely continuous* (see, e. g., [59]) like in the uniform case. This allows Ledrappier's definition of Sinai–Ruelle–Bowen measures [166] in that setting.
- Equality in the above formula holds if and only if μ is a Sinai–Ruelle–Bowen measure [167]. More generally the entropy can be computed as $\sum_{i=1}^d \gamma_i(\mu)\lambda_i^+(\mu)$ where the γ_i are some fractal dimensions related to the exponents.

Under the only assumption of hyperbolicity (i. e., no zero Lyapunov exponent almost everywhere), one gets further properties:

- Existence of an hyperbolic measure which is not periodic forces $h_{\mathrm{top}}(f) > 0$ [*However* $h(f, \mu)$ *can be zero.*] by [147].
- μ is *exact dimensional* [29,256]: the limit $\lim_{r\to 0} \log \mu(B(x, r))/ \log r$ exists μ-almost everywhere and is equal to the Hausdorff dimension of μ (the infimum of the Hausdorff dimension of the sets with full μ-measure). This is deduced from a more technical "asymptotic product structure" property of any such measure.

For hyperbolic Sinai–Ruelle–Bowen measures μ, one can then prove, e. g.,:

- *local ergodicity* [202]: μ has at most countably many ergodic components and μ-almost every point has a neighborhood whose Lebesgue-almost every point are contained in the basin of an ergodic component of μ.
- *Bernoulli* [198]: Each ergodic component of μ is conjugate in a measure-preserving way, up to a period, to a Bernoulli shift, that is, a full shift $\{1, \dots, N\}^{\mathbb{Z}}$ equipped with a product measure. This in particular implies mixing and sensitivity on initial conditions for a set of positive Lebesgue measure.

However, establishing even such a weak form of hyperbolicity is rather difficult. The fragility of this condition can be illustrated by the result [44,45] that the topologically generic area-preserving surface C^1-diffeomorphism is either uniformly hyperbolic or has Lebesgue almost everywhere vanishing Lyapunov exponents, hence is never non-uniformly hyperbolic! (but this is believed to be very specific to the very weak C^1 topology). Moreover, such weak hyperbolicity is not enough, with the current techniques, to build Sinai–Ruelle–Bowen measures or analyze maximum entropy measures only assuming non-zero Lyapunov exponents. Let us though quote two conjectures. The first one is from [251] [*We slightly strengthened Viana's statement for expository reasons.*]:

Conjecture 1 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact manifold. If Lebesgue-almost every point x has well-defined Lyapunov exponents in every direction and none of these exponents is zero, then there exists an absolutely continuous invariant σ-finite positive measure.*

The analogue of this conjecture has been proved for C^3 interval maps with unique critical point and negative Schwarzian derivative by Keller [150], but only partial results are available for diffeomorphisms [168].

We turn to measures of maximum entropy. As we said, C^∞ smoothness is enough to ensure their existence but this is through a functional-analytic argument (allowed by Yomdin theory [254]) which says nothing about their structure. Indeed, the following problem is open:

Conjecture 2 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact surface. If the topological entropy of f is nonzero then f has at most countably many ergodic invariant measures maximizing entropy.*

The analogue of this conjecture has been proved for $C^{1+\epsilon}$ interval maps [64,66,70]. In the above setting a classical result of Katok shows the existence of uniformly hyperbolic compact invariant subsets with topological entropy arbitrarily close to that of f implying the existence of many periodic points:

$$\limsup_{n\to\infty} \frac{1}{n} \log \#\{x \in M\colon f^n(x) = x\} \geq h_{\mathrm{top}}(f) \,.$$

The previous conjecture would follow from the following one:

Conjecture 3 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact manifold. There exists an invariant subset $X \subset M$, carrying all ergodic measures with maximum entropy, such that the restriction $f|X$ is conjugate to a countable state topological Markov shift (See the glossary.).*

Systems with Discontinuities

We now consider stronger assumptions to be able to build the relevant measures.

The simplest step beyond uniformity is to allow discontinuities, considering piecewise expanding maps. The discontinuities break the rigidity of the uniformly expanding situation. For instance, their symbolic dynamics are usually no longer subshifts of finite type though they still retain some "simplicity" in good cases (see [68]).

To understand the problem in constructing the absolutely continuous invariant measures, it is instructive to consider the pushed forwards of a smooth measure. Expansion tends to keep the measure smooth whereas discontinuities may pile it up, creating non-absolute continuity in the limit. One thus has to check that expansion wins. In dimension 1, a simple fact resolves the argument: under a high enough iterate, one can make the expansion arbitrarily large everywhere, whereas a small interval can be chopped into at most two pieces.

Lasota and Yorke [165] found a suitable framework. They considered C^2 interval maps with $|f'(x)| \geq \text{const} > 1$ except at finitely many points. They used the Ruelle transfer operator directly on the interval. Namely they studied

$$(L\phi)(x) = \sum_{y\in T^{-1}x} \frac{\phi(y)}{|T'(y)|}$$

acting on functions $\phi\colon [0,1] \to \mathbb{R}$ with bounded variation and obtained the invariant density as the eigenfunction associated to the eigenvalue 1. One can then prove a *Lasota–Yorke inequality* (which might more accurately be called Doeblin–Fortet since it was introduced in the theory of Markov chains much earlier):

$$\|L\phi\|_{\mathrm{BV}} \leq \alpha\|\phi\|_{\mathrm{BV}} + \beta\|\phi\|_1 \tag{3}$$

where $\|\cdot\|_{\mathrm{BV}}$, $\|\cdot\|_1$ are a strong and a weak norm, respectively and $\alpha < 1$ and $\beta < \infty$. One can then apply general theorems [143] or [196] (see [21] for a detailed presentation of this approach and its variants). Here α can essentially be taken as 2 (reflecting the locally simple discontinuities) divided by the minimum expansion: so $\alpha < 1$, perhaps after replacing T with an iterate. In particular, the existence of a finite statistical description then follows (see [61] for various generalizations and strengthenings of this result on the interval).

The situation in higher dimension is more complex for the reason explained above. One can obtain inequalities such as (3) on suitable if less simple functional spaces (see, e. g., [231]) but proving $\alpha < 1$ is another matter: discontinuities can get arbitrarily complex under iteration. [67,241] show that indeed, in dimension 2 and higher, piecewise uniform expansion (with a finite number of pieces) is not enough to ensure a finite statistical description if the pieces of the map have only finite smoothness. In dimension 2, resp. 3 or more, piecewise real-analytic, resp. piecewise affine, is enough to exclude such examples [65,240], resp. [242]. [82] has shown that, for any $r > 1$, an open and dense subset of piecewise C^r and expanding maps have a finite statistical description.

Piecewise hyperbolic diffeomorphisms are more difficult to analyze though several results (conditioned on technical assumptions that can be checked in many cases) are available [22,74,230,257].

Interval Maps with Critical Points

A more natural but also more difficult situation is a map for which the uniformity of the expansion fails because of the existence of critical points. [*Note that, by a theorem of Mañé a circle map without critical points or indifferent periodic point is either conjugate to a rotation or uniformly expanding* [181].]

A class which has been completely analyzed at the level of the above conjecture is that of *real-analytic families of maps of the interval* $f_t\colon [0,1] \to [0,1]$, $t \in I$, with a unique critical point, the main example being the quadratic family $Q_t(x) = tx(1-x)$ for $0 \leq t \leq 4$.

It is not very difficult to find quadratic maps with the following two types of behavior:

(stable) the orbit of Lebesgue-almost every $x \in [0,1]$ tends to an attracting periodic orbit;

(chaotic) there is an absolutely continuous invariant probability measure μ whose basin contains Lebesgue-almost every $x \in [0,1]$.

To realize the first it is enough to arrange the critical point to be periodic. One can easily prove that this stable behavior occurs on an open set of parameters –thus it is stable with respect to the parameter or the dynamical system. The second occurs for Q_4 with $\mu = dx/\sqrt{\pi x(1-x)}$. It is much more difficult to show that this chaotic behavior occurs for a set of parameters of positive Lebesgue measure. This is a theorem of Jakobson [145] for the quadratic family (see for a recent variant [265]). Let us sketch two main ingredients of the various proofs of this theorem. The first is *inducing*: around Lebesgue-almost every point $x \in [0,1]$ one tries to find a time $\tau(x)$ and an interval $J(x)$ such that $f^{\tau(x)}\colon J(x) \to f^{\tau(x)}(J(x))$ is a map with good expansion and distortion properties. This powerful idea appears in many disguises in the non-uniform hyperbolic theory (see for instance [133,262]). The second ingredient is *parameter exclusion*: one removes the parameters at which a good inducing scheme cannot be built. More precisely one proceeds inductively, performing simultaneously the inducing and the exclusion, the good properties of the early stage of the inducing allowing one to control the measure of the parameters that need to be excluded to continue [30,145]. Indeed, the expansion established at a given stage allows to transfer estimates in the dynamical space to the parameter space.

Using methods from complex analysis and renormalization theory one can go much further and prove the following difficult theorems (actually the product of the work of many people, including Avila, Graczyk, Kozlovski, Lyubich, de Melo, Moreira, Shen, van Strien, Swiatek), which in particular solves Palis conjecture in this setting:

Theorem 2 **([117,160,178])** *Stable maps (that is, such that Lebesgue almost every orbit converges to one of finitely many periodic orbits) form an open and dense set among C^r interval maps, for any $r \geq 2$. [In fact this is even true for polynomials.].*

The picture has been completed in the *unimodal case* (that is, with a unique critical point):

Theorem 3 **([19,20,117,159,178])** *Let $f_t\colon [0,1] \to [0,1]$, $t \in [t_0, t_1]$, be a real-analytic family of unimodal maps. Assume that it is not degenerate [f_{t_0} and f_{t_1} are not conjugate]. Then:*

- *The set of t such that f_t is chaotic in the above sense has positive Lebesgue measure;*
- *The set of t such that f_t is stable is open and dense;*
- *The remaining set of parameters has zero Lebesgue measure. [This set of "strange parameters" of zero Lebesgue measure has however positive Hausdorff dimension according to work of Avila and Moreira. In particular each of the following situations is realized on a set of parameters t of positive Hausdorff dimension: non-existence of the Birkhoff limit at Lebesgue-almost every point, the physical measure is δ_p for p a repelling fixed point, the physical measure is non-ergodic.]*

We note that the theory underlying the above theorem yields much more results, including a very paradoxical rigidity of typical analytic families as above. See [19].

Non-uniform Expansion/Contraction

Beyond the dimension 1, only partial results are available. The most general of those assume uniform contraction or expansion along some direction, restricting the non-uniform behavior to an invariant sub-bundle often one-dimensional, or "one-dimensional-like".

A first, simpler situation is when there is a *dominated decomposition* with a uniformly expanding term: there is a continuous and invariant splitting of the tangent bundle, $T_\Lambda M = E^{\mathrm{uu}} \oplus E^{\mathrm{cs}}$, over some Λ an attracting set: for all unit vectors $v^u \in E^{\mathrm{uu}}$, $v^c \in E^{\mathrm{cs}}$,

$$\|(f^n)'(x).v^u\| \geq C\lambda^n \quad \text{and}$$
$$\|(f^n)'(x).v^c\| \leq C\mu^n \|(f^n)'(x).v^u\| .$$

Standard techniques (pushing the Riemannian volume of a piece of unstable leaf and taking limits) allow the construction of Gibbs u-states as introduced by [205].

Theorem 4 (Alves–Bonatti–Viana [8]) *[A slightly different result is obtained in [49].] Let $f\colon M \to M$ be a C^2 diffeomorphism with an invariant compact subset Λ. Assume that there is a dominated splitting $T_\Lambda M = E^u \oplus E^{cs}$ such that, for some $c > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \prod_{k=0}^{n-1} \|f'(f^x)|E^{cs}\| \leq -c < 0$$

on a subset of Λ of positive Lebesgue measure. Then this subset is contained, up to a set of zero Lebesgue measure, in the union of the basins of finitely many ergodic and hyperbolic Sinai – Ruelle – Bowen measures.

The non-invertible, purely expansive version of the above theorem can be applied in particular to the following maps of the cylinder ($d \geq 16$, a is properly chosen close to 2 and α is small):

$$f(\theta, x) = (\mathrm{d}x \bmod 2\pi, a - x^2 + \epsilon \sin(\theta))$$

which are natural examples of maps with multidimensional expansion and critical lines considered by Viana [249]. A series of works have shown that the above maps fit in the above non-uniformly expanding setting with a proper control of the critical set and hence can be thoroughly analyzed through variants of the above theorem [4,249] and the references in [5]. For a, b properly chosen close to 2 and small ϵ, the following maps should be even more natural examples:

$$f(x, y) = (a - x^2 + \epsilon y, b - y^2 + \epsilon x) . \qquad (4)$$

However the inexistence of a dominated splitting has prevented the analysis of their physical measures. See [66,70] for their maximum entropy measures.

Cowieson and Young have used completely different techniques (thermodynamical formalism, Ledrappier–Young formula and Yomdin theory on entropy and smoothness) to prove the following result (see [13] for related work):

Theorem 5 (Cowieson–Young [83]) *Let $f\colon M \to M$ be a C^∞ diffeomorphism of a compact manifold. Assume that f admits an attractor $\Lambda \subset M$ on which the tangent bundle has an invariant continuous decomposition $T_\Lambda M = E^+ \oplus E^-$ such that all vectors of $E^+ \setminus \{0\}$, resp. $E^- \setminus \{0\}$, have positive, resp. negative, Lyapunov exponents. Then any zero-noise limit measure μ of f is a Sinai–Ruelle–Bowen measure and therefore, if it is ergodic and hyperbolic, a physical measure.*

One can hope, that typically, the latter ergodicity and hyperbolicity assumptions are satisfied (see, e. g., [27]).

By pushing classical techniques and introducing new ideas for generic maps with one expanding and one weakly contracting direction, Tsujii has been able to prove the following generic result (which can be viewed as a 2-dimensional extension of some of the one-dimensional results above: one adds a uniformly expanding direction):

Theorem 6 (Tsujii [243]) *Let M be a compact surface. Consider the space of C^{20} self-maps $f\colon M \to M$ which admits directions that are uniformly expanded [More precisely, there exists a continuous, forward invariant cone field which is uniformly expanded under the differential of f.]*

Then existence of a finite statistical description is both topologically generic and prevalent in this space of maps.

Hénon-Like Maps and Rank One Attractors

In 1976, Hénon [130] observed that the diffeomorphism of the plane

$$H_{a,b}(x,y) = (1 - ax^2 + y, bx)$$

seemed to present a "strange attractor" for $a = 1.4$ and $b = 0.3$, that is, the points of a numerically simulated orbit seemed to draw a set looking locally like the product of a segment with a Cantor set. This attractor seems to be supported by the unstable manifold $W^u(P)$ of the hyperbolic fixed point P with positive absciss. On the other hand, the Plykin classification [207] excluded the existence of a uniformly hyperbolic attractor for a dissipative surface diffeomorphism.

For almost twenty years the question of the existence of such an attractor (by opposition to an attracting periodic orbit with a very long period) remained open. Indeed, one knew since Newhouse that, for many such maps, there exist infinitely many such periodic orbits which are very difficult to distinguish numerically. But in 1991 Benedicks and Carleson succeeded in proposing an argument refining (with considerable difficulties) their earlier proof of Jakobson one-dimensional theorem and established the first part of the following theorem:

Theorem 7 (Benedicks–Carleson [31]) *For any $\epsilon > 0$, for $|b|$ small enough, there is a set A with* $\mathrm{Leb}(A) > 0$ *satisfying: for all $a \in A$, there exists $z \in W^u(P)$ such that*

- *The orbit of z is dense in $W^u(P)$;*
- $\liminf_{n\to\infty} \frac{1}{n} \log \|(f^n)'(z)\| > 0.$

Further properties were then established, especially by Benedicks, Viana, Wang, Young [32,34,35,264]. Let us quote the following theorem of Wang and Young which includes the previous results:

Theorem 8 ([264]) *Let $T_{\mathrm{ab}}\colon S^1 \times [-1,1] \to S^1 \times [-1,1]$ be such that*

- *$T_{a0}(S^1 \times [-1,1]) \subset S^1 \times \{0\}$;*
- *For $b > 0$, T_{ab} is a diffeomorphism on its image with*

  $$c^{-1} b \le |\det T_{\mathrm{ab}}(x,y)| \le c \cdot b$$

  for some $c > 1$ and all $(x,y) \in S^1 \times [-1,1]$ and all (a,b).

Let $f_a\colon S^1 \to S^1$ be the restriction of T_{a0}. Assume that $f = f_0$ satisfies:

- *Non-degenerate critical points: $f'(c) = 0 \implies f''(c) \neq 0$;*
- *Negative Schwarzian derivative: for all $x \in S^1$ non-critical, $f'''(x)/f'(x) - 3/2(f''(x)/f'(x))^2 < 0$;*
- *No indifferent or attracting periodic point, i. e., x such that $f^n(x) = x$ and $|(f)'(x)| \le 1$;*
- *Misiurewicz condition: $d(f^n c, d) > c > 0$ for all $n \ge 1$ and all critical points c, d.*

Assume the following transversality condition on f at $a = 0$: for every critical point c, $(\mathrm{d}/da)(f_a(c_a) - p_a) \neq 0$ if c_a is the critical point of f_a near c and p_a is the point having the same itinerary under f_a as $f(c)$ under c. Assume the following non-degeneracy of T: $f_0'(c) = 0 \implies \partial T_{00}(c,0)/\partial y \neq 0$.

- *T_{ab} restricted to a neighborhood of $S^1 \times \{0\}$ has a finite statistical description by a number of hyperbolic Sinai–Ruelle–Bowen measures bounded by the number of critical points of f;*
- *There is exponential decay of correlations and a Central Limit Theorem (see below) – except, in an obvious way, if there is a periodic interval with period > 1;*
- *There is a natural coding of the orbits that remains for ever close to $S^1 \times \{0\}$ by a closed invariant subset of a full shift.*

Very importantly, the above dynamical situation has been shown to occur near typical homoclinic tangencies: [190] proved that there is an open and dense subset of the set of all C^3 families of diffeomorphisms unfolding a first homoclinic tangency such that the above holds. However [201] shows that the set of parameters with a Henon-like attractor has zero Lebesgue density at the bifurcation itself, at least under an assumption on the so-called stable and unstable Hausdorff dimensions. [95] establishes positive density for another type of bifurcation. Furthermore [191] has related the Hausdorff dimensions to the abundance of uniformly hyperbolic dynamics near the tangency.

[248] is able to treat situations with more than one contracting direction. More recently [266] has proposed a rather general framework, with easily checkable assumptions in order to establish the existence of such dynamics in various applications. See also [122,252] for applications.

Statistical Properties

The ergodic theorem asserts that time averages of integrable functions converge to phase space averages for any ergodic system. The speed of convergence is quite arbitrary in that generality [161] (only upcrossing inequalities seem to be available [38,132]), however many results are available under very natural hypothesis as we are going to explain in this section. The underlying idea is that for sufficiently chaotic dynamics T and reasonably smooth observables ϕ, the time averages

$$A_n\phi(x) := \frac{1}{n}\sum_{k=0}^{n-1} \phi \circ T^k(x)$$

should behave as averages of independent and identically distributed random variables and therefore satisfy the classical limit theorems of probability theory.

The dynamical systems which are amenable to current technology are in a large part [*But other approaches are possible. Let us quote the work* [98] *on partially hyperbolic systems, for instance.*] those that can be reduced to the following type:

Definition 1 Let $T: X \to X$ be a nonsingular map on a probability, metric space $(X, \mathcal{B}, \mu, d)$ with bounded diameter, preserving the probability measure μ. This map is said to be **Gibbs–Markov** if there exists a countable (measurable) partition α of X such that:

1. For all $a \in \alpha$, T is injective on a and $T(a)$ is a union of elements of α.
2. There exists $\lambda > 1$ such that, for all $a \in \alpha$, for all points $x, y \in a$, $d(Tx, Ty) \geq \lambda d(x, y)$.
3. Let Jac be the *inverse* of the Jacobian of T. There exists $C > 0$ such that, for all $a \in \alpha$, for all points $x, y \in a$, $|1 - Jac(x)/Jac(y)| \leq Cd(Tx, Ty)$.
4. The map T has the "big image property": $\inf_{a\in\alpha} \mu(Ta) > 0$.

Some piecewise expanding and C^2 maps are obviously Gibbs–Markov but the real point is that many dynamics can be reduced to that class by the use of inducing and tower constructions as in [262], in particular. This includes possibly piecewise uniformly hyperbolic diffeomorphisms, Collet–Eckmann maps of the interval [21] (typical chaotic maps in the quadratic family), billiards with convex scatterers [262], the stadium billiard [71], Hénon-like mappings [266].

We note that in many cases one is led to first analyze mixing properties through *decay of correlations*, i. e., to prove inequalities of the type [21]:

$$\left|\int_X \phi \cdot \psi \circ T^n \, \mathrm{d}\mu - \int_X \phi \, \mathrm{d}\mu \int_X \psi \, \mathrm{d}\mu\right| \leq \|\phi\| \cdot \|\psi\|_w \cdot a_n \tag{5}$$

where $(a_n)_{n\geq 1}$ is some sequence converging to zero, e. g., $a_n = \mathrm{e}^{-\lambda n}, 1/n^\alpha, \ldots$ and $\|\cdot\|, \|\cdot\|_w$ a strong and a weak norm (e. g., the variation norm and the L^1 norm). These rates of decay are often linked with return times statistics [263]. Rather general schemes have been developed to deduce various limit theorems such as those presented below from sufficiently quick decay of correlations (see notably [175] based on a dynamical variant of [113]).

Probabilistic Limit Theorems

The foremost limit property is the following:

Definition 2 A class C of functions $\phi: X \to \mathbb{R}$ is said to satisfy the Central Limit Theorem if the following holds: for all $\phi \in C$, there is a number $\sigma = \sigma(\phi) > 0$ such that:

$$\lim_{n\to\infty} \mu\left(\left\{x \in X : \frac{A_n\phi(x) - \int \phi \, \mathrm{d}\mu}{\sigma n^{-1/2}} \leq t\right\}\right) = \int_{-\infty}^{t} \mathrm{e}^{-x^2/2\sigma^2} \frac{\mathrm{d}x}{\sqrt{2\pi}\sigma} \tag{6}$$

except for the degenerate case when $\phi(x) = \psi(Tx) - \phi(x) + \text{const}$.

The Central Limit Theorem can be seen in many cases as essentially a by-product of fast decay of correlations [175], i. e., if $\sum_{n\geq 0} a_n < \infty$ in the notations of Eq. (5). It has been established for Hölder-continuous observables for many systems together with their natural invariant measures including: uniformly hyperbolic attractors, piecewise expanding maps of the interval [174], Collet–Eckmann unimodal maps on the interval [152,260], piecewise hyperbolic maps [74], billiards with convex scatterers [238], Hénon-like maps [35].

Remark 5 The classical Central Limit Theorem holds for square-integrable random variables [193]. For maps exhibiting *intermittency* (e.g, interval maps like $f(x) = x + x^{1+\alpha} \bmod 1$ with an indifferent fixed point at 0) the invariant density has singularities and the integrability condition is non longer automatic for smooth functions. One can then observe convergence to stable laws, instead of the normal law [114].

A natural question is the speed of the convergence in (6). The *Berry–Esseen inequality*:

$$\left|\mu\left(\left\{\frac{A_n\phi(x) - \int \phi \, \mathrm{d}\mu}{\sigma n^{-1/2}} \leq t\right\}\right) - \int_{-\infty}^{t} \mathrm{e}^{-x^2/2\sigma^2} \frac{\mathrm{d}x}{\sqrt{2\pi}\sigma}\right| \leq \frac{C}{n^{\delta/2}}$$

for some $\delta > 0$. It holds with $\delta = 1$ in the classical, probabilistic setting.

The *Local Limit Theorem* looks at a finer scale, asserting that for any finite interval $[a, b]$, any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \sqrt{n}\mu\left(\left\{x \in X\colon nA_n\phi(x) \in [a,b] + \sqrt{n}t + n\int \phi \,\mathrm{d}\mu\right\}\right) = |b-a|\frac{\mathrm{e}^{-t^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\,.$$

Both the Berry–Esseen inequality and the local limit theorem have been shown to hold for non-uniformly expanding maps [115] (also [62,216]).

Almost Sure Results

It is very natural to try and describe the statistical properties of the averages $A_n(x)$ for almost every x, instead of the weaker above statements in probability over x. An important such property is the *almost sure invariance principle*. It asks for the discrete random walk defined by the increments of $\phi \circ T^n(x)$ to converge, after a suitable renormalization, to a Brownian motion. This has been proved for systems with various degrees of hyperbolicity [92,98,135,183]. Another one is the almost sure Central Limit Theorem. In the independent case (e.g., if $X_1, X_2, \ldots$ are independent and identically distributed random variables in L^2 with zero average and unit variance), the almost sure Central Limit Theorem states that, almost surely:

$$\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\delta_{\sum_{j=0}^{k-1} X_j/\sqrt{k}}$$

converges in law to the normal distribution. This implies, that, almost surely, for any $t \in \mathbb{R}$:

$$\lim_{n\to\infty}\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\delta_{\left\{\sum_{j=0}^{k-1} X_j/\sqrt{k}\le t\right\}} = \int_{-\infty}^{t} \mathrm{e}^{-x^2/2\sigma^2}\frac{\mathrm{d}x}{\sqrt{2\pi}\sigma}$$

compare to (6).

A general approach is developed in [73], covering Gibbs–Markov maps and those that can be reduced to it. They show in particular that the dynamical properties needed for the classical Central Limit Theorem in fact suffice to prove the above almost invariance principle and even the almost sure version of the Central Limit Theorem (using general probabilistic results, see [37,255]).

Other Statistical Properties

Essentially all the statistical properties of sums of independent identically distributed random variables can be established for tractable systems. Thus one can also prove large deviations [156,172,259], iterated law of the logarithm, etc. We note that the monograph [78] contains a nice introduction to the current work in this area.

Orbit Complexity

The orbit complexity of a dynamical system $f\colon M \to M$ is measured by its topological and measured entropies. We refer to ► Entropy in Ergodic Theory for detailed definitions.

The Variational Principle

Bowen–Dinaburg and Katok formulae can be interpreted as meaning that the topological entropy counts the number of arbitrary orbits whereas the measured entropy counts the number of orbits relevant for the given measure. In most situations, and in particular for continuous self-map of compact metric spaces, the following *variational principle* holds:

$$h_{\text{top}}(f) = \sup_{\mu\in M(f)} h(f,\mu)$$

where $M(f)$ is the set of all invariant probability measures.

This is all the more striking in light of the fact that for many systems, the set of points which are typical from the point of view of ergodic theory [*for instance, those x such that $\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x)$ exists for all continuous functions ϕ.*] is topologically negligible [*A subset of a countable union of closed sets with empty interior, that is, meager or first Baire category.*].

Strict Inequality For a fixed invariant measure, one can only assert that $h(f,\mu) \le h_{\text{top}}(f)$. One should be aware that this inequality may be strict even for a measure with full support. For instance, it is not difficult to check that the full tent map with $h_{\text{top}}(f) = \log 2$, admits ergodic invariant measures with full support and zero entropy.

There are also examples of dynamical systems preserving Lebesgue measure which have simultaneously positive topological entropy and zero entropy with respect to Lebesgue measure. That this occurs for C^1 surface diffeomorphisms preserving area is a simple consequence of a theorem of Bochi [44] according to which, generically in the C^1 topology, such a diffeomorphism is either uniformly hyperbolic or with Lyapunov exponents Lebesgue-almost everywhere zero. [*Indeed, it is easy to build such a diffeomorphism having both a uniformly hyperbolic compact invariant subset which will have robustly positive topological entropy and a non-degenerate elliptic fixed point which will prevent uniform hyperbolicity and therefore force all Lyapunov exponents to be zero. But Ruelle–Margulis* *inequality then implies that the entropy with respect to Lebesgue measure is zero.*] Smooth examples also exist [46].

Remark 6 Algorithmic complexity [170] suggests another way to look at orbit complexity. One obtains in fact in this way another formula for the entropy. However this point of view becomes interesting in some settings, like extended systems defined by partial differential equations in unbounded space. Recently, [47] has used this approach to build interesting invariant measures.

Orbit Complexity on the Set of Measures

We have considered the entropies of each invariant measure separately, sharing only the common roof of topological entropy. One may ask how these different complexity sit together. A first answer is given by the following theorem in the symbolic and continuous setting. Let .

Theorem 9 (Downarowicz–Serafin [102]) *Let K be a Choquet simplex and $H\colon K \to \mathbb{R}$ be a convex function.*

Say that H is realized by a self-map $f\colon X \to X$ and its set $M(f)$ of f-invariant probability measures equipped with the weak star topology if the following holds. There exists an affine homeomorphism $\Psi\colon M(f) \to K$ such that, if $h\colon M(f) \to [0,\infty]$ is the entropy function, $H = h \circ \Psi$.

Then

- *H is realized by some continuous self-map of a compact space if and only if it is an increasing limit of upper semi-continuous and affine functions.*
- *H is realized by some subshift on a finite alphabet, i. e., by the left shift on a closed invariant subset Σ of $\{1,2,\dots,N\}^{\mathbb{Z}}$ for some $N < \infty$, if and only if it is upper semi-continuous*

Thus, in both the symbolic and continuous settings it is possible to have a unique invariant measure with any prescribed entropy. This stands in contrast to surface $C^{1+\epsilon}$-diffeomorphisms for which the set of the entropies of ergodic invariant measures is always the interval $[0, h_{\text{top}}(f)]$ as a consequence of [147].

Local Complexity

Recall that the topological entropy can be computed as: $h_{\text{top}}(f) = \lim_{\epsilon\to 0} h_{\text{top}}(f,\epsilon)$ where:

$$h_{\text{top}}(f,\epsilon) := \lim_{n\to\infty} \frac{1}{n} \log s(\delta, n, X)$$

where $s(\delta, n, E)$ is the maximum cardinality of a subset S of E such that:

$$x \neq y \implies \exists 0 \le k < n \; d(f^k x, f^k y) \ge \epsilon$$

(see Bowen's formula of the topological entropy ▶ Entropy in Ergodic Theory). Likewise, the measure-theoretic entropy $h(T,\mu)$ of an ergodic invariant probability measure μ is $\lim_{\epsilon\to 0} h(f,\mu,\epsilon)$ where:

$$h(T,\epsilon) := \lim_{n\to\infty} \frac{1}{n} \log r(\delta, n, \mu)$$

where $r(\delta, n, \mu)$ is the minimum cardinality of $C \subset X$ such that

$$\mu\left(\left\{x \in X\colon \exists y \in C \quad \text{such that} \quad \forall 0 \le k < n \; d(f^k x, f^k y) < \epsilon\right\}\right) > 1/2\,.$$

One can ask *at which scales does entropy arise for a given dynamical system?*, i. e., how the above quantities $h(T,\epsilon)$, $h(T,\mu,\epsilon)$ converge when $\epsilon \to 0$.

An answer is provided by the *local entropy*. [*This quantity was introduced by Misiurewicz* [186] *under the name* conditional topological entropy *and is called* tail entropy *by Downarowicz* [100].] For a continuous map f of a compact metric space X, it is defined as:

$$\begin{aligned} h_{\text{loc}}(f) &:= \lim_{\epsilon\to 0} h_{\text{loc}}(f,\epsilon) \quad \text{with} \\ h_{\text{loc}}(f,\epsilon) &:= \sup_{x\in X} h_{\text{loc}}(f,\epsilon,x) \quad \text{and} \\ h_{\text{loc}}(f,\epsilon,x) &:= \lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log s\left(\delta, n, \left\{y \in X\colon \forall k \ge 0 \; d(f^k y, f^k x) < \epsilon\right\}\right) \end{aligned}$$

Clearly from the above formulas:

$$\begin{aligned} &h_{\text{top}}(f) \le h_{\text{top}}(f,\delta) + h_{\text{loc}}(f,\delta) \quad \text{and} \\ &h(f,\mu) \le h(f,\mu,\delta) + h_{\text{loc}}(f,\delta)\,. \end{aligned}$$

Thus the local entropy bounds the defect in uniformity with respect to the measure of the pointwise limit $h(f,\mu) = \lim_{\delta\to 0} h(f,\mu,\delta)$. An exercise in topology shows that the local entropy therefore also bounds the defect in upper semicontinuity of $\mu \mapsto h(f,\mu)$. In fact, by a result of Downarowicz [100] (extended by David Burguet to the non-invertible case), there is a *local variational principle*:

$$\begin{aligned} h_{\text{loc}}(f) &= \lim_{\delta\to 0} \left(\sup_{\mu} h(f,\mu) - h(f,\mu,\delta) \right) \\ &= \sup_{\mu} \limsup_{\nu\to\mu} h(f,\nu) - h(f,\mu) \end{aligned}$$

for any continuous self-map f of a compact metric space.

The local entropy is easily bounded for smooth maps using Yomdin's theory:

Theorem 10 ([64]) *For any C^r map f of a compact manifold, $h_{\text{loc}}(f) \leq \frac{d}{r} \log \sup_x \|f'(x)\|$. In particular, $h_{\text{loc}}(f) = 0$ if $r = \infty$.*

Thus C^∞ smoothness implies the existence of a maximum entropy measure (this was proved first by Newhouse) and the existence of *symbolic extension*: a subshift over a finite alphabet $\sigma : \Sigma \to \Sigma$ and a continuous and onto map $\pi : \Sigma \to M$ such that $\pi \circ \sigma = f \circ \sigma$. More precisely,

Theorem 11 (Boyle, Fiebig, Fiebig [56]) *Given a homeomorphism f of a compact metric space X, there exists a principal symbolic extension $\sigma : \Sigma \to \Sigma$, i. e., a symbolic extension such that, for every σ-invariant probability measure ν, $h(\sigma, \nu) = h(f, \nu \circ \pi^{-1})$, if and only if $h_{\text{loc}}(f) = 0$.*

We refer to [55,101] for further results, including a realization theorem showing that the continuity properties of the measured entropy are responsible for the properties of symbolic extensions and also results in finite smoothness.

Global Simplicity

One can marvel at the power of mathematical analysis to analyze such complex evolutions. Of course another way to look at this is to remark that this analysis is possible *once this evolution has been fitted in a simple setting*: one had to move focus away from an individual, unpredictable orbit, of, say, the full tent map to the set of all the orbits of that map, which is essentially the set of all infinite sequences over two symbols: a very simple set indeed corresponding to full *combinatorial freedom* [*[253] describes a weakening of this which holds for all positive entropy symbolic dynamics.*]. The complete description of a given typical orbit requires an infinite amount of information, whereas the set of all orbits has a finite and very tractable definition. The complexity of the individual orbits is seen now as coming from purely random choices inside a simple structure.

The classical systems, namely uniformly expanding maps or hyperbolic diffeomorphisms of compact spaces, have a *simple* symbolic dynamics. It is not necessarily a full shift like for the tent map, but it is a subshift of finite type, i. e., a subshift obtained from a full shift by forbidding finitely many finite subwords. What happens outside of the uniform setting?

A fundamental example is provided by *piecewise monotone maps*, i. e., interval maps with finitely many critical points or discontinuities. The partition cut by these points defines a symbolic dynamics. This subshift is usually not of finite type. Indeed, the topological entropy taking arbitrary finite nonnegative values [*For instance, the topological entropy of the β-transformation, $x \mapsto \beta x \bmod 1$, is $\log \beta$ for $\beta \geq 1$.*], a representation that respects it has to use an uncountable class of models. In particular models defined by finite data, like the subshifts of finite type, cannot be generally adequate. However there are tractable "almost finite representations" in the following senses:

Most symbolic dynamics $\Sigma(T)$ of piecewise monotone maps T can be defined by finitely many infinite sequences, the *kneading invariants* of Milnor and Thurston: $\kappa_0^+, \kappa_1^-, \kappa_1^+, \ldots, \kappa_{d+1}^- \in \{0, \ldots, d\}^{\mathbb{N}}$ if d is the number of critical/discontinuity points. [*The kneading invariants are the suitably defined (left and right) itineraries of the critical/discontinuity points and endpoints.*] Namely,

$$\Sigma(T) = \Big\{\alpha \in \{0, \ldots, d\}^{\mathbb{N}} : \forall n \geq 0 \qquad \kappa_{\alpha_n}^+ \preceq \sigma^n \alpha \preceq \kappa_{\alpha_n+1}^- \Big\}$$

where $\preceq$ is a total order on $\{0, \ldots, d\}^{\mathbb{N}}$ making the coding $x \mapsto \alpha$ non-decreasing. Observe how the kneading invariants determine $\Sigma(T)$ in an effective way: knowing their first n symbols is enough to know the sequences of length n which begin sequences of $\Sigma(T)$. We refer to [91] for the wealth of information that can be extracted from these kneading invariants following Milnor and Thurston [184].

This form of global simplicity can be extended to other classes of non-uniformly expanding maps, including those like Eq. (4) using the notions of subshifts and puzzles of *quasi-finite type* [68,69]. This leads to the notion and analysis of *entropy-expanding maps*, a new open class of non-uniformly expanding maps admitting critical hypersurfaces, defined purely in terms of entropies including the otherwise untractable examples of Eq. (4).

A generalization of the representation of uniform systems by subshifts of finite type is provided by *strongly positive recurrent countable state Markov shifts*, a subclass of Markov shifts (*see glossary*.) which shares many properties with the subshifts of finite type [57,123,124,224,229].

These "simple" systems admit a *classification result* which in particular identifies their measures with entropy close to the maximum [57]. Such a classification generalizes [164]. The "ideology" here is that complexity of individual orbits in a simple setting must come from randomness, but purely random systems are classified by their entropy according to Ornstein [197].

Stability

By definition, chaotic dynamical systems have orbits which are unstable and numerically unpredictable. It is all the more surprising that, once one accepts to consider their dynamics globally, they exhibit very good stability properties.

Structural Stability

A simple form of stability is *structural stability*: a system $f\colon M \to M$ is structurally C^r-stable if any system g sufficiently C^r-close to f is topologically the same as f, formally: g is topologically conjugate, i. e., there is some homeomorphism [*If h were* C^1, *the conjugacy would imply, among other things, that for every p-periodic point:* $\det((f^p)'(x)) = \det((g^p)'(h(p)))$, *a much too strong requirement*.] $h\colon M \to M$ mapping the orbits of f to those of g, i. e., $g \circ h = h \circ f$.

Andronov and Pontryaguin argued in the 1930's that only such structurally stable systems are physically relevant. Their idea was that the model of a physical system is always known only to some degree of approximation, hence mathematical model whose structure depends on arbitrarily small changes should be irrelevant.

A first question is: *What are these structurally stable systems?* The answer is quite striking:

Theorem 12 (Mañé [182]) *Let* $f\colon M \to M$ *be a* C^1 *diffeomorphism of a compact manifold.*

f is structurally stable among C^1*-diffeomorphisms of M if and only if f is uniformly hyperbolic on its chain recurrent set. [A point x is chain recurrent if, for all* $\epsilon > 0$, *there exists a finite sequence* $x_0, x_1, \dots, x_n$ *such that* $x_0 = x_n = x$ *and* $d(f(x_k), x_{k+1}) < \epsilon$. *The chain recurrent set is the set of all chain recurrent points.]*

A basic idea in the proof of the theorem is that failure of uniform hyperbolicity gives the opportunity to make an arbitrarily small perturbation contradicting the structural stability. In higher smoothness the required perturbation lemmas (e. g., the closing lemma [14,148,180]) are not available.

We note that uniform hyperbolicity without invertibility does not imply C^1-stability [210].

A second question is: *are these stable systems dense?* (So that one could offer structurally stable models for all physical situations). A deep discovery around 1970 is that this is not the case:

Theorem 13 (Abraham–Smale, Simon [3,234]) *For any* $r \geq 1$ *and any compact manifold M of dimension* ≥ 3, *the set of uniformly hyperbolic diffeomorphisms is not dense in the space of* C^r *diffeomorphisms of M. [They use the phenomenon called "heterodimensional homoclinic intersections".]*

Theorem 14 (Newhouse [194]) *For any* $r \geq 2$, *for any compact manifold M of dimension* ≥ 2, *the set of uniformly hyperbolic diffeomorphisms is not dense in the space of* C^r *diffeomorphisms of M. More precisely, there exists a nonempty open subset in this space which contains a dense* G_δ *subset of diffeomorphisms with infinitely many periodic sinks. [So these diffeomorphisms have no finite statistical description.]*

Observe that it is possible that uniform hyperbolicity could be dense among *surface* C^1*-diffeomorphisms* (this is the case for C^1 circle maps by a theorem of Jakobson [145]).

In light of Mañé C^1-stability theorem this implies that *structurally stable systems are not dense*, thus one can robustly see behaviors that are topologically modified by arbitrarily small perturbations (at least in the C^1-topology)! So one needs to look beyond these and face that topological properties of relevant dynamical system are not determined from "finite data". It is natural to ask whether the dynamics is almost determined by "sufficient data".

Continuity Properties of the Topological Dynamics

Structural stability asks the topological dynamics to remain *unchanged* by a small perturbation. It is probably at least as interesting to ask it to *change continuously*. This raises the delicate question of which topology should be put on the rather wild set of topological conjugacy classes. It is perhaps more natural to associate to the system a topological invariant taking value in a more manageable set and ask whether the resulting map is continuous.

A first possibility is Zeeman's *Tolerance Stability Conjecture*. He associated to each diffeomorphism the set of all the closures of all of its orbits and he asked whether the resulting map is continuous on a dense G_δ subset of the class of C^r-diffeomorphisms for any $r \geq 0$. This conjecture remains open, we refer to [85] for a discussion and related progress.

A simpler possibility is to consider our favorite topological invariant, the topological entropy, and thus ask whether the dynamical complexity as measured by the entropy is a stable phenomenon. $f \mapsto h_{\text{top}}(f)$ is lower semicontinuous for f among C^0 maps of the interval [*On the set of interval maps with a bounded number of critical points, the entropy is continuous* [188]. *Also* $t \mapsto h_{\text{top}}(Q_t)$ *is non-decreasing by complex arguments* [91], *though it is a non-smooth function*.] [187] and for f among $C^{1+\epsilon}$-diffeomorphisms of a compact surface [147]. [*It is an important open question whether this actually holds for* C^1*-diffeomorphisms. It fails for homeomorphisms* [214].] In both cases, one shows the existence of structurally stable invariant uniformly expanding or hyperbolic subsets with topological entropy close to that of the whole dynamics.

On the other hand $f \mapsto h_{\text{top}}(f)$ is upper semi-continuous for C^∞ maps [195,254].

Statistical Stability

Statistical stability is the property that *deterministic perturbations* of the dynamical system cause only small changes in the physical measures, usually with respect to the weak star topology on the space of measures. When the physical measure μ_g is uniquely defined for all systems g near f, statistic stability is the continuity of the map $g \mapsto \mu_g$ thus defined.

Statistical stability is known in the uniform setting and also in the piecewise uniform case, provided the expansion is strong enough [25,42,151] (otherwise there are counterexamples, even in the one-dimensional case).

It also holds in some non-uniform settings without critical behavior, in particular for maps with dominated splitting satisfying robustly a separation condition between the positive and negative Lyapunov exponents [247] (see also [7]).

However statistical unstability occurs in dynamics with critical behaviors [18]:

Theorem 15 *Consider the quadratic family $Q_t(x) = tx(1-x)$, $t \in [0,4]$. Let $I \subset [0,4]$ be the full measure subset of $[0,4]$ of good parameters t such that in particular Q_t admits a physical measure (necessarily unique). For $t \in I$, let μ_t be this physical measure.*

Lebesgue almost every $t \in I$ such that μ_t is not carried by a periodic orbit [At such parameters, statistical stability is easily proved.], is a discontinuity point of the map $t \mapsto \mu_t$ [$M([0,1])$ is equipped with the vague topology.]. However, for Lebesgue almost every t, there is a subset $I_t \subset I$ for which t is a Lebesgue density point [t is a Lebesgue density point if for all $r > 0$, the Lebesgue measure $m(I_t \cap [t-r, t+r]) > 0$ and $\lim_{\epsilon \to 0} m(I_t \cap [t-\epsilon, t+\epsilon])/2\epsilon = 1$.] and such that $\mu\colon I_t \to M([0,1])$ is continuous at t.

Stochastic Stability

A physically motivated and a technically easier approach is to study stability of the physical measure under *stochastic perturbations*. For simplicity let us consider a diffeomorphism of a compact subset of $\mathbb{R}^d$ allowing for a direct definition of *additive noise*. Let $\psi(x)\mathrm{d}x$ be an absolutely continuous probability law with compact support. [*Sometimes additional properties are required of the density, e. g., $\psi(x) = \phi(x)1_B$ where 1_B is the characteristic function of the unit ball and $C^{-1} \le \phi(x) \le C$ for some $C < \infty$.*] For $\epsilon > 0$, consider the Markov chain f_ϵ with state space M and transition probabilities:

$$P_\epsilon(x, A) = \int_A \psi\left(\frac{y - f(x)}{\epsilon}\right) \frac{\mathrm{d}y}{\epsilon^d} .$$

The evolution of measures is given by: $(f_\epsilon \mu)(A) = \int_M P_\epsilon(x, A)\, \mathrm{d}\mu$. Under rather weak irreducibility assumptions on f, f_ϵ has a *unique invariant measure* μ_ϵ (contrarily to f) and μ_ϵ is absolutely continuous. When f has a unique physical measure μ, it is said to be *stochastically stable* if $\lim_{\epsilon \to 0} \mu_\epsilon = \mu$ in the appropriate topology (the weak star topology unless otherwise specified).

It turns out that stochastic stability is a rather common property of Sinai–Ruelle–Bowen measures. It holds not only for uniformly expanding maps or hyperbolic diffeomorphisms [153,258], but also for most interval maps [25], for partially hyperbolic systems of the type $E^u \oplus E^{cs}$ or Hénon-like diffeomorphisms [6,13,33,40]. We refer to the monographs [21,41] for more background and results.

Untreated Topics

For reasons of space and time, many important topics have been left out of this article. Let us list some of them.

Other phenomena related to chaotic dynamics have been studied: entrance times [77,79], spectral properties and dynamical zeta functions [21], escape rates [104], dimension [204], differentiability of physical measures with respect to parameters [99,223], entropies and volume or homological growth rates [118,139,254].

As far as the structure of the setting is concerned, one can go beyond maps or diffeomorphisms or flows and study: more general group actions [128]; holomorphic and meromorphic structures [72] and the references therein; symplectic or volume-preserving [90,137] and in particular the Pugh–Shub program around stable ergodicity of partially hyperbolic systems [211]; random iterations [17,154,155].

A number of important problems have motivated the study of special forms of chaotic dynamics: equidistribution in number theory [105,109,138] and geometry [236]; quantum chaos [10,125]; chaotic control [227]; analysis of algorithms [245].

We have also omitted the important problem of *applying the above results*. Perhaps because of the lack of a general theory, this can often be a challenge (see for instance [244] for the already complex problem of verifying *uniform hyperbolicity* for a singular flow). Liverani has shown how theoretical results can lead to precise and efficient estimates for the toy model of piecewise expanding interval maps [176]. Ergodic theory implies that, in some settings at least, *adding noise may make some estimates more precise* (see [157]). We refer to [106] and the reference therein.

Future Directions

We conclude this article by a (very partial) selection of open problems.

General Theory

In dimension 1, we have seen that the analogue of the Palis conjecture (see above) is established (Theorem 2). However the description of the typical dynamics in Kolmogorov sense is only known in the unimodal, non-degenerate case by Theorem 3. Indeed, results like [63] suggest that the multimodal picture could be more complex.

In higher dimensions, our understanding is much more limited. As far as a general theory is concerned, a deep problem is the paucity of results on the generic dynamics in C^r smoothness with $r > 1$. The remarkable current progress in generic dynamics (culminating in the proof of the weak Palis conjecture, see [84], the references therein and [51] for background) seems restricted to the C^1 topology because of the lack of fundamental tools (e. g., closing lemmas) in higher smoothness. But Pesin theory requires higher smoothness at least technically. This is not only a hard technical issue but generic properties of physical measures, when they have been analyzed are often completely different between the C^1 case and higher smoothness [45].

Physical Measures

In higher dimensions, Benedicks and Carleson analysis of the Hénon map has given rise to a rather general theory of Hénon-like maps and more generally of the dynamical phenomena associated to homoclinic tangencies. However, the proofs are extremely technical. Could they be simplified? Current attempts like [266] center on the introduction of a simpler notion of critical points, possibly a non-inductive one [212].

Can this Hénon theory be extended to the weakly dissipative situation? to the conservative situation (for which the *standard map* is a well-known example defying analysis)? In the strongly dissipative setting, what are the typical phenomena on the complement of the Benedicks–Carleson set of parameters?

From a global perspective one of the main questions is the following:

> *Can infinitely many sinks coexist for a large set of parameters in a typical family or is Newhouse phenomenon atypical in the Kolmogorov or prevalent sense?*

This seems rather unlikely (see however [11]).

Away from such "critical dynamics", there are many results about systems with dominated splitting satisfying additional conditions. Can these conditions be weakened so they would be satisfied by typical systems satisfying some natural conditions (like robust transitivity)? For instance:

> *Could one analyze the physical measures of volume-hyperbolic systems?*

A more specific question is whether Tsujii's striking analysis of surface maps with one uniformly expanding direction can be extended to higher dimensions? can one weaken the uniformity of the expansion? The same questions for the corresponding invertible situation is considered in [83].

Maximum Entropy Measures and Topological Complexity

As we explained, C^∞ smoothness, by a Newhouse theorem, ensures the existence of maximum entropy measures, making the situation a little simpler than with respect to physical measures. This existence results allow in particular an easy formulation of the problem of the typicality of hyperbolicity:

> *Are maximum entropy ergodic measures of systems with positive entropy hyperbolic for most systems?*

A more difficult problem is that of the finite multiplicity of the maximum entropy measures. For instance:

> *Do typical systems possess finitely many maximum entropy ergodic measures?*

More specifically, can one prove intrinsic ergodicity (ie, uniqueness of the measure of maximum entropy) for an isolated homoclinic class of some diffeomorphisms (perhaps C^1-generic)? Can a generic C^1-diffeomorphism carry an infinite number of homoclinic classes, each with topological entropy bounded away from zero?

A perhaps more tractable question, given the recent progress in this area: Is a C^1-generic partially hyperbolic diffeomorphisms, perhaps with central dimension 1 or 2, intrinsically ergodic?

We have seen how uniform systems have simple symbolic dynamics, i. e., subshifts of finite type, and how interval maps and more generally entropy-expanding maps keep some of this simplicity, defining subshifts or puzzles of quasi-finite type [68,69]. [264] have defined symbolic dynamics for topological Hénon-like map which seems close to that of that of a one-dimensional system.

> *Can one describe Wang and Young symbolic dynamics of Hénon-like attractors and fit it in a class in which uniqueness of the maximum entropy measure could be proved?*

More generally, can one define nice combinatorial descriptions, for surface diffeomorphisms? Can one formulate variants of the entropy-expansion condition [*For instance building on our "entropy-hyperbolicity".*] of [66,70], that would be satisfied by a large subset of the diffeomorphisms?

Another possible approach is illustrated by the pruning front conjecture of [87] (see also [88,144]). It is an attempt to build a combinatorial description by trying to generalize the way that, for interval maps, kneading invariants determine the symbolic dynamics by considering the bifurcations from a trivial dynamics to an arbitrary one.

We hope that our reader has shared in our fascination with this subject, the many surprising and even paradoxical discoveries that have been made and the exciting current progress, despite the very real difficulties both in the analysis of such non-uniform systems as the Henon map and in the attemps to establish a general (and practical) ergodic theory of chaotic dynamical systems.

Acknowledgments

I am grateful for the advice and/or comments of the following colleagues: P. Collet, J.-R. Chazottes, and especially S. Ruette. I am also indebted to the anonymous referee.

Bibliography

Primary Literature

1. Aaronson J, Denker M (2001) Local limit theorems for partial sums of stationary sequences generated by Gibbs–Markov maps. Stoch Dyn 1:193–237
2. Abdenur F, Bonatti C, Crovisier S (2006) Global dominated splittings and the C^1 Newhouse phenomenon. Proc Amer Math Soc 134(8):2229–2237
3. Abraham R, Smale S (1970) Nongenericity of Ω-stability. In: Global Analysis, vol XIV. Proc Sympos Pure Math. Amer Math Soc
4. Alves JF (2000) SRB measures for non-hyperbolic systems with multidimensional expansion. Ann Sci Ecole Norm Sup 33(4):1–32
5. Alves JF (2006) A survey of recent results on some statistical features of non-uniformly expanding maps. Discret Contin Dyn Syst 15:1–20
6. Alves JF, Araujo V (2003) Random perturbations of nonuniformly expanding maps. Geometric methods in dynamics. I Astérisque 286:25–62
7. Alves JF, Viana M (2002) Statistical stability for robust classes of maps with non-uniform expansion. Ergod Theory Dynam Syst 22:1–32
8. Alves JF, Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly expanding. Invent Math 140:351–398
9. Anantharaman N (2004) On the zero-temperature or vanishing viscosity limit for certain Markov processes arising from Lagrangian dynamics. J Eur Math Soc (JEMS) 6:207–276
10. Anantharaman N, Nonnenmacher S (2007) Half delocalization of the eigenfunctions for the Laplacian on an Anosov manifold. Ann Inst Fourier 57:2465–2523
11. Araujo V (2001) Infinitely many stochastically stable attractors. Nonlinearity 14:583–596
12. Araujo V, Pacifico MJ (2006) Large deviations for non-uniformly expanding maps. J Stat Phys 125:415–457
13. Araujo V, Tahzibi A (2005) Stochastic stability at the boundary of expanding maps. Nonlinearity 18:939–958
14. Arnaud MC (1998) Le "closing lemma" en topologie C^1. Mem Soc Math Fr (NS) 74
15. Arnol'd VI (1988) Geometrical methods in the theory of ordinary differential equations, 2nd edn. Grundlehren der Mathematischen Wissenschaften, 250. Springer, New York
16. Arnol'd VI, Avez A (1968) Ergodic problems of classical mechanics. W.A. Benjamin, New York
17. Arnold L (1998) Random dynamical systems. Springer Monographs in Mathematics. Springer, Berlin
18. Avila A (2007) personal communication
19. Avila A, Moreira CG (2005) Statistical properties of unimodal maps: the quadratic family. Ann of Math 161(2):831–881
20. Avila A, Lyubich M, de Melo W (2003) Regular or stochastic dynamics in real analytic families of unimodal maps. Invent Math 154:451–550
21. Baladi V (2000) Positive transfer operators and decay of correlations. In: Advanced Series in Nonlinear Dynamics, 16. World Scientific Publishing Co, Inc, River Edge, NJ
22. Baladi V, Gouëzel S (2008) Good Banach spaces for piecewise hyperbolic maps via interpolation. preprint arXiv:0711.1960 available from http://www.arxiv.org
23. Baladi V, Ruelle D (1996) Sharp determinants. Invent Math 123:553–574
24. Baladi V, Tsujii M (2007) Anisotropic Hölder and Sobolev spaces for hyperbolic diffeomorphisms. Ann Inst Fourier (Grenoble) 57:127–154
25. Baladi V, Young LS (1993) On the spectra of randomly perturbed expanding maps. Comm Math Phys 156:355–385. Erratum: Comm Math Phys 166(1):219–220, 1994
26. Baladi V, Kondah A, Schmitt B (1996) Random correlations for small perturbations of expanding maps, English summary. Random Comput Dynam 4:179–204
27. Baraviera AT, Bonatti C (2003) Removing zero Lyapunov exponents, English summary. Ergod Theory Dynam Syst 23:1655–1670
28. Barreira L, Schmeling J (2000) Sets of "non-typical" points have full topological entropy and full Hausdorff dimension. Isr J Math 116:29–70
29. Barreira L, Pesin Ya, Schmeling J (1999) Dimension and product structure of hyperbolic measures. Ann of Math 149(2):755–783
30. Benedicks M, Carleson L (1985) On iterations of $1 - ax^2$ on $(-1, 1)$. Ann of Math (2) 122(1):1–25
31. Benedicks M, Carleson L (1991) The dynamics of the Hénon map. Ann of Math 133(2):73–169

32. Benedicks M, Viana M (2001) Solution of the basin problem for Hénon-like attractors. Invent Math 143:375–434
33. Benedicks M, Viana M (2006) Random perturbations and statistical properties of Hénon-like maps. Ann Inst H Poincaré Anal Non Linéaire 23:713–752
34. Benedicks M, Young LS (1993) Sinaĭ–Bowen–Ruelle measures for certain Hénon maps. Invent Math 112:541–576
35. Benedicks M, Young LS (2000) Markov extensions and decay of correlations for certain Hénon maps. Géométrie complexe et systèmes dynamiques, Orsay, 1995. Astérisque 261: 13–56
36. Bergelson V (2006) Ergodic Ramsey theory: a dynamical approach to static theorems. In: International Congress of Mathematicians, vol II, Eur Math Soc, Zürich, pp 1655–1678
37. Berkes I, Csáki E (2001) A universal result in almost sure central limit theory. Stoch Process Appl 94(1):105–134
38. Bishop E (1967/1968) A constructive ergodic theorem. J Math Mech 17:631–639
39. Blanchard F, Glasner E, Kolyada S, Maass A (2002) On Li-Yorke pairs, English summary. J Reine Angew Math 547:51–68
40. Blank M (1989) Small perturbations of chaotic dynamical systems. (Russian) Uspekhi Mat Nauk 44:3–28, 203. Translation in: Russ Math Surv 44:1–33
41. Blank M (1997) Discreteness and continuity in problems of chaotic dynamics. In: Translations of Mathematical Monographs, 161. American Mathematical Society, Providence
42. Blank M, Keller G (1997) Stochastic stability versus localization in one-dimensional chaotic dynamical systems. Nonlinearity 10:81–107
43. Blank M, Keller G, Liverani C (2002) Ruelle–Perron–Frobenius spectrum for Anosov maps. Nonlinearity 15:1905–1973
44. Bochi J (2002) Genericity of zero Lyapunov exponents. Ergod Theory Dynam Syst 22:1667–1696
45. Bochi J, Viana M (2005) The Lyapunov exponents of generic volume-preserving and symplectic maps. Ann of Math 161(2):1423–1485
46. Bolsinov AV, Taimanov IA (2000) Integrable geodesic flows with positive topological entropy. Invent Math 140:639–650
47. Bonano C, Collet P (2006) Complexity for extended dynamical systems. arXiv:math/0609681
48. Bonatti C, Crovisier S (2004) Recurrence et genericite. Invent Math 158(1):33–104
49. Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly contracting. Isr J Math 115:157–193
50. Bonatti C, Viana M (2004) Lyapunov exponents with multiplicity 1 for deterministic products of matrices. Ergod Theory Dynam Syst 24(5):1295–1330
51. Bonatti C, Diaz L, Viana M (2005) Dynamics beyond uniform hyperbolicity. A global geometric and probabilistic perspective. In: Mathematical Physics, III. Encyclopedia of Mathematical Sciences, 102. Springer, Berlin
52. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. In: Lecture Notes in Mathematics, vol 470. Springer, Berlin–New York
53. Bowen R (1979) Hausdorff dimension of quasi-circles. Inst Hautes Études Sci Publ Math 50:11–25
54. Bowen R, Ruelle D (1975) Ergodic theory of Axiom A flows. Invent Math 29:181–202
55. Boyle M, Downarowicz T (2004) The entropy theory of symbolic extensions. Invent Math 156:119–161
56. Boyle M, Fiebig D, Fiebig U (2002) Residual entropy, conditional entropy and subshift covers. Forum Math 14:713–757
57. Boyle M, Buzzi J, Gomez R (2006) Ricardo Almost isomorphism for countable state Markov shifts. J Reine Angew Math 592:23–47
58. Brain M, Berger A (2001) Chaos and chance. In: An introduction to stochastic aspects of dynamics. Walter de Gruyter, Berlin
59. Brin M (2001) Appendix A. In: Barreira L, Pesin Y (eds) Lectures on Lyapunov exponents and smooth ergodic theory. Proc Sympos Pure Math, 69, Smooth ergodic theory and its applications, Seattle, WA, 1999, pp 3–106. Amer Math Soc, Providence, RI
60. Brin M, Stuck G (2002) Introduction to dynamical systems. Cambridge University Press, Cambridge
61. Broise A (1996) Transformations dilatantes de l'intervalle et theoremes limites. Etudes spectrales d'operateurs de transfert et applications. Asterisque 238:1–109
62. Broise A (1996) Transformations dilatantes de l'intervalle et théorèmes limites. Astérisque 238:1–109
63. Bruin H, Keller G, Nowicki T, van Strien S (1996) Wild Cantor attractors exist. Ann of Math 143(2):97–130
64. Buzzi J (1997) Intrinsic ergodicity of smooth interval maps. Isr J Math 100:125–161
65. Buzzi J (2000) Absolutely continuous invariant probability measures for arbitrary expanding piecewise $\boldsymbol{R}$-analytic mappings of the plane. Ergod Theory Dynam Syst 20:697–708
66. Buzzi J (2000) On entropy-expanding maps. Unpublished
67. Buzzi J (2001) No or infinitely many a.c.i.p. for piecewise expanding C^r maps in higher dimensions. Comm Math Phys 222:495–501
68. Buzzi J (2005) Subshifts of quasi-finite type. Invent Math 159:369–406
69. Buzzi J (2006) Puzzles of Quasi-Finite Type, Zeta Functions and Symbolic Dynamics for Multi-Dimensional Maps. arXiv:math/0610911
70. Buzzi J () Hyperbolicity through Entropies. Saint-Flour Summer probability school lecture notes (to appear)
71. Bálint P, Gouëzel S (2006) Limit theorems in the stadium billiard. Comm Math Phys 263:461–512
72. Carleson L, Gamelin TW (1993) Complex dynamics. In: Universitext: Tracts in Mathematics. Springer, New York
73. Chazottes JR, Gouëzel S (2007) On almost-sure versions of classical limit theorems for dynamical systems. Probab Theory Relat Fields 138:195–234
74. Chernov N (1999) Statistical properties of piecewise smooth hyperbolic systems in high dimensions. Discret Contin Dynam Syst 5(2):425–448
75. Chernov N, Markarian R, Troubetzkoy S (2000) Invariant measures for Anosov maps with small holes. Ergod Theory Dynam Syst 20:1007–1044
76. Christensen JRR (1972) On sets of Haar measure zero in abelian Polish groups. Isr J Math 13:255–260
77. Collet P (1996) Some ergodic properties of maps of the interval. In: Dynamical systems, Temuco, 1991/1992. Travaux en Cours, 52. Hermann, Paris, pp 55–91
78. Collet P, Eckmann JP (2006) Concepts and results in chaotic dynamics: a short course. Theoretical and Mathematical Physics. Springer, Berlin
79. Collet P, Galves A (1995) Asymptotic distribution of entrance times for expanding maps of the interval. Dynamical systems

and applications, pp 139–152. World Sci Ser Appl Anal 4, World Sci Publ, River Edge, NJ

80. Collet P, Courbage M, Mertens S, Neishtadt A, Zaslavsky G (eds) (2005) Chaotic dynamics and transport in classical and quantum systems. Proceedings of the International Summer School of the NATO Advanced Study Institute held in Cargese, August 18–30, 2003. Edited by NATO Science Series II: Mathematics, Physics and Chemistry, 182. Kluwer, Dordrecht
81. Contreras G, Lopes AO, Thieullen Ph (2001) Lyapunov minimizing measures for expanding maps of the circle, English summary. Ergod Theory Dynam Syst 21:1379–1409
82. Cowieson WJ (2002) Absolutely continuous invariant measures for most piecewise smooth expanding maps. Ergod Theory Dynam Syst 22:1061–1078
83. Cowieson WJ, Young LS (2005) SRB measures as zero-noise limits. Ergod Theory Dynam Syst 25:1115–1138
84. Crovisier S (2006) Birth of homoclinic intersections: a model for the central dynamics of partially hyperbolic systems. http://www.arxiv.org/abs/math/0605387
85. Crovisier S (2006) Periodic orbits and chain-transitive sets of C^1-diffeomorphisms. Publ Math Inst Hautes Etudes Sci 104:87–141
86. Crovisier S (2006) Perturbation of C^1-diffeomorphisms and generic conservative dynamics on surfaces. Dynamique des diffeomorphismes conservatifs des surfaces: un point de vue topologique, 21, Panor Syntheses. Soc Math France, Paris, pp 1–33
87. Cvitanović P, Gunaratne GH, Procaccia I (1988) Topological and metric properties of Henon-type strange attractors. Phys Rev A 38(3):1503–1520
88. de Carvalho A, Hall T (2002) How to prune a horseshoe. Nonlinearity 15:R19-R68
89. de Castro A (2002) Backward inducing and exponential decay of correlations for partially hyperbolic attractors. Isr J Math 130:29–75
90. de la Llave R (2001) A tutorial on KAM theory. In: Smooth ergodic theory and its applications, Seattle, WA, 1999, pp 175–292. Proc Sympos Pure Math, 69. Amer Math Soc, Providence
91. de Melo W, van Strien S (1993) One-dimensional dynamics. In: Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 25. Springer, Berlin
92. Denker M, Philipp W (1984) Approximation by Brownian motion for Gibbs measures and flows under a function. Ergod Theory Dyn Syst 4:541–552
93. Detmers MF, Liverani C (2007) Stability of Statistical Properties in Two-dimensional Piecewise Hyperbolic Maps. Trans Amer Math Soc 360:4777–4814
94. Devaney RL (1989) An introduction to chaotic dynamical systems, 2nd ed. Addison–Wesley, Redwood
95. Diaz L, Rocha J, Viana M (1996) Strange attractors in saddle-node cycles: prevalence and globality. Invent Math 125: 37–74
96. Dolgopyat D (2000) On dynamics of mostly contracting diffeomorphisms. Comm Math Phys 213:181–201
97. Dolgopyat D (2002) On mixing properties of compact group extensions of hyperbolic systems. Isr J Math 130:157–205
98. Dolgopyat D (2004) Limit theorems for partially hyperbolic systems. Trans Amer Math Soc 356:1637–1689
99. Dolgopyat D (2004) On differentiability of SRB states for partially hyperbolic systems. Invent Math 155:389–449
100. Downarowicz T (2005) Entropy structure. J Anal Math 96: 57–116
101. Downarowicz T, Newhouse S (2005) Symbolic extensions and smooth dynamical systems. Invent Math 160:453–499
102. Downarowicz T, Serafin J (2003) Possible entropy functions. Isr J Math 135:221–250
103. Duhem P (1906) La théorie physique, son objet et sa structure. Vrin, Paris 1981
104. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
105. Eskin A, McMullen C (1993) Mixing, counting, and equidistribution in Lie groups. Duke Math J 71:181–209
106. Fiedler B (ed) (2001) Ergodic theory, analysis, and efficient simulation of dynamical systems. Springer, Berlin
107. Fiedler B (ed) (2002) Handbook of dynamical systems, vol 2. North-Holland, Amsterdam
108. Fisher A, Lopes AO (2001) Exact bounds for the polynomial decay of correlation, $1/f$ noise and the CLT for the equilibrium state of a non-Hölder potential. Nonlinearity 14:1071–1104
109. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. In: MB Porter Lectures. Princeton University Press, Princeton, NJ
110. Gallavotti G (1999) Statistical mechanics. In: A short treatise. Texts and Monographs in Physics. Springer, Berlin
111. Gatzouras D, Peres Y (1997) Invariant measures of full dimension for some expanding maps. Ergod Theory Dynam Syst 17:147–167
112. Glasner E, Weiss B (1993) Sensitive dependence on initial conditions. Nonlinearity 6:1067–1075
113. Gordin MI (1969) The central limit theorem for stationary processes. (Russian) Dokl Akad Nauk SSSR 188:739–741
114. Gouëzel S (2004) Central limit theorem and stable laws for intermittent maps. Probab Theory Relat Fields 128:82–122
115. Gouëzel S (2005) Berry–Esseen theorem and local limit theorem for non uniformly expanding maps. Ann Inst H Poincaré 41:997–1024
116. Gouëzel S, Liverani C (2006) Banach spaces adapted to Anosov systems. Ergod Theory Dyn Syst 26:189–217
117. Graczyk J, Świątek G (1997) Generic hyperbolicity in the logistic family. Ann of Math 146(2):1–52
118. Gromov M (1987) Entropy, homology and semialgebraic geometry. Séminaire Bourbaki, vol 1985/86. Astérisque No 145–146:225–240
119. Guivarc'h Y, Hardy J (1998) Théorèmes limites pour une classe de chaînes de Markov et applications aux difféomorphismes d'Anosov. Ann Inst H Poincaré 24:73:98
120. Guckenheimer J (1979) Sensitive dependence on initial conditions for one-dimensional maps. Commun Math Phys 70:133–160
121. Guckenheimer J, Holmes P (1990) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Revised and corrected reprint of the 1983 original. In: Applied Mathematical Sciences, 42. Springer, New York
122. Guckenheimer J, Wechselberger M, Young LS (2006) Chaotic attractors of relaxation oscillators. Nonlinearity 19:701–720
123. Gurevic BM (1996) Stably recurrent nonnegative matrices. (Russian) Uspekhi Mat Nauk 51(3)(309):195–196; translation in: Russ Math Surv 51(3):551–552
124. Gurevic BM, Savchenko SV (1998) Thermodynamic formalism for symbolic Markov chains with a countable number of

states. (Russian) Uspekhi Mat Nauk 53(2)(320):3–106; translation in: Russ Math Surv 53(2):245–344

125. Gutzwiller M (1990) Chaos in classical and quantum mechanics. In: Interdisciplinary Applied Mathematics, 1. Springer, New York
126. Hadamard J (1898) Les surfaces à courbures opposees et leurs lignes geodesiques. J Math Pures Appl 4:27–73
127. Hasselblatt B, Katok A (2003) A first course in dynamics. In: With a panorama of recent developments. Cambridge University Press, New York
128. Hasselblatt B, Katok A (eds) (2002,2006) Handbook of dynamical systems, vol 1A and 1B. Elsevier, Amsterdam
129. Hennion H (1993) Sur un théorème spectral et son application aux noyaux lipchitziens. Proc Amer Math Soc 118:627–634
130. Hénon M (1976) A two-dimensional mapping with a strange attractor. Comm Math Phys 50:69–77
131. Hesiod (1987) Theogony. Focus/R. Pullins, Newburyport
132. Hochman M (2006) Upcrossing inequalities for stationary sequences and applications. arXiv:math/0608311
133. Hofbauer F (1979) On intrinsic ergodicity of piecewise monotonic transformations with positive entropy. Isr J Math 34(3):213–237
134. Hofbauer F (1985) Periodic points for piecewise monotonic transformations. Ergod Theory Dynam Syst 5:237–256
135. Hofbauer F, Keller G (1982) Ergodic properties of invariant measures for piecewise monotonic transformations. Math Z 180:119–140
136. Hofbauer F, Keller G (1990) Quadratic maps without asymptotic measure. Comm Math Phys 127:319–337
137. Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser, Basel
138. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann of Math 161(2):397–488
139. Hua Y, Saghin R, Xia Z (2006) Topological Entropy and Partially Hyperbolic Diffeomorphisms. arXiv:math/0608720
140. Hunt FY (1998) Unique ergodicity and the approximation of attractors and their invariant measures using Ulam's method. (English summary) Nonlinearity 11:307–317
141. Hunt BR, Sauer T, Yorke JA (1992) Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces. Bull Amer Math Soc (NS) 27:217–238
142. Ibragimov I, Linnik Yu, Kingman JFC (ed) (trans) (1971) Independent and stationary sequences of random variables. Wolters-Noordhoff, Groningen
143. Ionescu Tulcea CT, Marinescu G (1950) Théorie ergodique pour des classes d'opérations non complètement continues. (French) Ann of Math 52(2):140–147
144. Ishii Y (1997) Towards a kneading theory for Lozi mappings. I. A solution of the pruning front conjecture and the first tangency problem. Nonlinearity 10:731–747
145. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Comm Math Phys 81:39–88
146. Jenkinson O (2007) Optimization and majorization of invariant measures. Electron Res Announc Amer Math Soc 13:1–12
147. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst Hautes Etudes Sci Publ Math No 51:137–173
148. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. In: Encyclopedia of Mathematics and its Applications, 54. With a supplementary chapter by: Katok A, Mendoza L. Cambridge University Press, Cambridge
149. Keller G (1984) On the rate of convergence to equilibrium in one-dimensional systems. Comm Math Phys 96:181–193
150. Keller G (1990) Exponents, attractors and Hopf decompositions for interval maps. Ergod Theory Dynam Syst 10:717–744
151. Keller G, Liverani C (1999) Stability of the spectrum for transfer operators. Ann Scuola Norm Sup Pisa Cl Sci 28(4):141–152
152. Keller G, Nowicki T (1992) Spectral theory, zeta functions and the distribution of periodic points for Collet-Eckmann maps. Comm Math Phys 149:31–69
153. Kifer Yu (1977) Small random perturbations of hyperbolic limit sets. (Russian) Uspehi Mat Nauk 32(1)(193):193–194
154. Kifer Y (1986) Ergodic theory of random transformations. In: Progress in Probability and Statistics, 10. Birkhäuser Boston, Inc, Boston, MA
155. Kifer Y (1988) Random perturbations of dynamical systems. In: Progress in Probability and Statistics, 16. Birkhäuser Boston, Inc, Boston, MA
156. Kifer Yu (1990) Large deviations in dynamical systems and stochastic processes. Trans Amer Math Soc 321:505–524
157. Kifer Y (1997) Computations in dynamical systems via random perturbations. (English summary) Discret Contin Dynam Syst 3:457–476
158. Kolyada SF (2004) LI-Yorke sensitivity and other concepts of chaos. Ukr Math J 56:1242–1257
159. Kozlovski OS (2003) Axiom A maps are dense in the space of unimodal maps in the C^k topology. Ann of Math 157(2):1–43
160. Kozlovski O, Shen W, van Strien S (2007) Density of Axiom A in dimension one. Ann Math 166:145–182
161. Krengel U (1983) Ergodic Theorems. De Gruyter, Berlin
162. Krzyżewski K, Szlenk W (1969) On invariant measures for expanding differentiable mappings. Studia Math 33:83–92
163. Lacroix Y (2002) Possible limit laws for entrance times of an ergodic aperiodic dynamical system. Isr J Math 132:253–263
164. Ladler RL, Marcus B (1979) Topological entropy and equivalence of dynamical systems. Mem Amer Math Soc 20(219)
165. Lastoa A, Yorke J (1973) On the existence of invariant measures for piecewise monotonic transformations. Trans Amer Math Soc 186:481–488
166. Ledrappier F (1984) Proprietes ergodiques des mesures de Sinaï. Inst Hautes Etudes Sci Publ Math 59:163–188
167. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. I Characterization of measures satisfying Pesin's entropy formula. II Relations between entropy, exponents and dimension. Ann of Math 122(2):509–539, 540–574
168. Leplaideur R (2004) Existence of SRB-measures for some topologically hyperbolic diffeomorphisms. Ergod Theory Dynam Syst 24:1199–1225
169. Li TY, Yorke JA (1975) Period three implies chaos. Amer Math Monthly 82:985–992
170. Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications. In: Graduate Texts in Computer Science, 2nd edn. Springer, New York
171. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
172. Liu PD, Qian M, Zhao Y (2003) Large deviations in Axiom A endomorphisms. Proc Roy Soc Edinb Sect A 133:1379–1388
173. Liverani C (1995) Decay of Correlations. Ann Math 142: 239–301

174. Liverani C (1995) Decay of Correlations in Piecewise Expanding maps. J Stat Phys 78:1111–1129
175. Liverani C (1996) Central limit theorem for deterministic systems. In: Ledrappier F, Lewowicz J, Newhouse S (eds) International conference on dynamical systems, Montevideo, 1995. Pitman Research Notes. In: Math 362:56–75
176. Liverani C (2001) Rigorous numerical investigation of the statistical properties of piecewise expanding maps – A feasibility study. Nonlinearity 14:463–490
177. Liverani C, Tsujii M (2006) Zeta functions and dynamical systems. (English summary) Nonlinearity 19:2467–2473
178. Lyubich M (1997) Dynamics of quadratic polynomials. I, II Acta Math 178:185–247, 247–297
179. Margulis G (2004) On some aspects of the theory of Anosov systems. In: With a survey by Richard Sharp: Periodic orbits of hyperbolic flows. Springer Monographs in Mathematics. Springer, Berlin
180. Mañé R (1982) An ergodic closing lemma. Ann of Math 116(2):503–540
181. Mañé R (1985) Hyperbolicity, sinks and measures in one-dimensional dynamics. Commun Math Phys 100:495–524
182. Mañé R (1988) A proof of the C^1 stability conjecture. Publ Math IHES 66:161–210
183. Melbourne I, Nicol M (2005) Almost sure invariance principle for nonuniformly hyperbolic systems. Comm Math Phys 260(1):131–146
184. Milnor J, Thurston W (1988) On iterated maps of the interval. Dynamical systems, College Park, MD, 1986–87, pp 465–563. Lecture Notes in Math, 1342. Springer, Berlin
185. Misiurewicz M (1976) A short proof of the variational principle for a Z^n_+ action on a compact space. Bull Acad Polon Sci Sér Sci Math Astronom Phys 24(12)1069–1075
186. Misiurewicz M (1976) Topological conditional entropy. Studia Math 55:175–200
187. Misiurewicz M (1979) Horseshoes for mappings of the interval. Bull Acad Polon Sci Sér Sci Math 27:167–169
188. Misiurewicz M (1995) Continuity of entropy revisited. In: Dynamical systems and applications, pp 495–503. World Sci Ser Appl Anal 4. World Sci Publ, River Edge, NJ
189. Misiurewicz M, Smítal J (1988) Smooth chaotic maps with zero topological entropy. Ergod Theory Dynam Syst 8:421–424
190. Mora L, Viana M (1993) Abundance of strange attractors. Acta Math 171:1–71
191. Moreira CG, Palis J, Viana M (2001) Homoclinic tangencies and fractal invariants in arbitrary dimension. CRAS 333:475–480
192. Murray JD (2002,2003) Mathematical biology. I and II An introduction, 3rd edn. In: Interdisciplinary Applied Mathematics, 17 and 18. Springer, New York
193. Nagaev SV (1957) Some limit theorems for stationary Markov chains. Theor Probab Appl 2:378–406
194. Newhouse SE (1974) Diffeomorphisms with infinitely many sinks. Topology 13:9–18
195. Newhouse SE (1989) Continuity properties of entropy. Ann of Math 129(2):215–235; Erratum: Ann of Math 131(2):409–410
196. Nussbaum RD (1970) The radius of the essential spectrum. Duke Math J 37:473–478
197. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
198. Ornstein D, Weiss B (1988) On the Bernoulli nature of systems with some hyperbolic structure. Ergod Theory Dynam Syst 18:441–456
199. Ovid (2005) Metamorphosis. W.W. Norton, New York
200. Palis J (2000) A global view of dynamics and a conjecture on the denseness of tinitude of attractors. Asterisque 261: 335–347
201. Palis J, Yoccoz JC (2001) Fers a cheval non-uniformement hyperboliques engendres par une bifurcation homocline et densite nulle des attracteurs. CRAS 333:867–871
202. Pesin YB (1976) Families of invariant manifolds corresponding to non-zero characteristic exponents. Math USSR Izv 10:1261–1302
203. Pesin YB (1977) Characteristic exponents and smooth ergodic theory. Russ Math Surv 324:55–114
204. Pesin Ya (1997) Dimension theory in dynamical systems. In: Contemporary views and applications. Chicago Lectures in Mathematics. University of Chicago Press, Chicago
205. Pesin Ya, Sinaĭ Ya (1982) Gibbs measures for partially hyperbolic attractors. Ergod Theory Dynam Syst 2:417–438
206. Piorek J (1985) On the generic chaos in dynamical systems. Univ Iagel Acta Math 25:293–298
207. Plykin RV (2002) On the problem of the topological classification of strange attractors of dynamical systems. Uspekhi Mat Nauk 57:123–166. Translation in: Russ Math Surv 57: 1163–1205
208. Poincare H (1892) Les methodes nouvelles de la mecanique céleste. Paris, Gauthier–Villars
209. Pollicott M, Sharp R (2002) Invariance principles for interval maps with an indifferent fixed point. Comm Math Phys 229: 337–346
210. Przytycki F (1977) On *U*-stability and structural stability of endomorphisms satisfying. Axiom A Studia Math 60:61–77
211. Pugh C, Shub M (1999) Ergodic attractors. Trans Amer Math Soc 312:1–54
212. Pujals E, Rodriguez–Hertz F (2007) Critical points for surface diffeomorphisms. J Mod Dyn 1:615–648
213. Puu T (2000) Attractors, bifurcations, and chaos. In: Nonlinear phenomena in economics. Springer, Berlin
214. Rees M (1981) A minimal positive entropy homeomorphism of the 2-torus. J London Math Soc 23(2):537–550
215. Robinson RC (2004) An introduction to dynamical systems: continuous and discrete. Pearson Prentice Hall, Upper Saddle River
216. Rousseau–Egele J (1983) Un théorème de la limite locale pour une classe de transformations dilatantes et monotones par morceaux. Ann Probab 11:772–788
217. Ruelle D (1968) Statistical mechanics of a one-dimensional lattice gas. Comm Math Phys 9:267–278
218. Ruelle D (1976) A measure associated with axiom-A attractors. Amer J Math 98:619–654
219. Ruelle D (1978) An inequality for the entropy of differentiable maps. Bol Soc Brasil Mat 9:83–87
220. Ruelle D (1982) Repellers for real analytic maps. Ergod Theory Dyn Syst 2:99–107
221. Ruelle D (1989) The thermodynamic formalism for expanding maps. Comm Math Phys 125:239–262
222. Ruelle D (2004) Thermodynamic formalism. In: The mathematical structures of equilibrium statistical mechanics, 2nd edn. Cambridge Mathematical Library, Cambridge University Press, Cambridge
223. Ruelle D (2005) Differentiating the absolutely continuous in-

variant measure of an interval map f with respect to f. Comm Math Phys 258:445–453

224. Ruette S (2003) On the Vere-Jones classification and existence of maximal measures for countable topological Markov chains. Pacific J Math 209:366–380
225. Ruette S (2003) Chaos on the interval. http://www.math.u-psud.fr/~ruette
226. Saari DG (2005) Collisions, rings, and other Newtonian *N*-body problems. In: CBMS Regional Conference Series in Mathematics, 104. Published for the Conference Board of the Mathematical Sciences, Washington, DC. American Mathematical Society, Providence
227. Saperstone SH, Yorke JA (1971) Controllability of linear oscillatory systems using positive controls. SIAM J Control 9: 253–262
228. Sarig O (1999) Thermodynamic formalism for countable Markov shifts. Ergod Theory Dynam Syst 19:1565–1593
229. Sarig O (2001) Phase Transitions for Countable Topological Markov Shifts. Commun Math Phys 217:555–577
230. Sataev EA (1992) Invariant measures for hyperbolic mappings with singularities. Uspekhi Mat Nauk 47:147–202. Translation in: Russ Math Surv 47:191–251
231. Saussol B (2000) Absolutely continuous invariant measures for multidimensional expanding maps. Isr J Math 116:223–48
232. Saussol B (2006) Recurrence rate in rapidly mixing dynamical systems. Discret Contin Dyn Syst 15(1):259–267
233. Shub M (1987) Global stability of dynamical systems. Springer, New York
234. Simon R (1972) A 3-dimensional Abraham-Smale example. Proc Amer Math Soc 34:629–630
235. Sinai Ya (1972) Gibbs measures in ergodic theory. Uspehi Mat Nauk 27(166):21–64
236. Starkov AN (2000) Dynamical systems on homogeneous spaces. In: Translations of Mathematical Monographs, 190. American Mathematical Society, Providence, RI
237. Stewart P (1964) Jacobellis v Ohio. US Rep 378:184
238. Szász D (ed) (2000) Hard ball systems and the Lorentz gas. Encyclopedia of Mathematical Sciences, 101. In: Mathematical Physics, II. Springer, Berlin
239. Tsujii M (1992) A measure on the space of smooth mappings and dynamical system theory. J Math Soc Jpn 44:415–425
240. Tsujii M (2000) Absolutely continuous invariant measures for piecewise real-analytic expanding maps on the plane. Comm Math Phys 208:605–622
241. Tsujii M (2000) Piecewise expanding maps on the plane with singular ergodic properties. Ergod Theory Dynam Syst 20:1851–1857
242. Tsujii M (2001) Absolutely continuous invariant measures for expanding piecewise linear maps. Invent Math 143:349–373
243. Tsujii M (2005) Physical measures for partially hyperbolic surface endomorphisms. Acta Math 194:37–132
244. Tucker W (1999) The Lorenz attractor exists. C R Acad Sci Paris Ser I Math 328:1197–1202
245. Vallée B (2006) Euclidean dynamics. Discret Contin Dyn Syst 15:281–352
246. van Strien S, Vargas E (2004) Real bounds, ergodicity and negative Schwarzian for multimodal maps. J Amer Math Soc 17: 749–782
247. Vasquez CH (2007) Statistical stability for diffeomorphisms with dominated splitting. Ergod Theory Dynam Syst 27: 253–283
248. Viana M (1993) Strange attractors in higher dimensions. Bol Soc Bras Mat (NS) 24:13–62
249. Viana M (1997) Multidimensional nonhyperbolic attractors. Inst Hautes Etudes Sci Publ Math No 85:63–96
250. Viana M (1997) Stochastic dynamics of deterministic systems. In: Lecture Notes 21st Braz Math Colloq IMPA. Rio de Janeiro
251. Viana M (1998) Dynamics: a probabilistic and geometric perspective. In: Proceedings of the International Congress of Mathematicians, vol I, Berlin, 1998. Doc Math Extra I:557–578
252. Wang Q, Young LS (2003) Strange attractors in periodically-kicked limit cycles and Hopf bifurcations. Comm Math Phys 240:509–529
253. Weiss B (2002) Single orbit dynamics. Amer Math Soc, Providence
254. Yomdin Y (1987) Volume growth and entropy. Isr J Math 57:285–300
255. Yoshihara KI (2004) Weakly dependent stochastic sequences and their applications. In: Recent Topics on Weak and Strong Limit Theorems, vol XIV. Sanseido Co Ltd, Chiyoda
256. Young LS (1982) Dimension, entropy and Lyapunov exponents. Ergod Th Dynam Syst 6:311–319
257. Young LS (1985) Bowen–Ruelle measures for certain piecewise hyperbolic maps. Trans Amer Math Soc 287:41–48
258. Young LS (1986) Stochastic stability of hyperbolic attractors. Ergod Theory Dynam Syst 6:311–319
259. Young LS (1990) Large deviations in dynamical systems. Trans Amer Math Soc 318:525–543
260. Young LS (1992) Decay of correlations for certain quadratic maps. Comm Math Phys 146:123–138
261. Young LS (1995) Ergodic theory of differentiable dynamical systems. Real and complex dynamical systems. Hillerad, 1993, pp 293–336. NATO Adv Sci Inst Ser C Math Phys Sci 464. Kluwer, Dordrecht
262. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 585–650
263. Young LS (1999) Recurrence times and rates of mixing. Isr J Math 110:153–188
264. Young LS, Wang D (2001) Strange attractors with one direction of instability. Comm Math Phys 218:1–97
265. Young LS, Wang D (2006) Nonuniformly expanding 1D maps. Commun Math Phys vol 264:255–282
266. Young LS, Wang D (2008) Toward a theory of rank one attractors. Ann Math 167:349–480
267. Zhang Y (1997) Dynamical upper bounds for Hausdorff dimension of invariant sets. Ergod Theory Dynam Syst 17: 739–756

Books and Reviews

Baladi V (2000) The magnet and the butterfly: thermodynamic formalism and the ergodic theory of chaotic dynamics. (English summary). In: Development of mathematics 1950–2000, Birkhäuser, Basel, pp 97–133

Bonatti C (2003) Dynamiques generiques: hyperbolicite et transitivite. In: Seminaire Bourbaki vol 2001/2002. Asterisque No 290:225–242

Kuksin SB (2006) Randomly forced nonlinear PDEs and statistical hydrodynamics in 2 space dimensions. In: Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich

Liu PD, Qian M (1995) Smooth ergodic theory of random dynamical systems. In: Lecture Notes in Mathematics, 1606. Springer, Berlin

Ornstein D, Weiss B (1991) Statistical properties of chaotic systems. With an appendix by David Fried. Bull Amer Math Soc (NS) 24:11–116

Young LS (1999) Ergodic theory of chaotic dynamical systems. XIth International Congress of Mathematical Physics, ICMP '97, Brisbane, Int Press, Cambridge, MA, pp 131–143

Chaotic Behavior of Cellular Automata

Julien Cervelle[1], Alberto Dennunzio[2], Enrico Formenti[3]

[1] Laboratoire d'Informatique de l'Institut Gaspard-Monge, Université Paris-Est, Marne la Vallée, France

[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy

[3] Laboratoire I3S, Université de Nice-Sophia Antipolis, Sophia Antipolis, France

Article Outline

Glossary

Equicontinuity All points are equicontinuity points (in compact settings).

Equicontinuity point A point for which the orbits of nearby points remain close.

Expansivity From two distinct points, orbits eventually separate.

Injectivity The next state function is injective.

Linear CA A CA with additive local rule.

Regularity The set of periodic points is dense.

Sensitivity to initial conditions For any point x there exist arbitrary close points whose orbits eventually separate from the orbit of x.

Strong transitivity There always exist points which eventually move from any arbitrary neighborhood to any point.

Surjectivity The next state function is surjective.

Topological mixing There always exist points which definitely move from any arbitrary neighborhood to any other.

Transitivity There always exist points which eventually move from any arbitrary neighborhood to any other.

Definition of the Subject

A discrete time dynamical system (DTDS) is a pair $\langle X, F \rangle$ where X is a set equipped with a distance d and $F\colon X \mapsto X$ is a mapping which is continuous on X with respect to the metric d. The set X and the function F are called the state space and the next state map. At the very beginning of the seventies, the notion of chaotic behavior for DTDS has been introduced in experimental physics [46]. Successively, mathematicians started investigating this new notion finding more and more complex examples. Although a general universally accepted theory of chaos has not emerged, at least some properties are recognized as basic components of possible chaotic behavior. Among them one can list: sensitivity to initial conditions, transitivity, mixing, expansively etc. [5,6,21,22,28,29,38,41,42,58,59,60].

In the eighties, S. Wolfram started studying some of these properties in the context of cellular automata (CA) [64]. These pioneering studies opened the way to a huge amount of successive paper which aimed to complete, precise and further develop the theory of chaos in the CA context (▶ Dynamics of Cellular Automata in Non–compact Spaces, ▶ Topological Dynamics of Cellular Automata, ▶ Ergodic Theory of Cellular Automata, [7,8,9,10,11,13,14,15,16,17,35,43,51]). This long quest has also been stimulated by the advent of more and more powerful computers which helped researchers in their investigations. However, remark that most of the properties involved in chaos definitions turned out to be undecidable [23,34,39,40,45]. Anyway, there are non-trivial classes of CA for which these properties are decidable. For this reason in the present work we focus on linear CA.

Introduction

Cellular automata are simple formal models for complex systems. They are used in many scientific fields ranging from biology to chemistry or from physics to computer science.

A CA is made of an infinite set of finite automata distributed over a regular lattice $\mathcal{L}$. All finite automata are identical. Each automaton assumes a value, chosen from a finite set A, called the *alphabet*. A *configuration* is a snapshot of the all automata values, i. e., a function from $\mathcal{L}$ to A.

In the present chapter, we consider the D dimensional lattice $\mathcal{L} = \mathbb{Z}^D$.

A *local rule* updates the value of an automaton on the basis of its current value and the ones of a fixed set of neighboring automata which are individuated by the *neighborhood frame* $N = \{\vec{u}_1, \dots, \vec{u}_s\} \subset \mathbb{Z}^D$. Formally, the local rule is a function $f \colon A^s \to A$.

All the automata of the lattice are updated synchronously in discrete time steps. In other words, the local rule f induces a *global rule* $F \colon A^{\mathbb{Z}^D} \to A^{\mathbb{Z}^D}$, describing the evolution of the whole system from a generic time $t \in \mathbb{N}$ to $t + 1$.

When one equips the configuration space with a metric, a CA can be viewed as a DTDS $\langle A^{\mathbb{Z}^D}, F \rangle$. The state space $A^{\mathbb{Z}^D}$ of a CA is also called the configuration space $A^{\mathbb{Z}^D}$. From now on, we identify a CA with the dynamical system induced by itself or with its global rule.

As we have already told, the studies on the chaotic behavior of CA have played a central role in the last twenty years. Unfortunately, most of the properties involved are undecidable. In this chapter, we illustrate these notions by means of a particular class of CA in which they turn out to be decidable constituting a valid source for examples and understanding. Indeed, we focus on *linear* CA, i. e., CA whose local rule is a linear function on a finite group G. For clarity's sake, we assume that G is the set $Z_m = \{0, 1, \dots, m-1\}$ of integers modulo m and $+$ the cell-wise addition.

Despite their simplicity, linear CA exhibit many of the complex features of general CA ranging from trivial to the most complicated behaviors.

Definitions

Given a DTDS $\langle X, g \rangle$ the next state function induces deterministic dynamics by its iterated application starting from a given initial state. Formally, for a fixed state $x \in X$, the *dynamical evolution* or *orbit of initial state* x is the sequence $\{F^n(x)\}_{n \in \mathbb{N}}$. A state $p \in X$ is a periodic point if there exists an integer $n > 0$ such that $F^n(p) = p$.

Deterministic Chaos

We are particularly interested in the properties which can be considered as components of "chaotic" behavior. Among them, sensitivity to initial conditions is the most intriguing, at least at a level of popular divulgement. It captures the feature that small errors in experimental measurements lead to large scale divergence in the evolution.

Definition 1 (Sensitivity) A DTDS $\langle X, F \rangle$ is *sensitive to the initial conditions* (or simply *sensitive*) if there exists a constant $\varepsilon > 0$ such that for any state $x \in X$ and any $\delta > 0$ there is a state $y \in X$ such that $d(y, x) < \delta$ and $d(F^n(y), F^n(x)) > \varepsilon$ for some $n \in \mathbb{N}$.

Intuitively, a DTDS is sensitive if for any state x there exist points arbitrarily close to x which eventually separate from x under iteration of F. Sensitivity is a strong form of instability. In fact, if a system is sensitive to the initial conditions and we are not able to measure with infinite precision the initial state, we cannot predict its dynamical evolution. This means that experimental or casual errors can lead to wrong results.

The following is another form of unpredictability.

Definition 2 (Positive Expansivity) A DTDS $\langle X, F \rangle$ is *positively expansive* if there exists a constant $\varepsilon > 0$ such that for any pair of distinct states $x, y \in X$ we have $d(F^n(y), F^n(x)) \geq \varepsilon$ for some $n \in \mathbb{N}$.

Remark that in perfect spaces (i. e., spaces without isolated points), expansive DTDS are necessarily sensitive to initial conditions.

Sensitivity alone, notwithstanding its intuitive appeal, has the drawback that, once taken as the unique condition for characterizing chaos, it appears to be neither sufficient nor as intuitive as it seems at a first glance.

Indeed, for physicists, a chaotic system must be necessarily nonlinear (the term linear in its classical meaning refers to the linearity of the equation which governs the behavior of a system). However, it is easy to find linear systems (on the reals, for instance) which are sensitive to initial conditions. Hence, a chaotic system has to satisfy further properties other than sensitivity.

Definition 3 (Transitivity) A DTDS $\langle X, F \rangle$ is (topologically) *transitive* if for all non empty open subsets U and V of X there exists a natural number n such that $F^n(U) \cap V \neq \emptyset$.

Intuitively, a transitive DTDS has points which eventually move under iteration of F from one arbitrarily small neighborhood to any other. As a consequence, the dynamical system cannot be decomposed into two disjoint clopen sets which are invariant under the iterations of F. Indecomposability is an important feature since, roughly speaking, it guarantees that the system behaves in the same way in the whole state space. Finally, note that transitivity implies surjectivity in the case of compact spaces.

Some DTDS may exhibit stronger forms of transitivity.

Definition 4 (Mixing) A DTDS $\langle X, F \rangle$ is *topologically mixing* if for all non-empty open subsets U, V of X there exists a natural number m such that for every $n \geq m$ it holds that $F^n(U) \cap V \neq \emptyset$.

The previous notion is the topological version of the well-known mixing property of ergodic theory. It is particularly useful when studying product of systems. Indeed, the product of two transitive systems is not necessarily transitive; it is transitive if one of the two systems is mixing [27].

Moreover, remark that any non-trivial (i. e. with at least two points) mixing system is sensitive to initial conditions [44].

Definition 5 (Strong Transitivity) A DTDS $\langle X, F\rangle$ is *strongly transitive* if for any nonempty open set $U \subseteq X$ we have that $\bigcup_{n=0}^{+\infty} F^n(U) = X$.

A strongly transitive map F has points which eventually move under iteration of F from one arbitrarily small neighborhood to any other *point*. As a consequence, a strongly transitive map is necessarily surjective.

Finally, remark that many well-known classes of transitive DTDS (irrational rotations of the unit circle, the tent map, the shift map, etc.) exhibit the following form of transitivity.

Definition 6 (Total Transitivity) A DTDS $\langle X, F\rangle$ is total transitive if for all integers $n > 0$, the system $\langle X, F^n\rangle$ is transitive.

Proposition 1 ([27]) *Any mixing DTDS is totally transitive.*

At the beginning of the eighties, Auslander and Yorke introduced the following definition of chaos [5].

Definition 7 (AY-Chaos) A DTDS is AY-chaotic if it is transitive and it is sensitive to the initial conditions.

This definition involves two fundamental characteristics: the undecomposability of the system, due to transitivity and the unpredictability of the dynamical evolution, due to sensitivity.

We now introduce a notion which is often referred to as an element of *regularity* a chaotic dynamical system must exhibit.

Definition 8 (DPO) A DTDS $\langle X, F\rangle$ has the *denseness of periodic points* (or, it is *regular*) if the set of its periodic points is dense in X.

The following is a standard result for compact DTDS.

Proposition 2 *If a compact DTDS has DPO then it is surjective.*

In his famous book [22], Devaney modified the AY-chaos adding the denseness of periodic points.

Definition 9 (D-Chaos) A DTDS is said to be D-chaotic if it is sensitive, transitive and regular.

An interesting result states that sensitivity, despite its popular appeal, is redundant in the Devaney definition of chaos.

Proposition 3 ([6]) *Any transitive and regular DTDS with an infinite number of states is sensitive to initial conditions.*

Note that neither transitivity nor DPO are redundant in the Devaney definition of chaos [37]. Further notions of chaos can be obtained by replacing transitivity or sensitivity with stronger properties (like expansively, strong transitivity, etc.).

Stability

All the previous properties can be considered as components of a chaotic, and then unstable, behavior for a DTDS. We now illustrate some properties concerning conditions of stability for a system.

Definition 10 (Equicontinuous Point) A state $x \in X$ of a DTDS $\langle X, F\rangle$ is an *equicontinuous point* if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $y \in X$, $d(y, x) < \delta$ implies that $\forall n \in \mathbb{N},\ d(F^n(y), F^n(x)) < \varepsilon$.

In other words, a point x is equicontinuous (or Lyapunov stable) if for any $\varepsilon > 0$, there exists a neighborhood of x whose states have orbits which stay close to the orbit of x with distance less than ε. This is a condition of local stability for the system.

Associated with this notion involving a single state, we have two notions of global stability based on the "size" of the set of the equicontinuous points.

Definition 11 (Equicontinuity) A DTDS $\langle X, F\rangle$ is said to be *equicontinuous* if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in X$, $d(y, x) < \delta$ implies that $\forall n \in \mathbb{N}$, $(F^n(y), F^n(x)) < \varepsilon$.

Given a DTDS, let E be its set of equicontinuity points. Remark that if a DTDS is equicontinuous then the set E of all its equicontinuity points is the whole X. The converse is also true in the compact settings. Furthermore, if a system is sensitive then $E = \emptyset$. In general, the converse is not true [43].

Definition 12 (Almost Equicontinuity) A DTDS is *almost equicontinuous* if the set of its equicontinuous points E is residual (i. e., it can be obtained by a infinite intersection of dense open subsets).

It is obvious that equicontinuous systems are almost equicontinuous. In the sequel, almost equicontinuous systems which are not equicontinuous will be called strictly almost equicontinuous. An important result affirms that transitive systems on compact spaces are almost equicontinuous if and only if they are not sensitive [3].

Topological Entropy

Topological entropy is another interesting property which can be taken into account in order to study the degree of chaoticity of a system. It was introduced in [2] as an invariant of topological conjugacy. The notion of topological entropy is based on the complexity of the coverings of the systems. Recall that an open covering of a topological space X is a family of open sets whose union is X. The join of two open coverings $\mathcal{U}$ and $\mathcal{V}$ is $\mathcal{U} \vee \mathcal{V} = \{U \cap V : U \in \mathcal{U}, V \in \mathcal{V}\}$. The inverse image of an open covering $\mathcal{U}$ by a map $F: X \mapsto X$ is $F^{-1}(\mathcal{U}) = \{F^{-1}(U) : U \in \mathcal{U}\}$. On the basis of these previous notions, the entropy of a system $\langle X, F\rangle$ over an open covering $\mathcal{U}$ is defined as

$$\begin{aligned} &H(X,F,\mathcal{U}) \\ &\quad = \lim_{n\to\infty} \frac{\log|\mathcal{U} \vee F^{-1}(\mathcal{U}) \cdots \vee F^{-(n-1)}(\mathcal{U})|}{n}\,, \end{aligned}$$

where $|U|$ is the cardinality of U.

Definition 13 (Topological Entropy) The topological entropy of a DTDS $\langle X, F\rangle$ is

$$\begin{aligned} &h(X,F) \\ &\quad = \sup\{H(X,F,\mathcal{U}) : \mathcal{U} \text{ is an open covering of } X\} \end{aligned} \tag{1}$$

Topological entropy represents the exponential growth of the number of orbit segments which can be distinguished with a certain good, but finite, accuracy. In other words, it measures the uncertainty of the system evolutions when a partial knowledge of the initial state is given.

There are close relationships between the entropy and the topological properties we have seen so far. For instance, we have the following.

Proposition 4 ([3,12,28]) *In compact DTDS, transitivity and positive entropy imply sensitivity.*

Cellular Automata

Consider the set of configurations C which consists of all functions from $\mathbb{Z}^D$ into A. The space C is usually equipped with the Thychonoff (or Cantor) metric d defined as

$$\forall a, b \in C, \quad d(a,b) = 2^{-n}, \qquad \text{with} \quad n = \min_{\vec{v}\in\mathbb{Z}^D} \left\{\|\vec{v}\|_\infty : a(\vec{v}) \neq b(\vec{v})\right\},$$

where $\|\vec{v}\|_\infty$ denotes the maximum of the absolute value of the components of $\vec{v}$. The topology induced by d coincides with the product topology induced by the discrete topology on A. With this topology, C is a compact, perfect and totally disconnected space.

Let $N = \{\vec{u}_1, \ldots, \vec{u}_s\}$ be an ordered set of vectors of $\mathbb{Z}^D$ and $f: A^s \mapsto A$ be a function.

Definition 14 (CA) The D-dimensional CA based on the local rule f and the neighborhood frame N is the pair $\langle C\rangle, F$ where $F: C \mapsto C$ is the global transition rule defined as follows:

$$\begin{aligned} &\forall c \in C, \quad \forall \vec{v} \in \mathbb{Z}^D, \\ &\qquad F(c)(\vec{v}) = f(c(\vec{v}+\vec{u}_1), \ldots, c(\vec{v}+\vec{u}_s))\,. \end{aligned} \tag{2}$$

Note that the mapping F is (uniformly) continuous with respect to the Thychonoff metric. Hence, the pair $\langle C\rangle, F$ is a proper discrete time dynamical system.

Let $Z_m = \{0, 1, \ldots, m-1\}$ be the group of the integers modulo m. Denote by C_m the configuration space C for the special case $A = Z_m$. When C_m is equipped with the natural extensions of the sum and the product operations, it turns out to be a linear space. Therefore, one can exploit the properties of linear spaces to simplify the proofs and the overall presentation.

A function $f: Z_m^s \mapsto Z_m$ is said to be linear if there exist $\lambda_1, \ldots, \lambda_s \in Z_m$ such that it can be expressed as:

$$\forall (x_1, \ldots, x_s) \in Z_m^s, \quad f(x_1, \ldots, x_s) = \left[\sum_{i=1}^{s} \lambda_i x_i\right]_m$$

where $[x]_m$ is the integer x taken modulo m.

Definition 15 (Linear CA) A D-dimensional linear CA is a CA $\langle C\rangle_m, F$ whose local rule f is linear.

Note that for the linear CA Equ. (2) becomes:

$$\forall c \in C, \ \forall \vec{v} \in \mathbb{Z}^D, \ F(c)(\vec{v}) = \left[\sum_{i=1}^{s} \lambda_i c(\vec{v}+\vec{u}_i)\right]_m .$$

The Case of Cellular Automata

In this section the results seen so far are specialized to the CA setting, focusing on dimension one. The following result allows a first classification of one-dimensional CA according to their degree of chaoticity.

Theorem 1 ([43]) *A one-dimensional CA is sensitive if and only if it is not almost equicontinuous.*

In other words, for CA the dichotomy between sensitivity and almost equicontinuity is true and not only under the transitivity condition. As a consequence, the family of all one-dimensional CA can be partitioned into four classes [43]:

(k1) equicontinuous CA;
(k2) strictly almost equicontinuous CA;
(k3) sensitive CA;
(k4) expansive CA.

This classification almost fits for higher dimensions. The problem is that there exist CA between the classes $K2$ and $K3$ (i. e., non sensitive CA without any equicontinuity point). Even relaxing $K2$ definition to "CA having some equicontinuity point", the gap persists (see, for instance [57]).

Unfortunately, much like most of the interesting properties of CA, the properties defining the above classification scheme are also affected by undecidability.

Theorem 2 ([23]) *For each $i = 1, 2, 3$, there is no algorithm to decide if a one-dimensional CA belongs to the class Ki.*

The following conjecture stresses the fact that nothing is known about the decidability for the membership in $K4$.

Conjecture 1 (Folklore) *Membership in class $K4$ is undecidable.*

Remark that the above conjecture is clearly false for dimensions greater than 1 since there do not exist expansive CA for dimension strictly greater than 1 [53].

Proposition 5 ([7,11]) *Expansive CA are strongly transitive and mixing.*

In CA settings, the notion of total transitivity reduces to the simple transitivity. Moreover, there is a strict relation between transitive and sensitive CA.

Theorem 3 ([49]) *If a CA is transitive then it is totally transitive.*

Theorem 4 ([28]) *Transitive CA are sensitive.*

As we have already seen, sensitivity is undecidable. Hence, in view of the combinatorial complexity of transitive CA, the following conjectures sound true.

Conjecture 2 *Transitivity is an undecidable property.*

Conjecture 3 [48] *Strongly transitive CA are (topologically) mixing.*

Chaos and Combinatorial Properties

In this section, when referring to a one-dimensional CA, we assume that $u_1 = \min N$ and $u_s = \max N$ (see also Sect. "Definitions"). Furthermore, we call *elementary* a one-dimensional CA with alphabet $A = \{0, 1\}$ and $N = \{-1, 0, 1\}$ (there exist 256 possible elementary CA which can be enumerated according to their local rule [64]).

In CA settings, most of the chaos components are related to some properties of combinatorial nature like injectivity, surjectivity and openness.

First of all, remark that injectivity and surjectivity are dimension sensitive properties in the sense of the following.

Theorem 5 ([4,39]) *Injectivity and surjectivity are decidable in dimension 1, while they are not decidable in dimension greater than 1.*

A one-dimensional CA is said to be a *right CA* (resp., *left CA*) if $u_1 > 0$ (resp., $u_s < 0$).

Theorem 6 ([1]) *Any surjective and right (or left) CA is topologically mixing.*

The previous result can be generalized in dimension greater than 1 in the following sense.

Theorem 7 *If for a given surjective D-dimensional CA there exists a $(D-1)$-dimensional hyperplane H (as a linear subspace of $\mathbb{Z}^D$) such that:*

1. *all the neighbor vectors stay on the same side of a H, and*
2. *no vectors lay on H,*

then the CA is topologically mixing.

Proof Choose two configurations c and d and a natural number r. Let U and V be the two distinct open balls of radius 2^{-r} and of center c and d, respectively (in a metric space (X, d) the open ball of radius $\delta > 0$ and center $x \in X$ is the set $B_\delta(x) = \{y \in X \mid d(y, x) < \delta\}$). For any integer $n > 1$, denote by N_n the neighbor frame of the CA F^n and with $d_n \in F^{-n}(d)$ any n-preimage of d. The values $c(\vec{x})$ for $\vec{x} \in O$ depend only on the values $d_n(\vec{x})$ for $\vec{x} \in O + N_n$, where $O = \{\vec{v} \mid \|\vec{v}\|_\infty \leq r\}$. By the hypothesis, there exists an integer $m > 0$ such that for any $n \geq m$ the sets O and $O + N_n$ are disjoint. Therefore, for any $n \geq m$ build a configuration $e_n \in C$ such that $e_n(\vec{x}) = d(\vec{x})$ for $\vec{x} \in O$, and $e_n(\vec{x}) = d_n(\vec{x})$ for $\vec{x} \in O + N_n$. Then, for any $n \geq m$, $e_n \in V$ and $F^n(e_n) \in U$. □

Injectivity prevents a CA from being strong transitive as stated in the following

Theorem 8 ([11]) *Any strongly transitive CA is not injective.*

Recall that a CA of global rule F is open if F is an open function. Equivalently, in the one-dimensional case, every configuration has the same numbers of predecessors [33].

Theorem 9 ([55]) *Openness is decidable in dimension one.*

Remark that mixing CA are not necessarily open (consider, for instance, the elementary rule 106, see [44]). The following conjecture is true when replacing strong transitivity by expansively [43].

Conjecture 4 *Strongly transitive CA are open.*

Recall that the shift map $\sigma : A^{\mathbb{Z}} \mapsto A^{\mathbb{Z}}$ is the one-dimensional linear CA defined by the neighborhood $N = \{+1\}$ and by the coefficient $\lambda_1 = 1$. A configuration of a one-dimensional CA is called jointly periodic if it is periodic both for the CA and the shift map (i. e., it is also spatially periodic). A CA is said to have the joint denseness of periodic orbits property (JDPO) if it admits a dense set of jointly periodic configurations. Obviously, JDPO is a stronger form of DPO.

Theorem 10 ([13]) *Open CA have JDPO.*

The common feeling is that (J)DPO is a property of a class wider than open CA. Indeed,

Conjecture 5 [8] *Every surjective CA has (J)DPO.*

If this conjecture were true then, as a consequence of Theorem 5 and Proposition 2, DPO would be decidable in dimension one (and undecidable in greater dimensions). Up to now, Conjecture 5 has been proved true for some restricted classes of one-dimensional surjective CA beside open CA.

Theorem 11 ([8]) *Almost equicontinuous surjective CA have JDPO.*

Consider for a while a CA whose alphabet is an algebraic group. A configuration is said to be *finite* if there exists an integer h such that for any i, $|i| > k$, $c(i) = 0$, where 0 is the null element of the group. Denote $s(c) = \sum_i c(i)$ the sum of the values of a finite configuration. A one-dimensional CA F is called *number conserving* if for any finite configuration c, $s(c) = s(F(c))$.

Theorem 12 ([26]) *Number-conserving surjective CA have DPO.*

If a CA F is number-conserving, then for any $h \in \mathbb{Z}$ the CA $\sigma^h \circ F$ is number-conserving. As a consequence we have that

Corollary 1 *Number-conserving surjective CA have JDPO.*

Proof Let F be a number-conserving CA. Choose $h \in \mathbb{Z}$ in such a way that the CA $\sigma^h \circ F$ is a (number-conserving) right CA. By a result in [1], both the CA $\sigma^h \circ F$ and F have JDPO. □

In a recent work it is proved that the problem of solving Conjecture 5 can be reduced to the study of mixing CA.

Theorem 13 ([1]) *If all mixing CA have DPO then every surjective CA has JDPO.*

As a consequence of Theorem 13, if all mixing CA have DPO then all transitive CA have DPO.

Permutivity is another easy-to-check combinatorial property strictly related to chaotic behavior.

Definition 16 (Permutive CA) A function $f : A^s \mapsto A$ is permutive in the variable a_i if for any $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_s) \in A^{s-1}$ the function $a \mapsto f(a_1, \ldots, a_{i-1}, a, a_{i+1}, \ldots, a_s)$ is a permutation.

In the one-dimensional case, a function f that is permutive in the leftmost variable a_1 (resp., rightmost a_s), it is called leftmost (resp. rightmost) permutive. CA with either leftmost or rightmost permutive local rule share most of the chaos components.

Theorem 14 ([18]) *Any one-dimensional CA based on a leftmost (or, rightmost) permutive local rule with $u_1 < 0$ (resp., $u_s > 0$) is topologically mixing.*

The previous result can be generalized to any dimension in the following sense.

Theorem 15 *Let f and N be the local rule and the neighborhood frame, respectively, of a given D-dimensional CA. If there exists i such that*

1. *f is permutive in the variable a_i, and*
2. *the neighbor vector $\vec{u}_i$ is such that $\|\vec{u}_i\|_2 = \max\{\|\vec{u}\|_2 \mid \vec{u} \in N\}$, and*
3. *all the coordinates of $\vec{u}_i$ have absolute value l, for some integer $l > 0$,*

then the given CA is topologically mixing.

Proof Without loss of generality, assume that $\vec{u}_i = (l, \ldots, l)$. Let U and V be two distinct open balls of equal radius 2^{-r}, where r is an arbitrary natural number. For any integer $n \geq 1$, denote with N_n the neighbor frame of the CA F_n, with f^n the corresponding local rule, and with $N_n(\vec{x})$ the set $\{\vec{x} + \vec{v} \mid \vec{v} \in N_n\}$ for a given $\vec{x} \in \mathbb{Z}^D$. Note that f^n is permutive in the variable corresponding to the neighbor vector $n\vec{u}_i \in N_n$. Choose two configurations $c \in U$ and $d \in V$. Let m be the smaller natural number such that $ml > 2r$. For any $n \geq m$, we are going to build a configuration $d_n \in U$ such that $F^n(d_n) \in V$. Set $d_n(\vec{z}) = c(\vec{z})$ for $\vec{z} \in O$ where $O = \{\vec{v} \mid \|\vec{v}\|_\infty \leq r\}$. In this way $d_n \in U$. In order to obtain $F^n(d_n) \in V$, it is required that $F^n(d_n)(\vec{x}) = d(\vec{x})$ for each $\vec{x} \in O$. We complete the configuration d_n by starting with $\vec{x} = \vec{y}$, where $\vec{y} = (-r, \dots, -r)$. Choose arbitrarily the values $d_n(\vec{z})$ for $\vec{z} \in (N_n(\vec{y}) \setminus O) \setminus \{\vec{y} + n\vec{u}_i\}$ (note that $O \subset N_n(\vec{y})$). By permutivity of f^n, there exists $a \in A$ such that if one set $d_n(\vec{y} + n\vec{u}_i) = a$, then $F^n(d_n)(\vec{y}) = d(\vec{y})$. Let now $\vec{x} = \vec{y} + \vec{e}_1$. Choose arbitrarily the values $d_n(\vec{z})$ for $\vec{z} \in (N_n(\vec{x}) \setminus N_n(\vec{y})) \setminus \{\vec{x} + n\vec{u}_i\}$. By the same argument as above, there exists $a \in A$ such that if one set $d_n(\vec{x} + n\vec{u}_i) = a$, then $F^n(d_n)(\vec{x}) = d(\vec{x})$. Proceeding in this way one can complete d_n in order to obtain $F^n(d_n) \in V$. □

Theorem 16 ([16]) *Any one-dimensional CA based on a leftmost (or, rightmost) permutive local rule with $u_1 < 0$ (resp., $u_s > 0$) has (J)DPO.*

Theorem 17 ([54]) *Any one-dimensional CA based on a leftmost and rightmost permutive local rule with $u_1 < 0$ and $u_s > 0$ is expansive.*

As a consequence of Proposition 5, we have the following result for which we give a direct proof in order to make clearer the result which follows immediately after.

Proposition 6 *Any one-dimensional CA based on a leftmost and rightmost permutive local rule with $u_1 < 0$ and $u_s > 0$ is strongly transitive.*

Proof Choose arbitrarily two configurations $c, o \in A^{\mathbb{Z}}$ and an integer $k > 0$. Let $n > 0$ be the first integer such that $nr > k$, where $r = \max\{-u_1, u_s\}$. We are going to construct a configuration $b \in A^{\mathbb{Z}}$ such that $d(b, c) < 2^{-k}$ and $F^n(b) = o$. Fix $b(x) = c(x)$ for each $x = nu_1, \dots, nu_s - 1$. In this way $d(b, c) < 2^{-k}$. For each $i \in \mathbb{N}$ we are going to find suitable values $b(nu_s + i)$ in order to obtain $F^n(b)(i) = o(i)$. Let us start with $i = 0$. By the hypothesis, the local rule f^n of the CA F^n is permutive in the rightmost variable nu_s. Thus, there exists a value $a_0 \in A$ such that, if one sets $b(nu_s) = a_0$, we obtain $F^n(b)(0) = o(0)$. By the same reasons as above, there exists a value $a_1 \in A$ such that, if one set $b(nu_s + 1) = a_1$, we obtain $F^n(b)(1) = o(1)$. Proceeding in this way one can complete the configuration b for any position $nu_s + i$. Finally, since f^n is permutive also in the leftmost variable nu_1 one can use the same technique to complete the configuration b for the positions $nu_1 - 1, nu_1 - 2, \dots$, in such a way that for any integer $i < 0$, $F^n(b)(i) = o(i)$. □

The previous result can be generalized as follows. Denote $\vec{e}_1, \vec{e}_2, \dots, \vec{e}_D$ the canonical basis of $\mathbb{R}^D$.

Theorem 18 *Let f and N be the local rule and the neighborhood frame, respectively, of a given D-dimensional CA. If there exists an integer $l > 0$ such that*

1. *f is permutive in all the 2^D variable corresponding to the neighbor vectors $(\pm l, \dots, \pm l)$, and*
2. *for each vector $\vec{u} \in N$, we have $\|u\|_\infty \leq l$,*

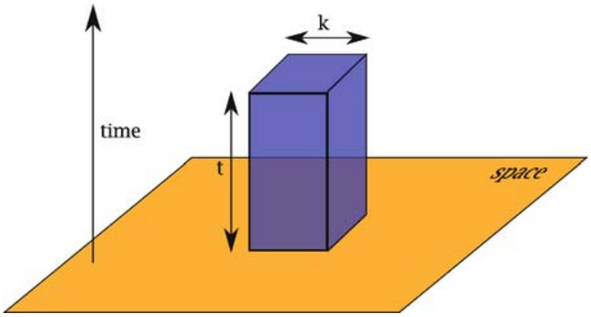
then the CA F is strongly transitive.

Proof For the sake of simplicity, we only trait the case $D = 2$. For higher dimensions, the idea of the proof is the same. Let $\vec{u}_2 = (l, l)$, $\vec{u}_3 = (-l, l)$, $\vec{u}_4 = (-l, -l)$, $\vec{u}_5 = (l, -l)$. Choose arbitrarily two configurations $c, o \in A^{\mathbb{Z}^2}$ and an integer $k > 0$. Let $n > 0$ be the first integer such that $nl > k$. We are going to construct a configuration $b \in A^{\mathbb{Z}^2}$ such that $d(b, c) < 2^{-k}$ and $F^n(b) = o$. Fix $b(x) = c(x)$ for each $\vec{x} \neq n\vec{u}_2$ with $\|\vec{x}\|_\infty \leq n$. In this way $d(b, c) < 2^{-k}$. For each $i \in \mathbb{Z}$ we are going to find suitable values for the configuration b in order to obtain $F^n(b)(i\vec{e}_1) = o(i\vec{e}_1)$. Let us start with $i = 0$. By the hypothesis, the local rule f^n of the CA F^n is permutive in the variable $n\vec{u}_2$. Thus, there exists a value $a_{(0,0)} \in A$ such that, if one set $b(n\vec{u}_1) = a_{(0,0)}$, we obtain $F^n(b)(\vec{0}) = o(\vec{0})$. Now, choose arbitrary values of b in the positions $(n + 1)\vec{e}_1 + j\vec{e}_2$ for $j = -n, \dots, n - 1$. By the same reasons as above, there exists a value $a_{(0,1)} \in A$ such that, if one sets $b(nu_2 + 1\vec{e}_1) = a_{(0,1)}$, we obtain $F^n(b)(\vec{e}_1) = o(\vec{e}_1)$. Proceeding in this way, at each step i ($i > 1$), one can complete the configuration b for all the positions $(n + i)\vec{e}_1 + j\vec{e}_2$ for $j = -n, \dots, n$, obtaining $F^n(b)(i\vec{e}_1) = o(i\vec{e}_1)$. In a similar way, by using the fact that the local rule f^n of the CA F^n is permutive in the variable $n\vec{u}_3$, for any $i < 0$ one can complete the configuration b for all the positions $(-n + i)\vec{e}_1 + j\vec{e}_2$ for $j = -n, \dots, n$, obtaining $F^n(b)(i\vec{e}_1) = o(i\vec{e}_1)$. Now, for each step $j = 1, 2, \dots$, choose arbitrarily the values of b in the positions $i\vec{e}_1 + (n + j)\vec{e}_2$ and $i\vec{e}_1 + (n + j)\vec{e}_2$ with $i = -n, \dots n - 1$. The permutivity of f^n in the variables $n\vec{u}_2$, $n\vec{u}_3$, $n\vec{u}_5$, and $n\vec{u}_4$ permits one to complete the configuration b in the positions $(n + i)\vec{e}_1 + (n + j)\vec{e}_2$ for all integers $i \geq 0$, $(-n + i)\vec{e}_1 + (-n + j)\vec{e}_2$ for all integer $i < 0$, $(-n + i)\vec{e}_1 + (-n - j)\vec{e}_2$ for all integers $i \geq 0$, and $(n + i)\vec{e}_1 + (-n - j)\vec{e}_2$ for all integers $i < 0$, so that for each step j we obtain $\forall i \in \mathbb{Z}, F^n(b)(i\vec{e}_1 + j\vec{e}_2) = o(i\vec{e}_1 + j\vec{e}_2)$. □

CA, Entropy and Decidability

In [34], it is shown that, in the case of CA, the definition of topological entropy can be restated in a simpler form than (1).

The space-time diagram $S(c)$ generated by a configuration c of a D-dimensional CA is the $(D + 1)$-dimensional infinite figure obtained by drawing in sequence the elements of the orbit of initial state c along the temporal axis.



Chaotic Behavior of Cellular Automata, Figure 1
$\mathcal{N}(k,t)$ is the number of distinct *blue blocks* that can be obtained starting from any initial configuration (*orange plane*)

Formally, $S(c)$ is a function from $\mathbb{N} \times \mathbb{Z}^D$ in A defined as $\forall t \in \mathbb{N}, \forall \vec{v} \in \mathbb{Z}^D$, $S(c)(t,\vec{v}) = F^t(c)(\vec{v})$. For a given CA, fix a time t and a finite square region of side of length k in the lattice. In this way, a finite $(D+1)$-dimensional figure (hyper-rectangle) is identified in all space-time diagrams. Let $\mathcal{N}(k,t)$ be the number of distinct finite hyper-rectangles obtained by all possible space-time diagrams for the CA (i. e., $\mathcal{N}(k,t)$ is the number of the all space-time diagrams which are distinct in this finite region). The topological entropy of any given CA can be expressed as

$$h(C,F) = \lim_{k\to\infty} \lim_{t\to\infty} \mathcal{N}(k,t)$$

Despite the expression of the CA entropy is simpler than for a generic DTDS, the following result holds.

Theorem 19 ([34]) *The topological entropy of CA is uncomputable.*

Nevertheless there exist some classes of CA where it is computable [20,45]. Unfortunately, in most of these cases it is difficult to establish if a CA is a member of these classes.

Results for Linear CA: Everything Is Detectable

In the sequel, we assume that a linear CA on C_m is based on a neighborhood frame $N = \vec{u}_1, \dots, \vec{u}_s$ whose corresponding coefficients of the local rule are $\lambda_1, \dots, \lambda_s$. Moreover, without loss of generality, we suppose $\vec{u}_1 = \vec{0}$. In most formulas the coefficient λ_1 does not appear.

Decidability Results for Chaotic Properties

The next results state that all chaotic properties introduced in Section III are decidable. Yet, one can use the formulas to build samples of cellular automata that has the required properties.

Theorem 20 ([17,19,47])

Sensitivity *a linear CA is sensitive to the initial conditions if there exists a prime number p such that $p|m$ and $p \nmid \gcd\{\lambda_2,\dots,\lambda_s\}$.*

Transitivity *a linear CA is topologically transitive if and only if $\gcd\{\lambda_2,\dots,\lambda_s\} = 1$.*

Mixing *a linear CA is topologically mixing if and only if it is topologically transitive.*

Strong transitivity *a linear CA is strongly transitive if for each prime p dividing m, there exist at least two coefficients λ_i, λ_j such that $p \nmid \lambda_i$ and $p \nmid \lambda_j$.*

Regularity (DPO) *a linear CA has the denseness of periodic points if it is surjective.*

Concerning positive expansively, since in dimensions greater than one, there are no such CA, the following theorem characterizes expansively for linear CA in dimension one. For this situation we consider a local rule f with expression $f(x_{-r},\dots,x_r) = \left[\sum_{i=-r}^{r} a_i x_i\right]_m$.

Theorem 21 ([47]) *A linear one dimensional CA is positively expansive if and only if $\gcd\{m, a_{-r},\dots,a_{-1}\} = 1$ and $\gcd\{m, a_1,\dots,a_r\} = 1$.*

Decidability Results for Other Properties

The next result was stated incompletely in [47] since the case of non sensitive CA without equicontinuity points is not treated, tough they exist [57].

Theorem 22 *Let F be a linear cellular automaton. Then the following properties are equivalent*

1. *F is equicontinuous*
2. *F has an equicontinuity point*
3. *F is not sensitive*
4. *for all prime p such that $p|m$, p divides $gcd(\lambda_2,\dots,\lambda_s)$.*

Proof 1) $\Longrightarrow$ 2) and 2) $\Longrightarrow$ 3) are obvious. 3) $\Longrightarrow$ 4) is done by negating the formula for sensitive CA in Theorem 20.

Let us prove that 4) $\Longrightarrow$ 1). Suppose that F is a linear CA. We decompose $F = G + H$ by separating the term in λ_1 from the others:

$$H(x)(\vec{v}) = \lambda_1 x(\vec{v}) \quad G(x)(\vec{v}) = \left[\sum_{i=2}^{s} \lambda_i c(\vec{v}+\vec{u}_i)\right]_m .$$

Let $m = p_1^{\alpha_1} \cdots p_l^{\alpha_l}$ be the decomposition in prime factors and $a = \text{lcm}\{\alpha_i\}$. The condition 4) gives that for all k, $\Pi_{i=1}^{l} p_i | \lambda_k$ and then m divides any product of a factors λ_i.

Let $\vec{v}$ be a vector such that for all i, $\vec{u}_i - \vec{v}$ has nonnegative coordinates. Classically, we represent local rules of linear CA by D-variable polynomials (this representation, together with the representation of configurations by formal power series allows to simplify the calculus of images through the iterates of the CA [36]). Let $X_1, \ldots, X_D$ be the variables. For $\vec{y} = (y_1, \ldots, y_D) \in \mathbb{Z}^D$, we note $X^{\vec{y}}$ the monomial $\Pi_{i=1}^{D} X_i^{y_i}$. We consider the polynomial P associated with G combined with a translation of vector $\vec{v}$, $P = \Pi_{i=2}^{s} \lambda_i X^{\vec{u}_i - \vec{v}}$. The coefficients of P^a are products of a factors λ_i hence $[P^a]_m = 0$. This means that the composition of G and the translation of vector $\vec{v}$ is nilpotent and then that G is nilpotent. As F is the sum of λ_1 times the identity and a nilpotent CA, we conclude that F is equicontinuous. □

The next theorem gives the formula for some combinatorial properties

Theorem 23 **([36]).**

Surjectivity *a linear CA is surjective if and only if* $\gcd\{\lambda_1, \ldots, \lambda_s\} = 1$.

Injectivity *a linear CA is injective if and only if for each prime p decomposing m there exists an unique coefficient λ_i such that p does not divide λ_i.*

Computation of Entropy for Linear Cellular Automata

Let us start by considering the one-dimensional case.

Theorem 24 *Let us consider a one dimensional linear CA. Let $m = p_1^{k_1} \cdots p_h^{k_h}$ be the prime factor decomposition of m. The topological entropy of the CA is*

$$h(C, F) = \sum_{i=1}^{h} k_i (R_i - L_i) \log(p_i)$$

where $L_i = \min P_i$ and $R_i = \max P_i$, with $P_i = \{0\} \cup \{j\colon \gcd(a_j, p_i) = 1\}$.

In [50], it is proved that for dimensions greater than one, there are only two possible values for the topological entropy: zero or infinity.

Theorem 25 *A D-dimensional linear CA $\langle C, F\rangle$ with $D \geq 2$ is either sensitive and $h(C, F) = \infty$ or equicontinuous and $h(C, F) = 0$.*

By a combination of Theorem 25 and 20, it is possible to establish if a D dimensional linear CA with $D \geq 2$ has either zero or infinite entropy.

Linear CA, Fractal Dimension and Chaos

In this section we review the relations between strong transitivity and fractal dimension in the special case of linear CA. The idea is that when a system is chaotic then it produces evolutions which are complex even from a (topological) dimension point of view.

Any linear CA F, can be associated with its *W-limit set*, a subset of the $(D+1)$-dimensional Euclid space defined as follows. Let t_n be a sequence of integers (we call them times) which tends to infinity. A subset $S_F(t_n)$ of $(D+1)$-dimensional Euclid space represents a space-time pattern until time $t_n - 1$:

$$S_F(t_n) = \left\{(t, i) \text{s.t.} F^t(e_1)_i \neq 0,\ t < t_n\right\} .$$

A *W-limit set* for F is defined by $\lim_{n\to\infty} S_F(t_n)/t_n$ if the limit exists, where $S_F(t_n)/t_n$ is the contracted set of $S_F(t_n)$ by the rate $\frac{1}{t_n}$ i. e. $S_F(t_n)/t_n$ contains the point $(t/t_n, i/t_n)$ if and only if $S_F(t_n)$ contains the point (t, i). The limit $\lim_{n\to\infty} S_F(t_n)/t_n$ exists when $\liminf_{n\to\infty} S_F(t_n)/t_n$ and $\limsup_{n\to\infty} S_F(t_n)/t_n$ coincide, where

$$\liminf_{n\to\infty} \frac{S_F(t_n)}{t_n} = \Bigg\{ x \in \mathbb{R}^{D+1} \ :\ \forall j, \exists x_j \in \frac{S_F(t_j)}{t_j}, x_j \to x \text{ when } j \to \infty \Bigg\}$$

and

$$\limsup_{n\to\infty} \frac{S_F(t_n)}{t_n} = \Bigg\{ x \in \mathbb{R}^{D+1} \ :\ \exists \left\{t_{n_j}\right\}, \forall j, \exists x_{n_j} \in \frac{S_F(t_{n_j})}{t_{n_j}}, x_{n_j} \to x \text{ when } j \to \infty \Bigg\} ,$$

for a subsequence $\left\{t_{n_j}\right\}$ of $\{t_n\}$.

For the particular case of linear CA, the W-limit set always exists [30,31,32,56]. In the last ten years, the W-limit set of additive CA has been extensively studied [61,62,63]. It has been proved that for most of additive CA, it has interesting dimensional properties which completely characterize the set of quiescent configurations [56]. Here we link dimension properties of a W-limit set with chaotic properties. Correlating dimensional properties of invariant sets to dynamical properties has become during the years a fruitful source of new understanding [52].

Let X be a metric space. The *Hausdorff dimension* D_H of $V \subseteq X$ is defined as:

$$D_H(V) = \sup\left\{ h \in \mathbb{R} |\ \lim_{\epsilon\to 0} \left(\inf \sum |U_i|^h\right) = \infty \right\}$$

where the infimum is taken over all countable coverings U_i of V such that the diameter $|U_i|$of each U_i is less than ϵ (for more on Hausdorff dimension as well as other definitions of fractal dimension see [24]).

Given a CA F we denote $D_H(F)$ the Hausdorff dimension of its W-limit set.

Proposition 7 ([25,47]) *Consider a linear CA F over Z_{p^k} where p is a prime number. If $1 < D_H(F) < 2$ then F is strongly transitive.*

The converse relation is still an open problem. It would be also an interesting research direction to find out similar notions and results for general CA.

Conjecture 6 *Consider a linear CA F over Z_{p^k}, where p is a prime number. If F is strongly transitive then $1 < D_H(F) < 2$.*

Future Directions

In this chapter we reviewed the chaotic behavior of cellular automata. It is clear from the results seen so far that there are close similarities between the chaotic behavior of dynamical systems on the real interval and CA. To complete the picture, it remains only to prove (or disprove) Conjecture 5. Due to its apparent difficulty, this problem promises to keep researchers occupied for some years yet.

The study of the decidability of chaotic properties like expansively, transitivity, mixing etc. is another research direction which should be further addressed in the near future. It seems that new ideas are necessary since the proof techniques used up to now have been revealed as unsuccessful. The solution to these problems will be a source of new understanding and will certainly produce new results in connected fields.

Finally, remark that most of the results on the chaotic behavior of CA are concerned with dimension one. A lot of work should be done to verify what happens in higher dimensions.

Acknowledgments

This work has been supported by the Interlink/MIUR project “Cellular Automata: Topological Properties, Chaos and Associated Formal Languages”, by the ANR Blanc Project “Sycomore” and by the PRIN/MIUR project “Formal Languages and Automata: Mathematical and Applicative Aspects”.

Bibliography

Primary Literature

1. Acerbi L, Dennunzio A, Formenti E (2007) Shifting and lifting of cellular automata. In: Third Conference on Computability in Europe, CiE 2007, Siena, Italy, 18–23 June 2007. Lecture Notes in Computer Science, vol 4497. Springer, Berlin, pp 1–10
2. Adler R, Konheim A, McAndrew J (1965) Topological entropy. Trans Amer Math Soc 114:309–319
3. Akin E, Auslander E, Berg K (1996) When is a transitive map chaotic? In: Bergelson V, March P, Rosenblatt J (eds) Convergence in Ergodic Theory and Probability. de Gruyter, Berlin, pp 25–40
4. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tesselation structures. J Comp Syst Sci 6:448–464
5. Auslander J, Yorke JA (1980) Interval maps, factors of maps and chaos. Tohoku Math J 32:177–188
6. Banks J, Brooks J, Cairns G, Davis G, Stacey P (1992) On Devaney's definition of chaos. Am Math Mon 99:332–334
7. Blanchard F, Maass A (1997) Dynamical properties of expansive one-sided cellular automata. Israel J Math 99:149–174
8. Blanchard F, Tisseur P (2000) Some properties of cellular automata with equicontinuity points. Ann Inst Henri Poincaré, Probabilité et Statistiques 36:569–582
9. Blanchard F, Kůrka P, Maass A (1997) Topological and measure-theoretic properties of one-dimensional cellular automata. Physica D 103:86–99
10. Blanchard F, Formenti E, Kůrka K (1998) Cellular automata in the Cantor, Besicovitch and Weyl topological spaces. Complex Syst 11:107–123
11. Blanchard F, Cervelle J, Formenti E (2005) Some results about chaotic behavior of cellular automata. Theor Comp Sci 349:318–336
12. Blanchard F, Glasner E, Kolyada S, Maass A (2002) On Li-Yorke pairs. J Reine Angewandte Math 547:51–68
13. Boyle M, Kitchens B (1999) Periodic points for cellular automata. Indag Math 10:483–493
14. Boyle M, Maass A (2000) Expansive invertible one-sided cellular automata. J Math Soc Jpn 54(4):725–740
15. Cattaneo G, Formenti E, Margara L, Mazoyer J (1997) A Shift-invariant Metric on S^Z Inducing a Non-trivial Topology. In: Mathmatical Foundations of Computer Science 1997. Lecture Notes in Computer Science, vol 1295. Springer, Berlin, pp 179–188
16. Cattaneo G, Finelli M, Margara L (2000) Investigating topological chaos by elementary cellular automata dynamics. Theor Comp Sci 244:219–241
17. Cattaneo G, Formenti E, Manzini G, Margara L (2000) Ergodicity, transitivity, and regularity for linear cellular automata. Theor Comp Sci 233:147–164. A preliminary version of this paper has been presented to the Symposium of Theoretical Computer Science (STACS'97). LNCS, vol 1200
18. Cattaneo G, Dennunzio A, Margara L (2002) Chaotic subshifts and related languages applications to one-dimensional cellular automata. Fundam Inform 52:39–80
19. Cattaneo G, Dennunzio A, Margara L (2004) Solution of some conjectures about topological properties of linear cellular automata. Theor Comp Sci 325:249–271
20. D'Amico M, Manzini G, Margara L (2003) On computing the entropy of cellular automata. Theor Comp Sci 290:1629–1646
21. Denker M, Grillenberger C, Sigmund K (1976) Ergodic Theory on Compact Spaces.Lecture Notes in Mathematics, vol 527. Springer, Berlin
22. Devaney RL (1989) An Introduction to chaotic dynamical systems, 2nd edn. Addison-Wesley, Reading
23. Durand B, Formenti E, Varouchas G (2003) On undecidability of equicontinuity classification for cellular automata. Dis-

crete Mathematics and Theoretical Computer Science, vol AB. pp 117–128

24. Edgar GA (1990) Measure, topology and fractal geometry. Undergraduate texts in Mathematics. Springer, New York
25. Formenti E (2003) On the sensitivity of additive cellular automata in Besicovitch topologies. Theor Comp Sci 301(1–3):341–354
26. Formenti E, Grange A (2003) Number conserving cellular automata II: dynamics. Theor Comp Sci 304(1–3):269–290
27. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in diophantine approximation. Math Syst Theor (now Theor Comp Syst) 1(1):1–49
28. Glasner E, Weiss B (1993) Sensitive dependence on initial condition. Nonlinearity 6:1067–1075
29. Guckenheimer J (1979) Sensitive dependence to initial condition for one-dimensional maps. Commun Math Phys 70:133–160
30. Haeseler FV, Peitgen HO, Skordev G (1992) Linear cellular automata, substitutions, hierarchical iterated system. In: Fractal geometry and Computer graphics. Springer, Berlin
31. Haeseler FV, Peitgen HO, Skordev G (1993) Multifractal decompositions of rescaled evolution sets of equivariant cellular automata: selected examples. Technical report, Institut für Dynamische Systeme, Universität Bremen
32. Haeseler FV, Peitgen HO, Skordev G (1995) Global analysis of self-similarity features of cellular automata: selected examples. Physica D 86:64–80
33. Hedlund GA (1969) Endomorphism and automorphism of the shift dynamical system. Math Sy Theor 3:320–375
34. Hurd LP, Kari J, Culik K (1992) The topological entropy of cellular automata is uncomputable. Ergodic. Th Dyn Sy 12:255–265
35. Hurley M (1990) Ergodic aspects of cellular automata. Ergod Theor Dyn Sy 10:671–685
36. Ito M, Osato N, Nasu M (1983) Linear cellular automata over z_m. J Comp Sy Sci 27:127–140
37. IV Assaf D, Gadbois S (1992) Definition of chaos. Am Math Mon 99:865
38. Kannan V, Nagar A (2002) Topological transitivity for discrete dynamical systems. In: Misra JC (ed) Applicable Mathematics in Golden Age. Narosa Pub, New Dehli
39. Kari J (1994) Reversibility and surjectivity problems of cellular automata. J Comp Sy Sci 48:149–182
40. Kari J (1994) Rice's theorem for the limit set of cellular automata. Theor Comp Sci 127(2):229–254
41. Knudsen C (1994) Chaos without nonperiodicity. Am Math Mon 101:563–565
42. Kolyada S, Snoha L (1997) Some aspect of topological transitivity – a survey. Grazer Mathematische Berichte 334:3–35
43. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergo Theor Dyn Sy 17:417–433
44. Kůrka P (2004) Topological and Symbolic Dynamics, vol 11 of Cours Spécialisés. Société Mathématique de France, Paris
45. Di Lena P (2006) Decidable properties for regular cellular automata. In: Navarro G, Bertolossi L, Koliayakawa Y (eds) Proceedings of Fourth IFIP International Conference on Theoretical Computer Science, pp 185–196. Springer, Santiago de Chile
46. Li TY, Yorke JA (1975) Period three implies chaos. Am Math Mon 82:985–992
47. Manzini G, Margara L (1999) A complete and efficiently computable topological classification of D-dimensional linear cellular automata over Z_m. Theor Comp Sci 221(1–2):157–177
48. Margara L (1999) On some topological properties of linear cellular automata. In: Kutylowski M, Pacholski L, Wierzbicki T (eds) Mathematical Foundations of Computer Science 1999 (MFCS99). Lecture Notes in Computer Science, vol 1672. Springer, Berlin, pp 209–219
49. Moothathu TKS (2005) Homogenity of surjective cellular automata. Discret Contin Dyn Syst 13:195–202
50. Morris G, Ward T (1998) Entropy bounds for endomorphisms commuting with k actions. Israel J Math 106:1–12
51. Nasu M (1995) Textile Systems for Endomorphisms and automorphisms of the shift, vol 114 of Memoires of the American Mathematical Society. American Mathematical Society, Providence
52. Pesin YK (1997) Dimension Theory in Dynamical Systems. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago
53. Shereshevsky MA (1993) Expansiveness, entropy and polynomial growth for groups acting on subshifts by automorphisms. Indag Math 4:203–210
54. Shereshevsky MA, Afraimovich VS (1993) Bipermutative cellular automata are topologically conjugate to the one-sided Bernoulli shift. Random Comput Dynam 1(1):91–98
55. Sutner K (1999) Linear cellular automata and de Bruijn automata. In: Delorme M, Mazoyer J (eds) Cellular Automata, a Parallel Model, number 460 in Mathematics and Its Applications. Kluwer, Dordrecht
56. Takahashi S (1992) Self-similarity of linear cellular automata. J Comp Syst Sci 44:114–140
57. Theyssier G (2007) Personal communication
58. Vellekoop M, Berglund R (1994) On intervals, transitivity = chaos. Am Math Mon 101:353–355
59. Walters P (1982) An Introduction to Ergodic Theory. Springer, Berlin
60. Weiss B (1971) Topological transitivity and ergodic measures. Math Syst Theor 5:71–5
61. Willson S (1984) Growth rates and fractional dimensions in cellular automata. Physica D 10:69–74
62. Willson S (1987) Computing fractal dimensions for additive cellular automata. Physica D 24:190–206
63. Willson S (1987) The equality of fractional dimensions for certain cellular automata. Physica D 24, 179–189
64. Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific, Singapore

Books and Reviews

Akin E (1993) The general topology of dynamical systems. Graduate Stud. Math 1 Am Math Soc, Providence

Akin E, Kolyada S (2003) Li-Yorke sensitivity. Nonlinearity 16:1421–1433

Block LS, Coppel WA (1992) Dynamics in One Dymension. Springer, Berlin

Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge

Kitchens PB (1997) Symbolic dynamics: One-Sided, Two-Sided and Countable State Markov Shifts. Universitext. Springer, Berlin

Kolyada SF (2004) Li-yorke sensitivity and other concepts of chaos. Ukr Math J 56(8):1242–1257

Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambidge University Press, Cambidge

Chaotic Dynamics in Nonequilibrium Statistical Mechanics

J. ROBERT DORFMAN
Institute for Physical Science and Technology
and Department of Physics, University of Maryland,
College Park, USA

Article Outline

Glossary

Chaotic systems The time evolution of a deterministic mechanical system defines a trajectory in the phase space of all the generalized coordinates and generalized momenta. Consider two infinitesimally separated points that lie on two different trajectories in this phase space. If these two trajectories typically separate exponentially with time, the systems is called chaotic provided the set of all points with an exponentially separating partner is of positive measure.

Chaotic hypothesis The hypothesis that systems of large numbers of particles interacting with short ranged forces can be treated mathematically as if the system were chaotic with no pathologies in the mathematical description of the systems' trajectories in phase space.

Dynamical systems theory The mathematical theory of the time evolution in phase space, or closely related spaces, of a deterministic system, such as a mechanical system obeying Hamiltonian equations of motion.

Ergodic systems A mechanical system is called ergodic if a typical trajectory in a phase space of finite total measure spends a fraction of its time in a set which is equal to the ratio of the measure of the set to the total measure of the phase space.

Escape rate formula Consider a chaotic dynamical system with a phase space constructed in such a way that the phase space has some kind of an absorbing boundary. The set of points, $\mathcal{R}$, in the phase space such that trajectories through them never escape through the absorbing boundary either in the forward or the backward motion is called a repeller. One can define a set of Lyapunov exponents, $\lambda_i(\mathcal{R})$ and a Kolmogorov–Sinai entropy, $h_{\text{KS}}(\mathcal{R})$ for motion on the repeller. Dynamical systems theory shows that the rate of escape, γ, of points, not on the repeller, through the boundary is given by

$$\gamma = \sum_i \lambda^+(\mathcal{R}) - h_{\text{KS}}(\mathcal{R}) \,, \tag{1}$$

where the sum is over all of the positive Lyapunov exponents on the repeller.

Gaussian thermostats A dynamical friction acting on the particles in a mechanical system which keeps the total energy, or the total kinetic energy of the system at a fixed value. It was invented by Gauss as the simplest solution to the problem of finding the equations of motion for a mechanical system with a constraint of fixed energy.

Gelfand triplet An operator with right and left hand eigenfunctions, possibly defined in different function spaces, and an inner product of one function from the *right* space and one from the *left* space. Generally one of these spaces contains singular functions such as Schwartz distributions and the other contains sufficiently smooth functions so that the inner product is well defined. The term *rigged Hilbert space* is also used to denote a Gelfand triplet.

Hyperbolic dynamical system A chaotic system where the tangent space to almost all trajectories in its phase space can be separated into well-defined stable and unstable manifolds, that intersect each other transversally.

Kolmogorov–Sinai entropy per unit time A measure of the rate at which information about the initial point of a chaotic trajectory is produced in time. The exponential separation of trajectories in phase space, characteristic of chaotic motion, implies that trajectories starting at very close-by, essentially indistinguishable, initial points will eventually be distinguishable. Hence as time evolves we can specify more precisely the initial point of the trajectory. Pesin has proved that for closed, hyperbolic systems, the Kolmogorov–Sinai entropy is equal to the sum of the positive Lyapunov exponents. The Kolmogorov–Sinai entropy is often called the metric entropy.

Lyapunov exponents Lyapunov exponents, λ_i, are the rates at which infinitesimally close trajectories separate or approach with time on the unstable and stable manifolds of a chaotic dynamical system. For closed phase spaces, that is, no absorbing boundaries present, Pesin theorem states that for hyperbolic dynamical system the Kolmogorov–Sinai entropy, h_{KS} is given by the sum of all the positive Lyapunov exponents.

$$h_{\mathrm{KS}} = \sum_i \lambda_i^+ \,. \tag{2}$$

Mixing systems Mixing systems are dynamical systems with stronger dynamical properties than ergodic systems in the sense that every mixing system is ergodic but the converse is not true. A system is mixing if the following statement about the time development of a set A_t is satisfied

$$\lim_{T\to\infty} \frac{\mu(A_T \cap B)}{\mu(B)} = \frac{\mu(A)}{\mu(\mathcal{E})} \,, \tag{3}$$

where B is any set of finite measure, and $\mu(\mathcal{E})$ is the measure of the full phase space. This statement is the mathematical expression of the fact that for a mixing system, every set of finite measure becomes uniformly distributed throughout the full phase space, with respect to the measure μ.

Normal variables Microscopic variables whose values are approximately constant on large regions of the constant energy surface in phase space.

Pseudo-chaotic systems A pseudo-chaotic system is a dynamical system where the separation of nearby trajectories is algebraic in time, rather than exponential. Pseudo-chaotic systems are weakly mixing as defined by the relation

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}\tau \left[\mu(A_\tau \cap B) - \frac{\mu(A)\mu(B)}{\mu(\mathcal{E})} \right] = 0 \,. \tag{4}$$

Sinai–Ruelle–Bowen (SRB) measure SRB measures for a chaotic system are invariant measures that are smooth on unstable manifolds and possibly singular on stable manifolds.

Stable manifold A stable manifold about a point P in phase space is the set of points that will approach P at time t approaches positive infinity, that is in the infinite future of the motion.

Transport coefficients Transport coefficients characterize the proportionality between the currents of particles, momentum, or energy in a fluid, and the gradients of density, fluid velocity or temperature in the fluid. The coefficients of diffusion, shear and bulk viscosity, and thermal conductivity are transport coefficients, and appear as coefficients of the second order gradients in the Navier–Stokes and similar equations.

Unstable manifold An unstable manifold about a point P in phase space is the set of points that will approach P as time approaches negative infinity, that is, as one follows the motion backwards in time to the infinitely remote past.

Definition of the Subject

For most of its history, non-equilibrium statistical mechanics has produced mathematical descriptions of irreversible processes by invoking one or another stochastic assumptions in order to obtain useful equations. Central to our understanding of transport in fluids, for example, are random walk processes, which typically are described by stochastic equations. These in turn lead to the Einstein relation for diffusion, and its generalizations to other transport processes. This relation, as formulated by Einstein, states that the mean square displacement of a diffusing particle grows linearly in time with a proportionality constant given by the coefficient of diffusion. If we assume that such a description applies to mechanical systems of many particles, we must explain the origins of irreversibility in deterministic – and time reversible – mechanical systems, and we must locate the source of stochasticity that is invoked to derive transport equations. For systems of particles that are chaotic, it is possible to make some progress on resolving these issues, and to gain some insights into the analogous properties of systems that are not chaotic or that have both chaotic and non-chaotic behaviors. This article summarizes the current status of the application of dynamical systems theory to non-equilibrium statistical mechanics, focuses on the behavior of chaotic systems, and presents some of the important open issues needing resolution.

Introduction

Statistical mechanics is devoted to the study of the collective properties of large numbers of particles, typically atoms, molecules, electrons, etc., but can also be stars, or even galaxies. The mechanical properties of the individual particles and their mutual interactions are presumed to be known, and the emphasis is on the statistical properties of collections of large numbers of them [1,2,3]. The study of statistical mechanics has always involved both an examination of its foundations and the devising of methods to compute thermodynamic, transport, and related quantities for systems of large numbers of particles. In recent years there has been a great deal of attention focused on the foundations of statistical mechanics prompted by developments in: (i) dynamical systems theory, particularly chaotic dynamics, (ii) the mathematics of Anosov and hard-ball systems, and in (iii) the computational physics of systems of particles undergoing processes of various kinds [4,5,6,7,8,9,10,11,12,13,14]. Moreover, modern developments in quantum physics, particularly those of some relevance for the foundations of quantum mechanics, for cosmology, and for quantum computation have some bearing on the foundations of quantum statistical mechanics [15].

Non-equilibrium statistical mechanics is faced with the task of describing the wide range of non-thermodynamic, generally time dependent behaviors of large systems, such as the transport of particles, momentum, and/or energy from one region in space to another over some time interval. Despite a large number of, as yet, unsolved problems, non-equilibrium statistical mechanics has been able to explain and provide quantitative predictions for a wide range of transport phenomena. Here we will discuss some of these applications and describe the role played by the microscopic dynamics of the constituent particles, particularly when the microscopic dynamics is classical and chaotic [14]. Most of this article will be devoted to a study of the role of chaotic dynamics in non-equilibrium transport for classical, chaotic systems. We will also consider, to some extent, classical, non-chaotic systems as well as quantum systems, where the usual notions of chaotic dynamics do not apply but one can nevertheless understand some main features of the behavior of a quantum system by examining its classical counterpart, if there is one.

This article is devoted to the role played by chaotic dynamics in our understanding non-equilibrium statistical mechanics. It will focus on two main topics: (1) Basic issues in statistical mechanics, namely the approach of systems of particles to thermodynamic equilibrium. Here we give an updated view of the role of ergodic and mixing properties, proposed by Boltzmann and Gibbs, respectively, as the basis for understanding the approach to equilibrium [1,2,3,4]. (2) Applications of dynamical systems theory to non-equilibrium statistical mechanics and the theory of transport processes. Here we show that for chaotic systems, at least, on can find some very deep relationships between macroscopic transport and microscopic dynamics [5,6,7,8,9,10,11,12,13]. We also discuss, briefly, the closely related topic of fluctuation theorems for non-equilibrium stationary states which apply even in far-from equilibrium situations [16,17,18,19,20,21].

By way of introduction we begin with some observations about transport processes in fluids in order to set the stage for describing the role of chaotic dynamics in transport theory. As we will discuss in more detail below, normal transport processes in fluids are characterized by a linear growth in time of the mean square fluctuations of an appropriate time dependent microscopic variable of the system. Typical of such a fluctuation formula is the Einstein relation which states that the mean square displacement of a particle undergoing Brownian motion in a fluid grows linearly in time with a coefficient that is, apart from a numerical factor, the diffusion coefficient of the Brownian particle [22]. That is

$$\left\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \right\rangle = 2dDt \,. \tag{5}$$

Here $\mathbf{r}(t)$ is the spatial location of the Brownian particle at some time t, and d is the number of spatial dimensions of the system. D is the diffusion coefficient appearing in the linear relation, known as Fick's Law, that relates the probability current, $\mathbf{J}(\mathbf{r}, t)$ for the Brownian particle to its probability density, $P(\mathbf{r}, t)$,

$$\mathbf{J}(\mathbf{r}, t) = -D\nabla P(\mathbf{r}, t) \,. \tag{6}$$

The angular brackets denote an average over an ensemble of trajectories of time duration t, at least. At the heart of the Einstein formula is the fact that the Brownian particle undergoes a random walk on sufficiently long time scales. The distinguishing feature of a random walk is the linear growth in time of the mean square displacement. Thus, normal transport in general is a form of a random walk process. This was formalized by Helfand in 1960 [23], who generalized the Einstein formula to other transport processes such as viscous flow, and heat conduction. By normal transport we mean transport that can be described macroscopically by linear relations between the currents of globally conserved quantities such as mass, energy, and/or momentum, and the gradients of the densities of these quantities. The coefficients of proportionality are transport coefficients, such as the coefficients of shear and bulk viscosity, thermal conductivity, diffusion, and so on. These coefficients, for normal transport, do not depend on time, but may depend on the local densities of mass and the local temperature of the fluid.

Helfand was able to show that each transport process may be regarded as a random walk process characterized by a linear growth as a function of time of the mean square fluctuation of a microscopic quantity, $M_\sigma(\Gamma, t)$, called a Helfand moment. The Helfand moments depend on all of the phase space variables of the system, now denoted collectively by Γ, and time t. For normal transport, the generalized Einstein relation becomes

$$\left\langle (M_\sigma(\Gamma, t) - M_\sigma(\Gamma, t = 0))^2 \right\rangle = \sigma C_\sigma t \,. \tag{7}$$

Here σ is a transport coefficient appearing in the Navier–Stokes or diffusion equations, C_σ is a constant, and the average is over an equilibrium ensemble. For diffusion of a Brownian particle, for example, the Helfand moment is simply the spatial location of the particle. This result implies that normal transport processes are essentially random walk processes with a generalized "displacement", namely the Helfand moment.

If we think of a system of particles, undergoing some kind of hydrodynamic flow as a deterministic dynamical system, with, typically but not exclusively, Hamiltonian dynamics, a deep question immediately presents itself: *Where does the randomness come from that is required for transport processes to be generalized random-walk processes?* While there are a variety of answers to this question, the main point of this article is to argue that for *chaotic systems* the randomness is an intrinsic property of the dynamics of the system [14], and then to illustrate some of the results that have been obtained recently which connect microscopic dynamics to macroscopic transport for such systems. We emphasize that many chaotic systems display the kind of randomness needed for good transport properties even though such systems are deterministic, and time reversible. However chaos is neither necessary nor sufficient for normal transport. There are examples of non-chaotic systems that exhibit normal diffusion The wind-tree model to be discussed below is an example [24,25]. There are also chaotic systems that exhibit abnormal transport. The two dimensional periodic Lorentz gas with circular scatterers is chaotic and exhibits both normal and abnormal diffusion, depending upon the lattice structure and the separation distance between neighboring scatterers [26,27].

Nevertheless, chaotic systems have sufficiently many nice properties that for them it is possible to obtain a number of rather, general connections between microscopic dynamics and macroscopic transport, and to understand the approach of systems to thermodynamic equilibrium from a dynamical standpoint. Such connections, should they exist, are not yet available for those non-chaotic systems which exhibit normal transport. The dynamical properties of most non-chaotic systems are not yet sufficiently well understood for the construction of a more or less general theory for generic connections between microscopic dynamics and macroscopic transport. Most realistic dynamical systems have a *mixed* phase space, where chaotic and non-chaotic regions, to be defined below, are intermingled [28,29]. It is generally supposed that for systems of large numbers of particles in macroscopic volumes, the non-chaotic regions occupy a very small part of phase space and can be ignored for all practical purposes. However this remains to be proved, and the effects of the non-chaotic regions require much more elucidation.

The plan of this article is as follows: We begin with a discussion of the foundations of statistical mechanics in Sect. "The Roles of Ergodicity and Mixing for the Approach to Equilibrium". There we discuss dynamical properties that are often considered to be important for statistical mechanics, such as ergodicity and mixing. There we argue that these properties, by themselves, are not sufficient to explain the approach to equilibrium of systems of large numbers of particles and that further notions are required. This will lead us in Sect. "Integrable, Pseudo-chaotic, and Chaotic Systems" to a discussion of some of the dynamical systems encountered in classical statistical mechanics, particularly integrable, pseudo-chaotic, and chaotic systems. In Sect. "Anosov and Anosov-like Systems; the Chaotic Hypothesis", we consider chaotic systems in more detail and show for a simple model of a chaotic system that the stochastic behavior needed for diffusion can be seen as a natural consequence of chaotic dynamics. Despite its simplicity and low dimensionality, this model will also allow us to provide a dynamical picture of the approach of a system to equilibrium. These models are simple examples of classes of dynamical systems called Anosov and Anosov-like systems. They have very useful mathematical properties for the description of non-equilibrium phenomena in simple systems, and it is convenient to assume, whenever possible, that the systems of interest to physicists have these good properties. These properties are then used for the applications of chaotic dynamics to non-equilibrium statistical mechanics in Sect. "Applications of Dynamical Systems Theory to Non-equilibrium". There we discuss a number of results that connect microscopic dynamical properties such as Lyapunov exponents and Kolmogorov–Sinai entropies [4,5,6,7,8,9,10,11,12,13,14] to macroscopic transport coefficients for chaotic systems. In Sect. "Discussion" we review the previous discussion and give brief comments about quantum systems, where the concepts of classical chaotic dynamics do not apply, but certain features of these systems can be understood in terms of the chaotic dynamics of their classical counterparts. Finally, we mention some future directions and open questions in Sect. "Future Directions".

The Roles of Ergodicity and Mixing for the Approach to Thermodynamic Equilibrium

Ergodic Systems

Boltzmann invented the notion of ergodicity in an attempt to reconcile the apparently irreversible behavior of large systems of particles, especially their approach to thermodynamic equilibrium, with the microscopic reversibility of

Newton's equations of motion [1,2,3]. Briefly put, Boltzmann argued as follows: Consider an isolated system of particles, and follow the system's trajectory on the appropriate constant energy surface in phase space. Suppose that the underlying microscopic dynamics of the system is such that the phase space trajectory of the system, over a long time interval, spends an amount of time in every region of phase space, A, that is proportional to the measure of that region, $\mu(A)$. Here the measure is the standard equilibrium phase space measure on a constant energy surface given by

$$\mu(A) = \int_A \frac{\mathrm{d}S}{|\nabla H|}\,, \tag{8}$$

where dS is the area of an infinitesimal region of the constant energy surface, and $|\nabla H|$ is the magnitude of the gradient of the Hamiltonian function for the system at the point of integration. Boltzmann's supposition above is called the *ergodic hypothesis*, and in mathematical terms it can be stated as [4]

$$\lim_{T\to\infty} \frac{\tau(A)}{T} = \frac{\mu(A)}{\mu(\mathcal{E})}\,. \tag{9}$$

Here $\tau(A)$ is the amount of time the system spends in region A during a time interval T, and $\mu(\mathcal{E})$ is the measure of the entire constant energy surface, which we assume to be finite. If one accepts the hypothesis, then one can easily show, by approximating integrals by sums, for example, that the *time average* of any well behaved dynamical quantity approaches its equilibrium, micro-canonical ensemble average as the time interval over which the average is taken approaches infinity. Boltzmann's hypothesis holds equally well for the time reversed motion of the system, and is consistent with the reversibility of the microscopic dynamics.

The kinds of elementary physical systems one studies in physics courses, such as coupled systems of harmonic oscillators, are generally not ergodic, although some simple systems, such as a single, one dimensional harmonic oscillator, can be shown to be ergodic. Another simple example is the discrete time motion of a point particle on the circumference of a circle, where the particle moves a fixed irrational fraction of the length of the circumference at successive time steps. In the long time limit, the circumference is uniformly covered by points visited by the particle [30]. A great deal of mathematical research over the past few decades, and longer, has been devoted to a study of more complicated ergodic systems. The first dramatic example of a system with ergodic properties, and one that has influenced most of the more recent efforts in this direction is the geodesic motion of a point on a surface of constant negative curvature. This was proved by E. Hopf [31], and the techniques employed remain useful today. Sinai and coworkers [32], as well as many other workers, particularly Simányi [33], have given careful mathematical proofs of ergodic behavior of various systems composed of hard disks or hard spheres, generally referred to as hard-ball systems. Turaev and Rom-Kedar, Donnay and others [28,29] have shown that by softening the hard sphere potential one may change an ergodic system into a non-ergodic one. Thus, the problem of proving that a system of interest to physicists is actually ergodic, or ergodic for practical purposes, is still far from a general solution.

Mixing Systems

Gibbs took a different approach to the problem of irreversibility. He used the analogy of an ink drop being stirred in a container of glycerin to introduce the stronger notion of a *mixing system* in his efforts to understand the approach of a non-equilibrium ensemble distribution function to an equilibrium distribution [4,8]. In considering a non-equilibrium phase space distribution one has to follow the trajectories of a set of points in phase space, not just a typical trajectory, as in the study of ergodic behavior. Gibbs suggested that the an initial set of points concentrated in a small region of phase space might spread out over the entire available phase space in the course of time, and become finely mixed throughout the phase space in much the same way as the ink drop will become mixed in the glycerin, in such a way that a coarse grained observation of the ink would lead to the conclusion that it is uniformly mixed, but a fine grained observation would reveal the individual threads of ink. Returning to phase space, Gibbs argued that although the measure of phase space occupied by the set of points should remain constant in time, in accordance with Liouville's theorem, the set eventually gets distributed over the constant energy surface such that a coarse grained observation of the phase space would lead to the conclusion that the set uniformly covers the energy surface, while a fine grained observation would reveal that the coverage consists of one long, thin set of total measure much less than that of the whole energy surface. Of course, mechanical reversibility ensures that one can recover the initial distribution of points by time reversing the motion but eventually mixing takes place for the time reversed motion, as well.

The mathematical definition of a mixing system [30] is given by looking at the time development of the initial set of phase points. We denote by A, the initial set of phase points on which the initial ensemble is concentrated, and the set to which this initial set evolves after a time T, by A_T. Then, to examine the mixing of the set in phase space, we consider some arbitrary set of positive measure B and consider the intersection $B \cap A_T$. If the dynamics of the system is mixing, then eventually the fraction of the set B occupied by A_T should approach the fraction of A in the entire phase space, no matter what sets of positive measure A and B we consider. That is, a system is mixing if for any sets A and B of positive measure

$$\lim_{T\to\infty} \frac{\mu(A_T \cap B)}{\mu(B)} = \frac{\mu(A)}{\mu(\mathcal{E})} \,. \tag{10}$$

Here we used the conservation of phase space measure, namely that $\mu(A) = \mu(A_T)$. One can easily prove that a mixing system is also ergodic, but the reverse need not be true [30].

For a system of particles with the mixing property one can prove that a non-equilibrium phase space distribution will approach an equilibrium phase space distribution in a *weak*, long time limit [8]. That is, average quantities taken with respect to the non-equilibrium distribution approach equilibrium averages. The proof can easily be constructed by approximating integrals by sums in the usual way.

Mathematicians have considered proofs of the mixing property for various systems of interest to physicists. The most important of such systems are hard ball systems mentioned above, where the proofs of ergodicity and mixing are consequences of proving a stronger dynamical property, the Bernoulli property of the system, which implies that the system is mixing and therefore ergodic as well. We leave a proper definition of a Bernoulli system to the literature [34] and we will not consider it any further here. Needless to say, the class of systems which can be proved to be mixing is not yet large enough to encompass the typical systems studied in statistical mechanics or in kinetic theory, although considerable progress has been made in this direction over the past several years.

What is the Relevance of These Notions for Statistical Mechanics?

It would appear that with the proof that a system of a large number of particles in a reasonable container is mixing, the foundation for the applications of statistical mechanics to such a system would be secure. That this is not a complete explanation may be seen for the following reasons, among others:

1. We have assumed that our systems have fixed energies and that all the forces acting between the particles or on the system, due to the walls of the container, say, are conservative and known. Strictly speaking, this is not true of any laboratory system.
2. We have not examined how long it might take for the time average to be reasonably close to the ensemble average for an ergodic system or how long it would take a reasonable initial set to get uniformly mixed over the appropriate phase space, for a mixing system. One can argue that the appropriate times are very long, typically very much longer than the duration of an experiment. A partial but useful answer to this objection to the use of ergodicity and/or mixing properties is to consider Reduced distribution functions, particularly the single particle and two-particle distribution functions. These reduced distribution functions approach equilibrium, or more generally, local equilibrium forms, on much more realistic time scales.
   The notion of local equilibrium arises in the context of the decay of a non-equilibrium state in a many particle system to total equilibrium, through hydrodynamic processes. In the usual picture, due to Chapman, Enskog, and Bogoliubov [35,36], the initial state becomes, on the time scale of the duration of a collision, one that can be described by reduced distribution functions. Then on the time scale of the time between collisions, set by the microscopic properties of the system such as the density and the size of the particles, the system becomes close to a state of local equilibrium with a local temperature, density and mean velocity. Finally, on a time scale set by the physical size of the container, the system relaxes to total equilibrium. The approach to local equilibrium is set by the dynamical interactions between the particles.
   In addition, many non-equilibrium as well as equilibrium properties of a many-particle system can be formulated in terms of these reduced functions. This being the case, the questions are: What is the reason that reduced distribution functions approach local equilibrium forms, and what are the time scales involved in the approach to local equilibrium, and eventually to total equilibrium? For chaotic systems, at least, there is reason to believe that one can provide satisfactory answers to these questions.
   Another approach to resolving the issue of time scales in applying these notions to laboratory systems, and one closely related to the use of reduced distribution functions is to focus attention on a set of microscopic variables called *normal variables* [1]. These variables are defined by the requirement that they not vary much over the constant energy surface. These variables then have the property that their value is almost the same

no matter where the phase point happens to be on the constant energy surface. In this picture an initial nonequilibrium state would be one where the values of the normal variables are far from their equilibrium or average values, but the time evolution of the system leads to regions where these variables have values closer to their equilibrium averages. In this picture the relevant time scale is the average time needed for the system to evolve to regions of phase space where the normal variables are close to their average values. Presumably this time will be much shorter than the time needed for the trajectory to cover the phase space. However, we need a detailed description of these variables and a description of their behavior as a system approaches local, then total equilibrium via kinetic and hydrodynamic processes, respectively. In the absence of such an understanding of normal variables, in general, the reduced distribution functions provide a clearer way to address the questions of time scales for the approach of distribution functions to their equilibrium values.

One can imagine, that both the reduced distribution resolution and the normal variable resolution of the time scale issue become easier to justify as the number of degrees of freedom, or the number of particles become large. In either case, although the phase space measure will also grow, and the time needed to cover it will increase, the relative fluctuations in the reduced distribution functions and normal variables will become smaller. It should be emphasized that the notions of ergodicity and mixing have an important role to play in statistical mechanics, despite the issues raised here concerning the time needed for covering the full phase space. Using ergodicity one can motivate the use of the micro-canonical ensemble and ultimately all of the Gibbs ensembles that provide methods for computing equilibrium properties of systems. Similarly, the mixing property can be used to show that time correlation functions decay in time, an important property needed for many applications of non-equilibrium statistical mechanics, particularly for the Green–Kubo formalism [2,3].

3. We have used classical mechanics as our description of dynamics, but nature is fundamentally quantum mechanical.

In the following sections we will address point (2) in some detail, and point (3), briefly. It is important to note, however, that the answers we shall provide, while encouraging, are very incomplete and much more work needs to be done to make these answers believable and secure. Here we consider point (1). As mentioned in the previous section, in order to produce a random walk of the Helfand moments needed for normal transport, some sort of stochastic or stochastic-like mechanism is required. Possible sources of stochastic behavior could be random external influences or some built-in structural randomness, however subtle and difficult to identify. Evidence for the effectiveness of built-in randomness is provided by a useful model of a statistical system, such as the Ehrenfest wind-tree model [24] where diamond shaped scatterers (trees) are placed at random in a plane, with their diagonals aligned parallel to the x- and y-axes. Then point particles (wind) move among the trees with velocities only in the $\pm x$ and $\pm y$ directions. The random placement of the trees provides a clear stochastic mechanism leading to normal diffusion and an approach to equilibrium and thereby simulate a mixing system. Such systems are not chaotic and some source of randomness beyond the dynamics is necessary for normal transport in such non-chaotic models. Moreover there are periodic versions of the wind-tree model that also exhibit normal diffusion [12,25], with more subtle sources of randomness, possibly connected with the splitting of two nearby trajectories at the corners of the scatterers. Below we turn our attention to the case of chaotic systems and argue that for such systems, the dynamics can be enough to allow normal transport and the approach of an ensemble of such systems to equilibrium, even when the system is spatially periodic.

Integrable, Pseudo-chaotic, and Chaotic Systems

The kind of systems one studies in courses on classical mechanics are typically *integrable system*. These are systems with N canonical coordinates, N canonical momenta, and N independent constants of the motion [2]. The canonical coordinates of such systems can be transformed to a set of action-angle variables, and when expressed in these variables, the systems become quasi-periodic in time, and the motion is confined to regions in phase space called invariant tori. Systems of coupled harmonic oscillators are good examples of integrable systems. There is another class of mechanical systems, of which the Ehrenfest wind-tree model described above is an example, called *pseudo-chaotic* systems [37]. These systems have the property that two infinitesimally close trajectories will separate algebraically with time, for sufficiently long times. For example, two distinct trajectories moving in initially parallel directions in a wind-tree system will eventually separate algebraically with time, no matter how close they are to each other initially, due to the random placement of scatterers and the fact that one of the two trajectories will eventually hit a scatterer that is "just missed" by the other one. Pseudo-chaotic systems have a *weak mixing* property [30,37], namely that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathrm{d}\tau \left[\mu(A_\tau \cap B) - \frac{\mu(A_\tau)\mu(B)}{\mu(\mathcal{E})}\right] = 0\,. \quad (11)$$

Chaotic systems are usually defined by the condition that two infinitesimally close trajectories will, in the appropriate long time limit, separate exponentially [4]. We will give a more careful discussion of chaotic systems in the next section.

Anosov and Anosov-like Systems; the Chaotic Hypothesis

Earlier in this discussion we argued that the notions of ergodicity and mixing are not, in themselves, fully sufficient for understanding the approach to thermodynamic equilibrium for systems of many particles due to problems of time scales. In addition we argued that in many cases reduced distribution functions may very well be better indicators of the approach to equilibrium than the full phase space distribution function. We also suggested that for chaotic systems at least, these observations can be verified to the extent that we can make some more detailed statements concerning the approach to equilibrium and transport properties. Here we will provide some justification for these comments.

It is helpful to consider simple model systems that display the kind of irreversible behavior that we would like to be able to describe for more realistic, many particle systems. Two model systems that have this feature are the baker's map [31,38] and the Arnold Cat Map, which itself is an example of more general models called hyperbolic toral automorphisms [14]. We will explain this terminology as we proceed.

The baker's map is a map of a two dimensional "unit square", $0 \le x, y \le 1$ onto itself, given by (see Fig. 1)

$$\begin{aligned} \begin{pmatrix} x' \\ y' \end{pmatrix} &= \begin{pmatrix} 2x \\ y/2 \end{pmatrix} \quad \text{for } 0 \le x \le 1/2\,; \\ \text{and} \quad &= \begin{pmatrix} 2x-1 \\ (y+1)/2 \end{pmatrix} \quad \text{for } 1/2 < x < 1\,. \end{aligned} \quad (12)$$

This map consists of a stretching of the square in the x-direction by a factor of 2, and squeezing in the y-direction by a factor of 1/2. The elongated image of the unit square is then cut in half and the right side is put on top of the left side so that a unit square is reconstructed, after each application of the map. Of course this is an area preserving map, and it has a discontinuity at $x = 1/2$.

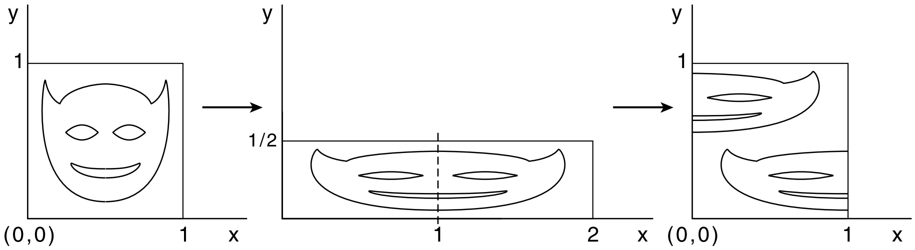
Suppose the x, y-plane is our (two-dimensional) phase space and the motion of a phase point is confined to the unit square. We now apply some of the usual techniques of statistical mechanics to baker-map distribution functions defined on the unit square. The phase space distribution function changes at each time step according to

$$\begin{aligned} \rho_n(x, y,) &= \rho_{n-1}(x/2, 2y) \quad \text{for } 0 \le y \le 1/2\,; \\ \text{and} \quad &= \rho_{n-1}((x+1)/2, 2y-1) \quad \text{for } 1/2 < y < 1\,. \end{aligned} \quad (13)$$

For statistical mechanics, we often are interested in the reduced distribution functions of fewer variables. Here there are only two variables, so we consider the reduced distribution function, $W_n(x)$ obtained by integrating $\rho_n(x, y)$ over the y coordinate. Using Eq. (13), we obtain the equation [38]

$$W_n(x) = \frac{1}{2}\left[W_{n-1}\left(\frac{x}{2}\right) + W_{n-1}\left(\frac{x+1}{2}\right)\right]. \quad (14)$$

If one assumes that the initial value, $W_0(x)$, of the reduced distribution function is a reasonably smooth function of x, for example, if it can be represented by a convergent Fourier series, then $W_n(x)$ approaches a constant value as $n \to \infty$! The reason for this is that Eq. (14) says that the



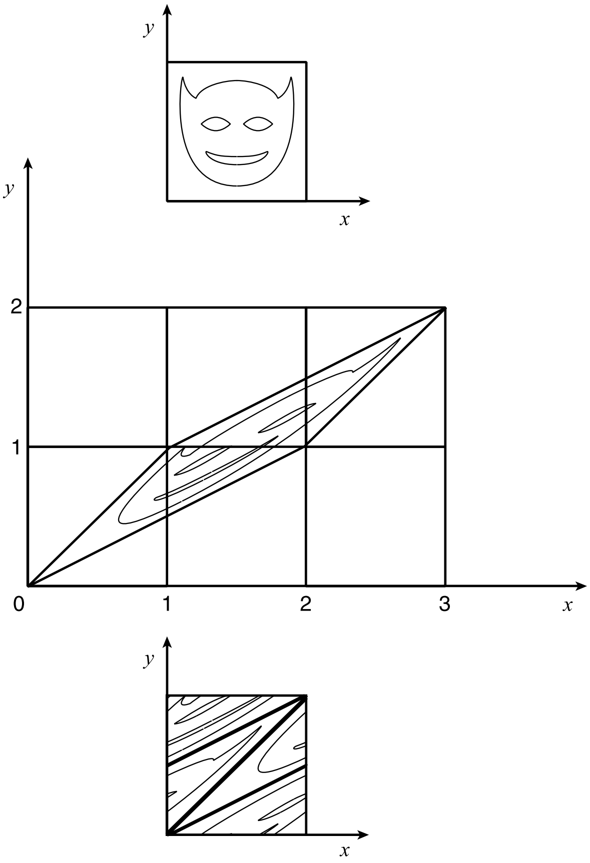
Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 1
The baker's map: The unit square is mapped onto itself by stretching by a factor *2* in the *x*-direction and by a compression by a factor of 1/2 in the *y*-direction. This is followed by cutting and re-arranging the resulting rectangle

reduced distribution at time n at a point x is the average value of the distribution at two points, $x/2$ and $(x+1)/2$ at the previous time, $n-1$. This averaging, if carried on long enough, produces a function which is constant in x. One can readily estimate the time it takes to reach this uniform state in the following way. Suppose that $l < 1$ is some length scale on which $W_0(x)$ varies with x. Then since the baker's map stretches sets of length l into sets of length $2l$ we can estimate the time to reach equilibrium as the time it takes a set of length l to be stretched to a set of length 1, which is $(-\ln l)/(\ln 2)$. This can be much shorter than the time it takes for the ergodic or mixing properties of the baker's map to make themselves manifest, which takes an additional time of order $\ln N/\ln 2$ to produce N horizontal strips of unit length, stacked in the y-direction. If we were to consider the reduced distribution function in the y variable, we would not get a nice equation with a distribution function that approaches equilibrium. Why that is so will be clear in a moment.

We have not introduced any stochastic features into our derivation of Eq. (14), but only integrated the Liouville equation over one of the variables. Thus we have been able to start from the Liouville equation, and by introducing nothing but an initial condition and integrating over the appropriate number of unmeasured variables, obtain an irreversible equation for a reduced distribution function which approaches equilibrium. Incidentally, the quantity $S = -\int_0^1 \mathrm{d}x\, W_n(x) \ln W_n(x)$ exhibits a monotonic increase with time n [8].

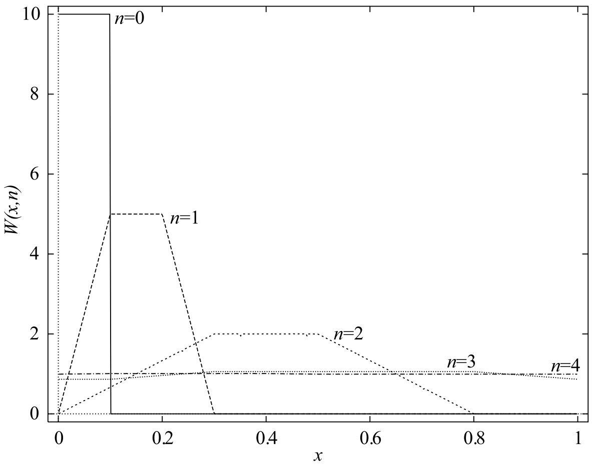
Our analysis of the approach to equilibrium of the function $W_n(x)$ made strong use of the stretching nature of the map in the x-direction. To get some further insight into the importance of the stretching of the phase space regions, we consider another model, the Arnold Cat Map [14] (see Fig. 2). Here the unit square, with opposite sides identified, represents a torus. This transformation maps the torus (that is, there is no cutting and moving of any section as there is in the baker's map) onto itself and may be described by a 2×2 matrix with unit determinant and integer elements. Such maps are called *toral automorphisms* [14], of which the Cat Map is a specific example. This map has an additional, *hyperbolic*, property, namely, that one of its eigenvalues be greater than unity. It follows from the fact that the determinant is unity, that the other eigenvalue is less than unity, and the product of the two is 1. The standard version of the map is given by the symmetric matrix

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \mathbf{T} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}, \quad \text{modulo } 1. \tag{15}$$

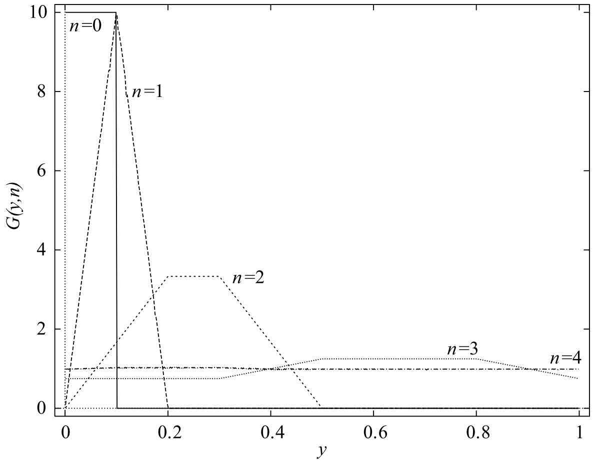


Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 2
The Arnold Cat Map of the unit torus onto itself

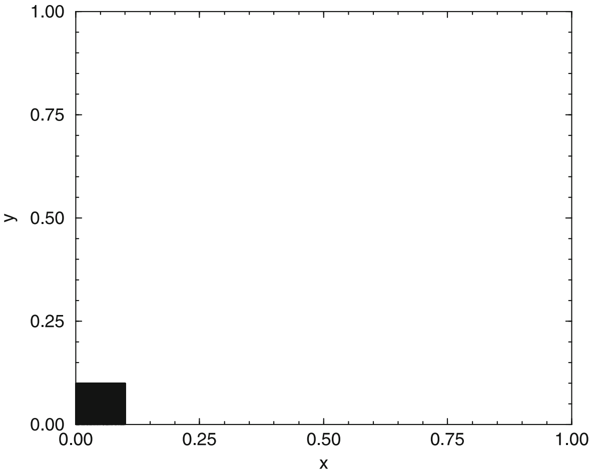
The eigenvalues of $\mathbf{T}$ are $(3 \pm \sqrt{5})/2$. The eigendirections are perpendicular to each other, and make non-zero angles with the x- and y-axes. There is a stretching direction which corresponds to the direction of the larger eigenvalue, and a contracting direction which corresponds to the smaller eigenvalue. While the equation for the projection of the "phase space" distribution function onto the x- or y-directions is not as simple as that for $W_n(x)$ given above, it is not difficult to write a computer program which shows the values of these projections after a few time steps, starting with some initial distribution on the unit torus. Figures 3 and 4 show the behavior of the projected distribution functions, $W_n(x)$ and $G_n(y)$, onto the x- and y-directions, respectively, for an initial set of points that is uniform over the region $0 \leq x, y \leq 0.1$ and zero everywhere else. Both these distribution functions become uniform after three or four time steps. In Figs. 4 and 5, we show the evolution of the phase space distribution function at some of the same times. Here we clearly see the difference on



Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 3
The projection of the Cat Map distribution onto the x-axis, $W_n(x)$, at various times, n. By $n = 3, 4$ the distribution is essentially uniform



Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 4
The projection of the Cat Map distribution onto the y-axis, $G_n(y)$ at various times, n. By $n = 3, 4$ the distribution is essentially uniform



Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 5
The initial phase space distribution used to obtain Figs. 3 and 4

the rate of evolution of projected vs. full phase space distribution functions, and why, for simple models at least, the notions of ergodicity and mixing demand more than is physically required for an approach to equilibrium for the reduced distribution functions.

The important feature that both the baker's map and the cat map have in common is that they are both area preserving maps with a stretching direction, or *unstable* direction, and a contracting direction, called a *stable* direction [14]. These directions are associated with *stretching* and *contracting* factors that depend exponentially upon time. The logarithms of these factors per unit time are the *Lyapunov exponents* [4,6,14,39]. For the baker's map the two Lyapunov exponents are $\lambda_{\pm} = \pm \ln 2$, and for the cat map above, the two Lyapunov exponents are $\lambda_{\pm} = \ln[(3 \pm \sqrt{5})/2]$. In general, for higher dimensional systems, the stable and unstable directions span the so-called stable and unstable manifolds, respectively. For the baker's map the y-axis and the x-axis are stable and unstable manifolds, respectively, for the baker's map, as are the axes in the eigen-directions of the cat map. The importance of the unstable manifolds resides in the fact that on them non-uniform functions become smoothed with time, if they are not too singular, on a time scale that is determined by the positive Lyapunov exponents. In order to get a reduced distribution function that approaches an equilibrium distribution for long enough times, one clearly must project the Liouville distribution onto manifolds that are not orthogonal to all of the unstable directions. This is one reason to examine the Cat Map, namely, to show that projections onto either direction, x or y equally lead to a uniform distribution after a few time steps. This would be the case for the baker's map, if we were to project on any direction that makes a non-zero angle with the y-direction. Generally, the unstable directions in phase space are so complex that practically any projected distribution will not be orthogonal to them. Under time reversal the unstable and stable manifolds are interchanged, and the picture remains essentially the same. It is worth mentioning that if we were to project the distribution function onto a stable manifold we would not obtain an equilibrium distribution function, but rather a very complicated function since any irregularities in the initial distribution function become more irregular with time, along stable directions [6,8,40,41]. This is why a projection onto the y-axis would not lead to an equilibrium distribution for the baker's map, without some sort of coarse graining to "smooth out" an otherwise singular function.

Dynamical systems that have properties similar to the cat map or the baker's map are called *Anosov* or *Anosov-like* systems, respectively [6,14,18]. *Anosov* systems are continuous dynamical systems such that at every point in phase space there are stable and unstable directions, with positive and negative Lyapunov exponents, respectively. There may also be some *neutral* directions with zero Lyapunov exponents, but there must be at least one positive Lyapunov exponent. The presence of positive Lyapunov exponents is characteristic of a *hyperbolic* system. Furthermore there must be at least one trajectory in phase space that is dense, a requirement which is called *transitivity*. Systems of particles with some manageable singularities such as hard spheres are not strictly Anosov systems but are close enough to be considered *Anosov-like*. The Arnold Cat Map is an example of an Anosov system, while the baker's map is Anosov-like. Of course in a multidimensional Anosov system, there are a number of stable and unstable directions, and all stable manifolds intersect all unstable manifolds transversally. Systems with at least one positive Lyapunov exponents are commonly referred to as being *chaotic*. We have chose simple systems to illustrate the main ideas, but the reader should be aware that it is possible to define Lyapunov exponents for motion on invariant sets in phase space, even for sets of measure zero, such as periodic orbits, or as we discuss below, fractal repellers.

Returning to statistical mechanics, we see that the simple examples of the baker's map and the Cat Map give us some reason to believe that, for chaotic systems at least, reduced distribution functions will approach equilibrium values, or more precisely, local equilibrium values, on reasonable time scales. This conclusion depends, of course, on the assumption that the underlying microscopic dynamics is of the Anosov or Anosov-like variety. It is therefore useful to assume that systems of large numbers of particles with short ranged interactions, of general interest for statistical mechanics, are ergodic, Anosov-like systems as far as their dynamical properties are concerned. This assumption has been made a central feature of the dynamical systems approach to non-equilibrium statistical mechanics by Gallavotti and Cohen [18], who have called it *the chaotic hypothesis*. This hypothesis allows one to apply the results of chaotic dynamics, such as those discussed in the next section, to realistic systems. In actuality we know little a priori about the dynamics of such systems, such as their Lyapunov exponents, or the structure and properties of unstable and stable manifolds in phase space. Furthermore, due to the work of Turaev, Rom-Kedar, Donnay [28,29], and others, we know that if the intermolecular potential is smooth, there may exist regions in phase space where the dynamics is not chaotic. For these systems, the chaotic hypothesis implies the additional statement that the total volume of the non-chaotic regions in phase space is small enough so that for determining the important averages, they can be ignored.

Fractal Dimensions

An important concept used in applications of dynamical systems theory to statistical mechanics, among many other applications, is the notion of *fractal dimensions*. As we will mention in the next section, for chaotic systems many of the connections between macroscopic transport coefficients and microscopic dynamical quantities such as Lyapunov exponents can be related to the dimensions of fractal structures that form in the phase spaces of nonequilibrium systems. Generally, but not exclusively, fractals are defined by the statement that their dimension is not an integer. However, this definition is complicated by the fact that there are many ways to define the dimension of a set, and some fractals may have a dimension that is an integer by some definition and non-integer by others. Fractals that have a variety of different dimensions are referred to as *multifractals*. There are also other definitions of fractals based upon self-similarity at all scales, or upon differentiability and continuity properties. We refer to the literature for more details [42].

In order to present one definition of the fractal dimension of a set, imagine that the set is embedded an a d-dimensional space. Cover the set with cubes of length ϵ on a side, each labeled by an index i and suppose that one needs at least $N(\epsilon)$ of these cubes for a full coverage. Suppose further that we have a way to assign a measure μ_i to each cube. We can suppose that the total measure is normalized to some constant, say unity,

$$\sum_{1}^{N(\epsilon)} \mu_i = 1 \,. \tag{16}$$

The measure μ_i may be a normalized Lebesgue measure of the cube, or it may be the fraction of time that a typical trajectory on the fractal spends in cube i, for example. In terms of this measure we can define a dimension D_q that depends on a continuous variable q, as

$$D_q = \frac{-1}{1-q} \lim_{\epsilon \to 0} \frac{I(q,\epsilon)}{\ln \epsilon} \,, \tag{17}$$

where $I(q,\epsilon)$ is given by

$$I(q,\epsilon) = \sum_{i=1}^{N(\epsilon)} \mu_i^q \,. \tag{18}$$

In this construction the quantity, D_0 is called the *box-counting dimension*, D_1 is called the *information dimension*, and D_2 is called the *correlation dimension*. The dimension D_q is a monotonic non-decreasing function of q. Other dimensions frequently employed to characterize fractals include the *Hausdorff dimension*, and we refer to the literature for a careful definition and discussion of this and other dimensions. As an example of a fractal, consider the *middle third Cantor Set*. This set is constructed by taking a unit interval, discarding the middle third of it, then take each of the two remaining pieces, discarding the middle third of each of them, and so on. It is easy to show that the dimensions, D_q, of the remaining set, are all equal to $\ln 2/ln 3$.

Applications of Dynamical Systems Theory to Non-equilibrium Statistical Mechanics

The connections between normal transport, random walk processes, and the generalized Einstein formulae have allowed results from dynamical systems theory to be applied to non-equilibrium statistical mechanics in a direct way, at least for chaotic systems. However, one must study transport processes in a way that takes advantage of the fundamental properties of chaotic dynamics. Among the useful properties of chaotic systems are the formation of fractal structures in phase space. If the phase space is of sufficiently low dimensions then the fractal structures can be studied both by analytical methods and by computer simulated molecular dynamics. In addition to providing some insights into the approach to equilibrium for large systems, the dynamical systems approach gives us deep results of a more practical sort. Among these results are: (i) relations between transport coefficients and dynamical quantities, such as Lyapunov exponents and Kolmogorov–Sinai entropies, to be defined below [5,6,7,8,9,12,42,43,44,45,46]; (ii) a theory of entropy production in non-equilibrium steady states and in the approach of a system to equilibrium [5,6,7,8,9,10,16,17,18,20,40,49,50,51,52,53,54]; and (iii) a number of fluctuation theorems – Evans–Searles–Gallavotti–Cohen theorems – which describe fluctuations in entropy production in non-equilibrium steady states [16,17,18,19,20,21,52,55]. Here we will summarize these results, and we refer the reader to the literature for a more complete discussion of these and related topics.

Microscopic Dynamical Quantities and Macroscopic Transport Coefficients

Among the many reasons dynamical systems theory attracted the attention of workers in non-equilibrium statistical mechanics were the connections between macroscopic transport coefficients of a system of particles and quantities that characterize the system's chaotic behavior discovered by Gaspard and Nicolis [6,8,43,44,45,46,47], for Hamiltonian systems, and by Evans, Morriss, Hoover, Posch, and coworkers, for dissipative systems with Gaussian thermostats [5,7,9,12,48,49]. We discuss each case separately.

The Escape-Rate Formulae for Transport Coefficients Previously we noted that the Helfand moments related to transport undergo a random walk motion in phase space. Gaspard and Nicolis pointed out that if one considers a random walk in a space with *absorbing* boundaries one can combine results from random walk theory and from dynamical systems theory to obtain very interesting previously unknown connections between transport coefficients and quantities that characterize the microscopic chaotic dynamics.

The random walk theory that one needs for this connection is based on a Fokker–Planck equation for the probability distribution of the Helfand moments, $P(M_\sigma, t)$, where M_σ is a Helfand moment. We assume that the transport process is normal, that is, that the mean square fluctuations of the Helfand moment grows linearly with time. In such a situation the Fokker–Planck equation takes the simple form of a diffusion equation:

$$\frac{\partial P(M_\sigma, t)}{\partial t} = \alpha \frac{\partial^2 P(M_\sigma, t)}{\partial M_\sigma^2} \tag{19}$$

Here α is a constant proportional to the transport coefficient σ. Suppose now that we solve this diffusion equation in M-space with boundary conditions that when M_σ reaches the values $\pm\chi/2$, $P(M_\sigma, t) = 0$. That is, the Helfand moments undergo a Brownian motion in a space with absorbing boundaries such that the probability distribution vanishes whenever $|M_\sigma|$ reaches a specific value for the first time. Under these conditions, the probability distribution will decay exponentially, with a decay rate γ

given by

$$\gamma = \alpha \left(\frac{\pi}{\chi} \right)^2 . \tag{20}$$

A remarkable result from dynamical systems theory is that for chaotic systems, another expression holds for the same escape rate in terms of Lyapunov exponents and a quantity called the Kolmogorov–Sinai entropy per unit time [6,14,56,57]. This microscopic escape-rate formula for the escape-rate, γ_{mic} is

$$\gamma_{\text{mic}} = \sum_i \lambda_i^+(\mathcal{R}) - h_{\text{KS}}(\mathcal{R}) . \tag{21}$$

All of the terms on the right hand side of Eq. (21) require some explanation. To get some insight into this formula, we now consider the microscopic dynamics of the system of particles and imagine that there is a set of initial conditions for all the particles for which the Helfand moment is in the region $-\chi/2 < M_\sigma < \chi/2$, and such that as the initial system evolves in time, either forward or backward, the Helfand moment will never reach the boundary. The set of all such initial conditions is called a *repeller* and denoted by $\mathcal{R}$. It is typically a fractal set of points in phase space, usually of measure zero in the set of all initial phase points, and a highly unstable set since an arbitrarily small displacement of the initial phase of the system from a point in $\mathcal{R}$ will lead to escape, unless the new initial point is also on the repeller. Now imagine that we consider an infinitesimally small set of points in $\mathcal{R}$, and observe how these points separate in phase space in the course of time. If the dynamics on $\mathcal{R}$ is Anosov-like, there will be a set of positive Lyapunov exponents which we have denoted by $\lambda_i^+(\mathcal{R})$ in Eq. (21). The positive Lyapunov exponents on the repeller describe the rate with which trajectories on the repeller move apart. apart. The other term in the escape-rate formula, $h_{\text{KS}}(\mathcal{R})$, is the Kolmogorov–Sinai entropy per unit time of the trajectories on $\mathcal{R}$, and can be understood in the following way. In general, dynamic entropies characterize the rate at which information about the exact trajectory of a system is gained or lost in the course of time. For example, if we know that the initial phase of a system is within some small region of phase space, then the *stretching* of the small region with time due to the dynamical instability allows us to locate the initial location of the trajectory ever more precisely as we follow the motion of the small initial set. If the system is closed, that is, there is no possibility of escape, then the amount of information about the initial location of the trajectory grows exponentially as the sum of the positive exponents. This result is known as Pesin's theorem. On the other hand, if the system is open and there is a possibility of escape, then the escaping trajectories lead to a loss of information [58]. Hence a better way to write Eq. (21) is

$$h_{\text{KS}}(\mathcal{R}) = \sum_i \lambda_i^+(\mathcal{R}) - \gamma_{\text{mic}} . \tag{22}$$

In any case we now have two expressions for the same escape-rate for a hyperbolic system, and by equating them, we obtain an expression for the transport coefficient in terms of microscopic dynamical quantities, as

$$\alpha = \lim_{\chi \to \infty} \frac{\chi^2}{\pi^2} \left[\sum_i \lambda_i^+(\mathcal{R}) - h_{\text{KS}}(\mathcal{R}) \right] . \tag{23}$$

We have taken the large system limit to remove any possible dependence of the right hand side of Eq. (23) on the shape of the boundaries or on the size of the system. This result is due to Gaspard and Nicolis [43].

For two dimensional systems, one can reformulate the escape rate formula in terms of the information dimension, d_1, of the repeller along the unstable direction [6,59], which is given by $d_1 = h_{\text{KS}}/\lambda^+$, so that $\gamma = \lambda^+(1 - d_1)$.

The escape-rate formula has been used in molecular dynamics studies to obtain values for transport coefficients in a number of cases such as diffusion coefficients in periodic Lorentz gases, viscosities of simple systems and chemical reaction rates [27,47,60,61,62]. The results for the transport coefficients obtained by using the escape-rate methods agree with those obtained by other methods, often based on the Green–Kubo time correlation method, or, as discussed below, on Gaussian thermostat methods. There are as yet very few theoretical methods to compute the Kolmogorov–Sinai entropy for the repeller [6,27,47], while both the transport coefficients, and, as it turns out, the sum of the positive Lyapunov exponents on the repeller are sometimes amenable to treatment by the usual methods of statistical mechanics and kinetic theory [63,64,65].

Transport Coefficients and Phase-Space Contraction in Gaussian Thermostatted Systems An alternative method for relating macroscopic transport coefficients to Lyapunov exponents was developed by Evans, Hoover, Posch and co-workers in their work on developing computer algorithms to simulate non-equilibrium flows in many-particle systems [5,7,9,48,49,66,67,68,69,70,71,72,73,74,75]. A problem arose in these simulations because the systems tended to heat up considerably due to the presence of viscous friction or, for charged particle systems, ohmic heating. To counteract this heating, these authors introduced a fictitious thermostat which maintains a constant value for either the total energy or the total kinetic energy of the system. The thermostat was introduced by a modification of the equations of motion by the addition of a frictional force in such a way that the Hamiltonian nature of the system, in the usual coordinate, momentum, and time variables, is lost. This kind of frictional force was first introduced by Gauss in his study of mechanical systems with various constraints.

For systems with Gaussian thermostats, Liouville's theorem, which states that the phase space density at a point moving under the equations of motion does not change with time, is no longer satisfied. It is replaced by a conservation equation that relates the time derivative of the distribution function to the parameters characterizing the frictional forces. As a consequence phase space volumes for thermostatted systems do not remain constant in time, but rather, on the average the phase space volume of a set decreases with time, the system approaches a *non-equilibrium* stationary state, and the system's distribution function approaches a distribution with fractal properties. The average rate of decrease of the phase space volume is given by a negative value of the sum of all of the Lyapunov exponents of the system. The non-equilibrium stationary state distribution is characterized by a special type of measure, called a Sinai–Ruelle–Bowen (SRB) measure, which describes the probability of finding a system in different regions of phase space. The SRB measure is characterized by the fact that it is smooth in the unstable directions in phase space but typically is fractal in the stable directions [14]. Simple examples of this type of measure can be found in the literature [41]. The resulting fractal is called an *attractor*.

It is important to note that the decrease of the phase space volume and the concomitant formation of an attractor with a non-trivial SRB measure, does not mean that the fractal has zero dimension. Indeed due to the decrease in phase space volume the *Lebesgue measure* of the phase space region where the trajectories are located approaches zero. However the dimension of the resulting fractal is not zero. The box counting dimension, D_0, of the fractal may even coincide with the dimension of phase space, and the information dimension, D_1, can often be expressed in terms of the Lyapunov exponents. For example for two-dimensional, thermostatted systems, with one positive Lyapunov exponent, λ^+ and one negative exponent, $-|\lambda^-|$, with $\lambda^+ - |\lambda^-| < 0$, the information dimension of the attractor is given by the *Kaplan–Yorke–Young formula* [4,42,69,71] for two dimensional ergodic systems, as $D_1 = 1 + \lambda^+/|\lambda|$.

The decrease of the phase space volume can also be related to the entropy production in the system+thermostat. This entropy production can also be related to the transport coefficients via the usual macroscopic laws of irreversible thermodynamics. Thus one has a way to relate transport coefficients to the sum of the Lyapunov exponents for these thermostatted systems.

To make this discussion more concrete, we consider the example of a hard-ball Lorentz gas [53,69,70,74,75]. We assume that the moving particles have a bounded free path between collisions with the fixed scatterers, and that the scatterers do not form traps without escape for the moving particles. We suppose the moving particles have mass. m, and carry a charge, q, (but still do not interact with each other) and are placed in a constant, external electric field, $\boldsymbol{E}$. Ordinarily, the field will accelerate the moving particles in such a way that their energy increases over time. To avoid this we add a frictional force to the equations of motion of the moving particle, and adjust the frictional force so as to keep the kinetic energy of the moving particle constant in time. The equations of motion for the position, $\boldsymbol{r}$ and velocity, $\boldsymbol{v}$, of the moving particle between collisions with scatterers are

$$\dot{\boldsymbol{r}} = \boldsymbol{v} \tag{24}$$

$$\dot{\boldsymbol{v}} = \frac{q\boldsymbol{E}}{m} - \alpha\boldsymbol{v}\,, \tag{25}$$

where α is a dynamical variable defined by the constant energy condition, $\dot{\boldsymbol{v}} \cdot \boldsymbol{v} = 0$. Thus we find that

$$\alpha = \frac{q\boldsymbol{v} \cdot \boldsymbol{E}}{m\boldsymbol{v}^2}\,. \tag{26}$$

The equations of motion given above for the moving particle must be supplemented by the equations describing the collisions of the moving particles with the scatterers. To simplify matters, we will suppose that the particles make instantaneous, elastic, and specular collisions with the scatterers.

If we define the Gibbs entropy for this system in terms of the phase space distribution function, $\rho(\boldsymbol{r}, \boldsymbol{v}, t)$, by

$$S_G(t) = -k_{\mathrm{B}} \int \mathrm{d}\boldsymbol{r} \int \mathrm{d}\boldsymbol{v}\rho(\boldsymbol{r}, \boldsymbol{v}, t)[\ln \rho(\boldsymbol{r}, \boldsymbol{v}, t) - 1]\,, \tag{27}$$

where k_{B} is Boltzmann's constant, one finds that

$$\frac{\mathrm{d}S_G}{\mathrm{d}t} = -k_{\mathrm{B}} \int \mathrm{d}\boldsymbol{r} \int \mathrm{d}\boldsymbol{v}\alpha\rho = -k_{\mathrm{B}}\langle\alpha\rangle\,. \tag{28}$$

The average value is taken with respect to the phase space distribution function for the moving particles, ρ. This entropy actually decreases with time, since on the average α is positive. This decrease in entropy must be matched, at least, by an increase in entropy of the reservoir that is responsible for the additional frictional force. Thus

$$\frac{\mathrm{d}S_{\text{reservoir}}}{\mathrm{d}t} \geq k_{\mathrm{B}}\langle\alpha\rangle \geq 0\,. \tag{29}$$

We now take the entropy production in the reservoir to be given by the usual macroscopic laws, in particular,

$$\frac{\mathrm{d}S_{\text{reservoir}}}{\mathrm{d}t} = \frac{\boldsymbol{J}\cdot\boldsymbol{E}}{T} = \frac{\sigma\boldsymbol{E}^2}{T}\,, \tag{30}$$

where $\boldsymbol{J} = \sigma\boldsymbol{E}$ is the electrical current produced by the moving particles and σ is the coefficient of electrical conductivity. Then, by combining Eqs. (29) and (30), and assuming that we take the entropy production in the reservoir to be exactly $k_{\mathrm{B}}\langle\alpha\rangle$, we obtain and expression for σ as

$$\sigma = \frac{k_{\mathrm{B}}T\langle\alpha\rangle}{\boldsymbol{E}^2}\,. \tag{31}$$

The final step in this process is to relate the average friction coefficient, $\langle\alpha\rangle$, to the Lyapunov exponents for the trajectories of the moving particles. We note, first of all, that the volume, $\mathcal{V}$ of a small region in phase space will change exponentially in time with an exponent equal to the sum of all the Lyapunov exponents, as

$$\mathcal{V}(t) = \mathcal{V}(0)\exp\sum_i \lambda_i\,, \tag{32}$$

since the Lyapunov exponents describe the rates of stretching or of contracting of small distances in phase space. A simple argument shows that

$$\left\langle\frac{\mathrm{d}\ln\mathcal{V}}{\mathrm{d}t}\right\rangle = -\langle\alpha\rangle = \left\langle\sum_i \lambda_i\right\rangle\,. \tag{33}$$

In the non-equilibrium stationary state, all averages remain constant with time, so that we can now use Eq. (31) to obtain the desired connection between a transport coefficient, in this case the conductivity σ, and the average Lyapunov exponents,

$$\sigma = -\frac{k_{\mathrm{B}}T}{\boldsymbol{E}^2}\left\langle\sum_i \lambda_i\right\rangle\,. \tag{34}$$

This result has been used to determine the conductivity and the diffusion coefficient of the moving particles in the Lorentz gas by means of efficient computer algorithms for determining the Lyapunov exponents [70,74,75]. Similar methods have been used to determine the shear viscosity [48] and other transport properties of systems with short range, repulsive potentials.

One important result of the analysis of these thermostatted systems is the *conjugate pairing rule*. This is a generalization of the result for chaotic Hamiltonian systems that non-zero Lyapunov exponents come in pairs, with the same magnitude but with opposite signs so that members of each conjugate pair sum to zero [42]. This is called the *symplectic conjugate pairing rule*. For many chaotic systems with Gaussian thermostats it is possible to prove another conjugate pairing rule where the Lyapunov exponents form conjugate pairs which sum to a non-zero value, independent of which pair of exponents is chosen [48,76,77]. The conjugate pairing rule is very helpful for analyzing relations such as that given by Eq. (34), since the right hand side is completely determined by the sum of any one conjugate pair of exponents. The easiest pair to use is the pair formed by the positive and negative exponents with the largest magnitude.

Ruelle–Pollicott Resonances and Irreversible Processes

For a Hamiltonian N-particle system, the phase space distribution function, $\rho(\Gamma, t)$ evolves in time according to the relation given by the Liouville equation, in terms of a time evolution operator, $S(-t)$, as

$$\rho(\Gamma, t) = S(-t)\rho(\Gamma, 0) = \exp - t\mathcal{L}\rho(\Gamma, 0) = \rho(\Gamma(-t), 0), \tag{35}$$

where we indicate that the phase space variables Γ evolve as

$$\Gamma(t) = S(t)\Gamma = \exp t\mathcal{L}\Gamma\,. \tag{36}$$

Here $\mathcal{L}$ is the Liouville operator, when acting on differentiable functions, is given, for an isolated system of N particles, by

$$\mathcal{L} = \sum_{i=1}^{N} \boldsymbol{p}_i\cdot\frac{\partial}{\partial\boldsymbol{r}_i} + \sum_i^N \boldsymbol{F}_i\cdot\frac{\partial}{\partial\boldsymbol{p}_i}\,, \tag{37}$$

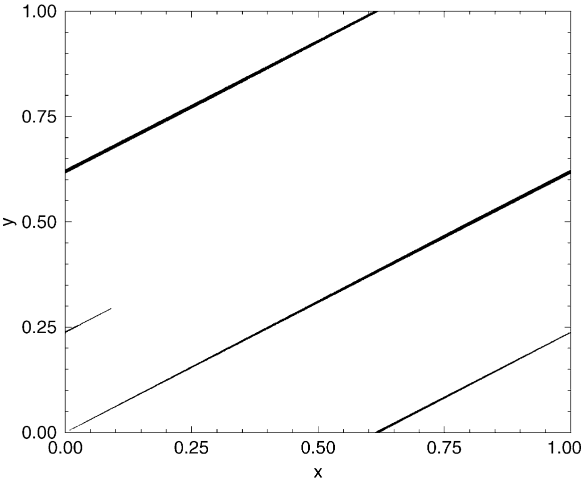
where $\boldsymbol{F}_i$ is the total force on particle i due to the other particles, and external sources such as walls of a container.

If one insists that the functions on which the time displacement operator acts be ordinary functions with well behaved derivatives, the exponential operators, S appearing in Eqs. (35), (36) are unitary with spectrum on the unit circle. In view of this observation it is difficult to see how a satisfactory theory of irreversible behavior, including decays in time, might be obtained from such an operator. The answer is to be found by enlarging the space of functions to include singular functions such as Schwartz distributions [6]. The discussion in the previous subsection has indicated that distributions may evolve to fractals, which are typically non-differentiable functions, so it is very natural to include singular functions in the space of functions on which the time evolution operator may act. In this enlarged function space, the time evolution operator may have properties that do not obtain in a more restricted space of functions. Under these circumstances one studies the right and left eigenfunctions of the Liouville operator, $\mathcal{L}$ in a structure called a *Gelfand triplet* or *rigged Hilbert space* [78]. There are two sets of such triplets corresponding to forward and time reversed motion.

A Gelfand triplet is an operator, together with the two spaces spanned by its right and left eigenfunctions, and an inner product involving one function from each of the two spaces. If the right eigenfunctions are singular functions, the left eigenfunctions must be sufficiently well behaved so that the inner product of a function in the "singular" space with a function in the "smooth" space is well defined. The eigenvalues for the time displacement operator appear as two different sets of poles of the resolvent operator, $(z - \mathcal{L})^{-1}$, in the complex plane, one for forward and one for time-reversed motion. The poles are called *Ruelle–Pollicott* resonances [79,80,81,82], and they give the relaxation rates for various processes that can take place in the system. The time reversal symmetry of the motion provides the relation between the two sets of resonances in the complex plane. For simple models it is possible to construct the functions, singular and smooth, in a Gelfand triplet and to locate the resonances in the complex plane [83,84,85]. In addition to poles of the resolvent one may, in general, expect to find branch cuts, etc. in the complex plane [6]. The existence of these resonances provides an argument that the rates of relaxation to equilibrium found by using traditional methods of statistical mechanics, including kinetic theory, are not artifacts of the approximations made in arriving at these results, but are reflections of the existence of Ruelle–Pollicott resonances for the system.

Fractal Forms in Diffusion and a Dimension Formula

For chaotic systems, the relation between the mean square displacement of a diffusing particle and the diffusion coefficient conceals a fractal function. This can be understood on the basis of the following observation: Since the actual displacement of the diffusing particle depends on the initial phases of the particles in the system, even the slightest change in these phases will make a large change in the displacement of the Brownian particle. One way to try to capture some features of this variation of the displacement is to consider simple models of diffusion that are deterministic, chaotic, and diffusive [86,87]. The process of diffusion in these models is often referred to as deterministic diffusion, since no stochastic elements are introduced in the dynamics of the models. Here we illustrate this idea using the two dimensional periodic Lorentz gas with finite free paths of the moving particle between collisions, so that diffusion is well defined in this system [86,87,88].



Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 6

The evolution of the phase space distribution of Figure 5 after three iterations of the Arnold Cat Map

It is very convenient to use the van Hove intermediate scattering function as a way of describing the diffusive process. This function, denoted by $F_{\boldsymbol{k}}(t)$, is defined as

$$F_{\boldsymbol{k}}(t) = \left\langle \mathrm{e}^{[i\boldsymbol{k}\cdot(\boldsymbol{r}(t)-\boldsymbol{r}(0))]} \right\rangle , \tag{38}$$

where the average is taken over an equilibrium ensemble. The van Hove function is related to the probability of finding the diffusing particle at point $\boldsymbol{r}$ at time t, $P(\boldsymbol{r}, t)$ by

$$P(\boldsymbol{r}, t) = \int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^{\mathrm{d}}} P_{\boldsymbol{k}} F_{\boldsymbol{k}}(t) , \tag{39}$$

where d is the number of spatial dimensions of the system and $P_{\boldsymbol{k}}$ is the Fourier transform of the initial probability distribution for the diffusing particle. One can use a cumulant expansion in the exponent to show that $F_{\boldsymbol{k}}(t)$ is given by

$$F_{\boldsymbol{k}}(t) = \mathrm{e}^{s(k)t} , \tag{40}$$

where $s(k)$ is the wave-number dependent decay rate for diffusive motion. Thus, for wave numbers, k much smaller than the inverse of a lattice spacing, in this case, unity, this function takes the form

$$s(k) = -Dk^2 + \tilde{D}k^4 + O(k^6) \,, \tag{41}$$

where D is the diffusion coefficient, $\tilde{D}$ is called the super-Burnett diffusion coefficient, etc. In order to see the physics that is obscured by taking the equilibrium average, consider the microscopic quantity that is averaged, namely, $\exp[i\boldsymbol{k} \cdot (\boldsymbol{r}(t) - \boldsymbol{r}(0))]$. If the motion of the particle is chaotic, the displacement over a time t is a very rapidly varying function of the initial point $\boldsymbol{r}(0)$. As a result, the exponential containing this displacement will be a rapidly oscillating function of the initial point. To capture these oscillations and to explore the fractal structure of the exponential function, we consider a partially averaged quantity, which we will call a normalized, incomplete van Hove function, $F_{\boldsymbol{k}}(\theta, t)$, defined by

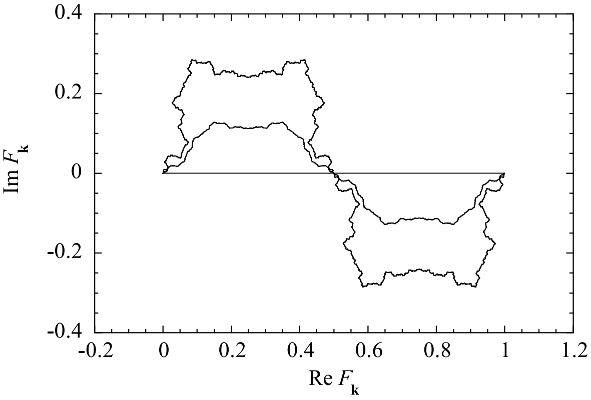
$$F_{\boldsymbol{k}}(\theta, t) = \frac{\int_0^{\theta} \mathrm{d}\theta' \mathrm{e}^{[i\boldsymbol{k}\cdot(\boldsymbol{r}(\theta',t)-\boldsymbol{r}(\theta',0))]}}{\int_0^{2\pi} \mathrm{d}\theta' \mathrm{e}^{[i\boldsymbol{k}\cdot(\boldsymbol{r}(\theta',t)-\boldsymbol{r}(\theta',0))]}} \,. \tag{42}$$

Here we take initial points to be uniformly distributed just outside the surface of one of the scatterers in the periodic Lorentz gas, with initial velocity directed radially outward. The point on the surface is indicated by the angle θ taken with respect to some fixed direction. If one waits for a sufficient number of collisions to take place so that the motion is diffusive, and then plots $\mathrm{Im}\, F_{\boldsymbol{k}}(\theta, t)$ vs $\mathrm{Re}\, F_{\boldsymbol{k}}(\theta, t)$ one finds a fractal curve for small values of the wave number, $|\boldsymbol{k}|$. Computer results obtained by Claus et al. [89] for the hard disk Lorentz gas are plotted in Fig. 7 for various values of the wave number.

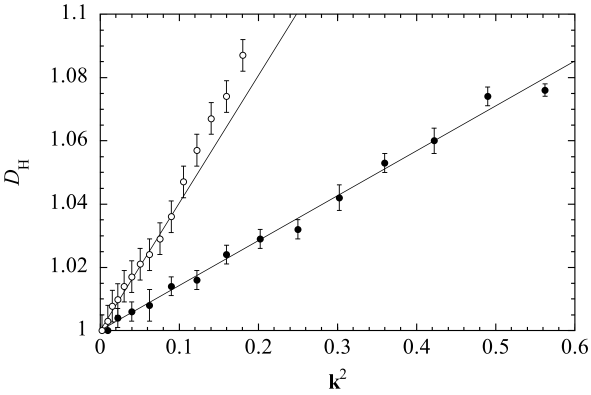
It is possible to prove that there exists a striking connection between the fractal Hausdorff dimension, $\mathcal{D}_H$, of the curve $(\mathrm{Re}\, F_{\boldsymbol{k}}(x), \mathrm{Im}\, F_{\boldsymbol{k}}(x))$, for the incomplete van Hove function [87,88], for small k, the diffusion coefficient, D, and the positive Lyapunov exponent, λ, characterizing the chaotic process underlying the diffusive motion of the particle. This connection is illustrated in Fig. 8 and given by the equation

$$\mathcal{D}_H(k) = 1 + \frac{D}{\lambda}k^2 + O(k^4) \,. \tag{43}$$

Also illustrated in Fig. 8 is the analogous curve obtained from the incomplete van Hove function when the hard disk potential is replaced by a repulsive Coulomb potential. For sufficiently high energy of the moving particle, the motion of the moving particle is chaotic and the van Hove function also shows fractal properties. Thus the fractal curves are not artifacts of the hard disk potential. Equation (43) illustrates the fact that for chaotic models such as the one discussed here, the incomplete van Hove function encodes in its structure both a macroscopic property of the system, the diffusion coefficient, and a microscopic property, the Lyapunov exponent. Such interesting connections are central to a deeper understanding of the microscopic foundations of transport processes.



Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 7
The incomplete van Hove function for a periodic Lorentz gas where a point particle of unit mass and velocity undergoes elastic collisions with hard disks of unit radius forming a triangular lattice with interdisk distance $d = 2.3$: Curves of the cumulative functions for wavenumber $k_x = 0.0$, 0.5, and 0.9 with $k_y = 0$. Note that the fractality increases with the wavenumber



Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 8
Hausdorff dimension D_H of the incomplete van Hove function versus $k^2 = k_x^2$ ($k_y = 0$) for both periodic Lorentz gases with hard-disk scatterers (filled circles) and for the case where the hard disks are replaced by repulsive Coulomb scatterers on a square lattice, and the energy of the particle is sufficiently high for the motion to be chaotic (open circles). Both solid lines have slopes equal to $\mathcal{D}/\lambda$ for the respective diffusion coefficient $\mathcal{D}$ and Lyapunov exponent λ of the Lorentz gases

Entropy Production in Non-equilibrium, Hamiltonian Systems

One of the subjects of most active research in recent years has been the theory of entropy production in non-equilibrium systems. Dynamical systems theory has provided some new insights into this old problem. Briefly formulated, the problem is to explain the positive, irreversible production of entropy using the fundamental ideas in statistical mechanics, particularly the Liouville equation. The obstacle in this direction that must be overcome somehow is the fact that the Gibbs entropy

$$S_G(t) = -k_{\mathrm{B}} \int \mathrm{d}\Gamma \rho(\Gamma, t)[\ln \rho(\Gamma, t) - 1] \,, \tag{44}$$

remains constant in time if $\rho(\Gamma, t)$ satisfies the Liouville equation for Hamiltonian systems

$$\frac{\mathrm{d}\rho(\Gamma, t)}{\mathrm{d}t} = 0 \,. \tag{45}$$

Here Γ represents all of the coordinate and momentum variables of the system. The usual way around this difficulty is to introduce a coarse grained entropy, obtained by either defining an entropy in terms of reduced distribution functions, as is done in Boltzmann's famous H-theorem, or by coarse graining the phase space itself, and defining an entropy in terms of the average phase space distribution in each of the coarse graining regions [9]. In either case, it is possible to show that the rate of production of the redefined entropy is positive, and for the Boltzmann case at least, is in agreement with the predictions of irreversible thermodynamics when the gas is close to a local equilibrium state.

Ideas from dynamical systems theory have not changed the fundamental need for coarse graining. Instead they have provided strong reasons for doing it, reasons that were lacking in the previous approaches, where the motivation for coarse graining seemed only to be that it was necessary to coarse grain in some way to get a positive entropy production. The central new idea in this area is that the phase space description of a non-equilibrium process requires, in the thermodynamic limit at least, the use of distribution functions that are defined on fractal sets. A simple example is provided by a system in which diffusion can take place in a region between two particle reservoirs, each reservoir being maintained at a different density of particles [6,40]. Then if the dynamics that leads to transport of the particles between the reservoirs is Anosov-like, one can argue, and in simple enough cases show explicitly, that the regions in phase space that correspond to regions of different density in the system get so tangled up and enmeshed that the distribution function is a wildly varying fractal function. As such, it is no longer differentiable and the steps that lead to the proof of the constancy of the Gibbs entropy can no longer be justified. The only way to treat this kind of fractal behavior of the distribution function is to smooth the function in one or another way, typically by defining cumulative distribution functions over small regions of phase space, leading to the construction of SRB measures. In simple cases, one can show that fractal forms appear in the phase space distribution function for systems in non-equilibrium stationary states produced by particle reservoirs, or in the relaxation of a system with particle diffusion to equilibrium, and that the rate of entropy production agrees with the predictions of irreversible thermodynamics [6,40,50,90,91]. However, much needs to be done to extend this work to other hydrodynamic processes and to understand why the results of macroscopic theory are obtained in this way. Recently, Gaspard has presented an expression for the rate of entropy production in chaotic as well as in stochastic systems as the difference in two dynamical entropies per unit time, one a form of the Kolmogorov–Sinai entropy per unit time, and the other, an entropy per unit time for the time reversed motion but as measured by the measure of the forward process. It is possible to prove that this rate of entropy production is positive, and for the cases studied so far, agrees with the results of irreversible thermodynamics [51,52].

Kinetic Theory Methods for Analytical Calculations of Lyapunov Spectra

As we demonstrated in the examples given above for the applications of chaos theory to non-equilibrium processes, an important role is played by Lyapunov exponents and Kolmogorov–Sinai entropies per unit time. Apart from computer simulations, it is not always clear how one might obtain expressions and values for these quantities. In some simple cases such as the baker's map and toral automorphisms, the calculation of the Lyapunov exponents is a simple matter. However these cases are very special and rarely are realistic models for physical systems. Thus the problem remains of finding general methods for an analytic determination of Lyapunov exponents and Kolmogorov–Sinai entropies. One such method is kinetic theory. Using the Boltzmann and related equations [8,35], van Beijeren and co workers have been able to obtain expressions for Lyapunov exponents, and in some cases, Kolmogorov–Sinai entropies, for dilute gases with short range potentials, such as hard ball systems, and for dilute Lorentz gases [8,63,64,65,92,93]. Results obtained this way are in good agreement with the results of computer simulations. We refer to the literature for details.

Entropy Production in Systems with Gaussian Thermostats and the Cohen–Gallavotti Fluctuation Theorem

As a final example of the applications of dynamical systems theory to non-equilibrium statistical mechanics we briefly discuss one example of a number of closely related fluctuation theorems, first obtained by Evans, Searles, and by Gallavotti, and Cohen, and generalized by Spohn and Lebowitz, among others [16,17,18,19,20,21]. These fluctuation theorems are very closely related to a class of work-free energy fluctuation results due to Jarzynski [94], and Crooks [20], as well as to the expression for entropy production given by Gaspard [51,52], as mentioned above. By now the literature is quite extensive and the reader is directed there for more details. Here we discuss, briefly, the Gallavotti–Cohen version of a fluctuation theorem appropriate for systems with Gaussian thermostats [18]. In the discussion of systems with energy conserving thermostats, we have introduced another idea for entropy production, namely entropy production in a reservoir associated with phase space contraction of a thermostatted system. Here the central idea is that the thermostatted dynamics produces an attractor for the system's trajectories in phase space. The friction coefficient, which we have denoted by α, is a dynamical function taking on positive or negative values depending on the trajectory of the system in phase space. To produce an overall phase space contraction, the friction coefficient should be positive on average, but from time to time it may be negative. Loosely speaking, we might say that the entropy production in the system is positive when the value of α is positive at the time, and negative when α is negative. For example, in the Lorentz gas that we have been discussing, when the particle moves in the direction of the field, α is positive, and when it moves opposite to the direction of the field, α is negative, provided the charge q is positive. One might imagine some time interval τ, say, and ask for the probability that the time-averaged entropy production per unit time, ϵ_τ, as measured by the phase space contraction, has a value a over this time interval. This probability can be expressed in terms of the SRB measure on the attractor of the steady state system. The fluctuation theorem for this system is a result for the ratio of the probability that the time-average entropy production per unit time over an interval τ will be the value a, to the probability that this value will be $-a$. For a reversible, Anosov-like system with a Gaussian thermostat, the theorem is

$$\frac{\mathrm{Prob}(\epsilon_\tau = a)}{\mathrm{Prob}(\epsilon_\tau = -a)} = \mathrm{e}^{a\tau} \tag{46}$$

Note that for long times this ratio approaches zero or infinity depending upon the sign of a.

This result was first discovered by means of computer simulations by Evans, Cohen, and Morriss [16], and this observation was explained on the basis of Anosov-like dynamics by Gallavotti and Cohen [18]. A closely related fluctuation formula was derived by Evans and Searles [17]. By now this class of fluctuation theorems has been generalized considerably to include analogous results for stochastic and other kinds of systems. We refer to the literature mentioned above for further details.

Discussion

Here we described some aspects of the theory of irreversible processes, as seen from the point of view of dynamical systems theory. We described the notions of ergodic and mixing properties of a dynamical system and argued that they alone are insufficient for a full explanation of the approach to equilibrium as seen on laboratory time scales. However, we tried to argue that if the microscopic system has positive Lyapunov exponents connected to unstable manifolds in the phase space, then one can argue that an approach to equilibrium of reduced distribution functions can occur on much shorter times scales than those needed to establish the mixing of phase space regions throughout the entire phase space. Moreover the existence of a chaotic microscopic dynamics allows us to derive some very striking connections between macroscopic transport coefficients and quantities such as Lyapunov exponents, fractal dimensions, Kolmogorov–Sinai entropies, that characterize the underlying chaotic, microscopic dynamics of the system. Much, but not all, of our discussion was based on some very simple examples of classical chaotic systems with few degrees of freedom, the baker's map and the Arnold Cat Map. These are simple Anosov-like and Anosov systems where one can analyze many points in some detail. However, this analysis is still very far from an analysis of systems of real interest to statistical mechanics, where the number of degrees of freedom is generally much, much larger, even for systems of particles studied on computers, and where the dynamics is not always chaotic. Therefore, it is very much an open question to show that this picture persists when the systems studied are of macroscopic size, and the dynamics is treated in a more realistic way.

Finally we should say something about the role that quantum mechanics plays in our understanding of the approach of a system to equilibrium. "Quantum chaos" is a subject that has developed an enormous literature [15,95,96,97], primarily associated with the quantum behavior of systems with few degrees of freedom which are chaotic in the classical limit of $\hbar \to 0$. This work is clearly of great importance for understanding the behavior of *mesoscopic* quantum devices such as quantum dots and related materials. However, there is no analog of dynamical chaos in quantum mechanics and no direct translation of the results described here to quantum systems. This is a result of the fact that the limit as $\hbar$ approaches zero *does not commute* with the limit of t approaching infinity. Therefore one cannot typically study the asymptotically long time behavior of a quantum system whose classical counterpart is chaotic, and then by taking the limit of $\hbar$ approaching zero, obtain the correct, chaotic behavior of the classical system. Instead one may look for some sign in the quantum motion that the classical system is chaotic. Typically a quantum system will exhibit some signs of the chaotic behavior of the classical system in the semi-classical region of small, but not zero $\hbar$, for some period of time known as the *Ehrenfest time*. The Ehrenfest time is essentially the time that it takes an initially small wave packet to expand to some characteristic length in the system. While the packet is small, the semi-classical behavior is essentially that of a classical system. The rate of the expansion of the wave packet is determined by the classical Lyapunov exponents. When the wave packet reaches a certain size, then the motion of the system is governed by interference and diffraction effects that have no classical counterpart. For example, while a classical particle moving among a random array of fixed scatterers exhibits normal diffusion, its quantum counterpart can be localized, or move diffusively, depending upon the spatial dimension of the system and the particle's energy.

The comments above suggest that chaotic dynamics has a large role to play in quantum systems when one looks at the semi-classical regime, namely the regime where $\hbar$, in proper dimensionless units, can be considered very small [95,96]. One important result, known as *Schnirelman's Theorem* [98]states that in the semi-classical limit, most of the wave functions for a quantum system, whose classical counterpart is chaotic and ergodic, become "equidistributed" on the constant energy surface. This means that the probability of finding of finding the system in some region becomes equal to the ratio of the measure of that region to the total measure of the region available to the system. Nevertheless, there are some wavefunctions for classically chaotic systems that exhibit *scars* in the semi-classical limit, where the wave function is concentrated on periodic orbits of the classical motion. The scarred wave functions then form a special class which do not satisfy Schnirelman's Theorem. In this connection we also mention *Berry's conjecture* [99] which states that the high energy wavefunctions for a classically chaotic, ergodic system can be represented as a Gaussian random function, such as a superposition of plane waves with random phases. For further details we refer to the literature listed above.

The understanding of the approach to equilibrium in macroscopic systems typically requires a treatment of quantum systems with a very large number of degrees of freedom. Such systems are not likely to be in one or in a superposition of a few energy eigenstates. Instead, such systems are likely to be in a state that is a superposition of a huge number of such quantum states, and destructive interference of the phase relations among the states is an important ingredient of the behavior of such systems [100]. In fact there is a large literature that deals with the phenomenon of the loss of coherence, or *decoherence*, of superpositions of large numbers of quantum states for a macroscopic system, and helps us understand a bit more clearly why classical mechanics is a good approximation for describing systems that we know to be intrinsically quantum mechanical.

Future Directions

We conclude with a list of open questions that will provide some indication of what directions might be fruitful further work in the future. This list is far from exhaustive, but focuses upon the particular issues addressed here.

1. Most of the results described here that connect microscopic dynamics and macroscopic transport, such as the escape rate formulae, Eq. (23), the dimension formula Eq. (43), and the Lyapunov exponents-transport coefficient relations for Gaussian thermostatted systems Eq. (34), have been proven for purely chaotic, Anosov or Anosov-like systems, only. The phase spaces of most realistic systems are not likely to be purely chaotic. Instead one expects realistic systems to have mixed phase spaces with both chaotic regions and non-chaotic regions. So far a satisfactory treatment of transport in mixed systems is in a very rudimentary shape, except for some models which have been studied extensively. It would be useful to have estimates of the size and importance of non-chaotic regions in many-particle systems and to determine what corrections, if any, are needed to generalize the results mentioned above to mixed systems.
2. Most of the results given here have been illustrated, and in some cases only derived, for low dimensional systems. How much of this discussion is relevant for

systems with many particles where the dimension of phase space is very large? For example, the formula connecting diffusion and Lyapunov exponents, given by Eq. (43), has only been derived for diffusion in two dimensional chaotic systems Lorentz gases. What is the generalization of this formula to higher dimensions and to general transport processes?

3. The fact that our detailed understanding of the role of chaos in non-equilibrium processes is for systems of low dimensionality suggests that applications of chaos to nanoscale systems might be very fruitful [52].
4. Pseudo-chaotic systems present a great challenge both to physicists and to mathematicians [101,102]. Are there general statements one can make about the motion of a particle in a collection of scatterers, such as hard squares or other scatterers where the motion is not chaotic at all, and the rate of separation of trajectories is at best algebraic? How do these properties depend on the geometrical structure and arrangements of the scatterers? Are there generalizations of the escape-rate formula, and the others mentioned here, to pseudo-chaotic systems?
5. There are at least a few logical gaps in the applications of dynamical systems theory to non-equilibrium statistical mechanics. One one hand we have argued that as far as the approach to equilibrium is concerned it makes sense to look at projected distributions since these distribution functions reach equilibrium forms long before the full phase space distribution function does. This statement itself needs a careful proof for realistic systems. Moreover, the derivations of the formulae that connect transport properties with microscopic dynamical quantities rely upon the use of the full phase space. This suggests that there is a fundamental issue of relevant time scales that needs to be resolved, in order to determine the time scales on which these results become valid.

Acknowledgments

I would like to thank Henk van Beijeren for reading a draft of this article and for his very helpful remarks. I would also like to thank Rainer Klages for his new book (Reference [12]), which was very helpful when preparing this article.

Bibliography

Primary Literature

1. Uhlenbeck GE, Ford GW (1963) Lectures in Statistical Mechanics, 2nd edn. Cambridge University Press, Cambridge
2. Toda M, Kubo R, Saito N (1992) Statistical Physics, vol I. Springer, Berlin
3. Kubo R, Toda M, Hashitsume (1992) Statistical Physics, vol II. Springer, Berlin
4. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617
5. Evans DJ, Morriss GM (1990) Statistical Mechanics of Nonequilibrium Liquids, 2nd edn. Cambridge Univ Press, Cambridge
6. Gaspard P (1998) Chaos, Scattering, and Statistical Mechanics. Cambridge University Press, Cambridge
7. Hoover WG (1999) Time Reversibility, Computer Simulation, and Chaos. World Scientific Publishing Co, Singapore
8. Dorfman JR (1999) An Introduction to Chaos in Nonequilibrium Statistical Mechanics. Cambridge University Press, Cambridge
9. Ruelle D (1999) Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. J Stat Phys 95:393
10. Gallavotti G (1999) Statistical Mechanics – A Short Treatise. Springer, Berlin
11. Szasz D (ed) (2000) Hard-ball Systems and the Lorentz Gas. Encyclopedia of mathematical sciences, vol 101. Springer, Berlin
12. Klages R (2007) Microscopic Chaos, Fractals and Transport in Nonequilibrium Statistical Mechanics. World Scientific Publishing Co, Singapore
13. Klages R, van Beijeren H, Dorfman JR, Gaspard P (eds) (2004) Microscopic chaos and transport in many-particle systems. Special Issue of Physica D, vol 187:1–391
14. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge
15. Srednicki M (1999) The approach to thermal equilibrium in quantized chaotic systems. J Phys A 32:1163
16. Evans DJ, Cohen EGD, Morriss GP (1993) Probability of second law violations in shearing steady flows. Phys Rev Lett 71:2401
17. Evans DJ, Searles DJ (2002) The fluctuation theorem. Adv Physics 51:1529
18. Gallavotti G, Cohen EGD (1995) Dynamical ensembles in stationary states. J Stat Phys 80:931
19. Lebowitz JL, Spohn H (1999) A Gallavotti–Cohen type symmetry in the large deviation functional for stochastic dynamics. J Stat Phys 95:333
20. Crooks GE (1999) Entropy fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys Rev E 60:2721
21. Kurchan J (1998) Fluctuation theorem for stochastic dynamics. J Phys A 31:3719
22. Mazo RM (2002) Brownian Motion: Fluctuations, Dynamics, and Applications. Oxford University Press, Clarendon
23. Helfand E (1960) Transport coefficients from dissipation in a canonical ensemble. Phys Rev 119:1
24. Ehrenfest P, Ehrenfest T (1959) The Conceptual Foundations of the Statistical Approach in Mechanics. Cornell University Press, Ithaca
25. Dettmann CP, Cohen EGD (2000) Microscopic chaos and diffusion. J Stat Phys 101:775
26. Bunimovich L, Sinai YG (1981) Statistical properties of the Lorentz gas with periodic configuration of scatterers. Commun Math Phys 78:478
27. Gaspard, Baras F (1995) Chaotic scattering and diffusion in the Lorentz gas. Phys Rev E 51:5332

28. Turaev D, Rom-Kedar V (1998) Elliptic islands appearing in near-ergodic flows. Nonlinearity 11:575
29. Donnay VJ (1996) Elliptic islands in generalized Sinai billiards. Ergod Theory Dyn Syst 16:975
30. Walters P (1982) An Introduction to Ergodic Theory. Springer, Berlin
31. Hopf E (1937) Ergodentheorie. Springer, Berlin
32. Sinai YG (ed) (1991) Dynamical Systems, A Collection of Papers. World Scientific Publishing Co, Singapore
33. Simányi N (2004) Proof of the ergodic hypothesis for typical hard ball systems. Ann Henri Poincaré 5:203
34. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic Theory. Springer, Berlin
35. Chapman S, Cowling TG (1970) The Mathematical Theory of Non-uniform Gases, 3rd edn. Cambridge University Press, Cambridge
36. Bogoliubov NN (1962) Problems of a dynamical theory in statistical physics. In: Studies in Statistical Mechanics, vol 1. North Holland Publishing Co, Amsterdam
37. Zaslavsky GM (2007) The Physics of Chaos in Hamiltonian Systems. Imperial College Press, London
38. Berry MV (1978) Regular and irregular motion. In: Jorna S (ed) Topics in Nonlinear Dynamics: A Tribute to Sir Edward Bullard, American Institute of Physics, New York
39. Tél T, Gruiz M (2006) Chaotic Dynamics: An Introduction Based on Classical Mechanics. Cambridge University Press, Cambridge
40. Gaspard P (1997) Entropy production in open vol preserving systems. J Stat Phys 88:1215
41. Tasaki S, Gilbert T, Dorfman JR (1998) An analytical construction of the SRB measures for baker-type maps. Chaos 8: 424
42. Ott E (2002) Chaos in Dynamical Systems. Cambridge University Press, Cambridge
43. Gaspard P, Nicolis G (1990) Transport properties, Lyapunov exponents and entropy per unit time. Phys Rev Lett 65:1693
44. Gaspard P (1992) Diffusion, effusion and chaotic scattering. J Stat Phys 68:673
45. Dorfman JR, Gaspard P (1995) Chaotic scattering theory of transport and reaction-rate coefficients. Phys Rev E 51:28
46. Gaspard P, Dorfman JR (1995) Chaotic scattering theory, thermodynamic formalism, and transport coefficients. Phys Rev E 52:3525
47. Viscardy S, Gaspard P (2003) Viscosity in the escape-rate formalism. Phys Rev E 68:041205
48. Evans DJ, Cohen EGD, Morriss GP (1990) Viscosity of a simple liquid from its maximal Lyapunov exponents. Phys Rev A 42:5990
49. Hoover WG, Posch HA (1994) Second-law irreversibility and phase space dimensionality loss from time-reversible nonequilibrium steady-state Lyapunov spectra. Phys Rev E 49:1913
50. Dorfman JR, Gaspard P, Gilbert T (2002) Entropy production of diffusion in spatially periodic deterministic systems. Phys Rev E 66:026110
51. Gaspard P (2004) Time reversed dynamical entropy and irreversibility in Markovian random processes. J Stat Phys 117:599
52. Gaspard P (2006) Hamiltonian dynamics, nanosystems, and nonequilibrium statistical mechanics. Phys A 369:201
53. Tél T, Vollmer J (2000) Entropy balance, multibaker maps, and the dynamics of the Lorentz gas. In: Szasz D (ed) Hard Ball Systems and the Lorentz Gas. Springer, Berlin
54. Vollmer J (2002) Chaos, spatial extension, transport, and nonequilibrium thermodynamics. Phys Rep 372:131
55. van Zon R, Cohen EGD (2004) Extended heat fluctuation theorems for a system with deterministic and stochastic forces. Phys Rev E 69:056121
56. Gaspard P, Rice SA (1989) Scattering from a classically chaotic repeller. J Chem Phys 90:2225
57. Gaspard P (1993) What is the role of chaotic scatttering in irreversible processes? Chaos 3:427
58. Dorfman JR, van Beijeren H (1997) Physica A 240:12
59. Tél T, Vollmer J, Breymann W (1996) Transient chaos: The origin of chaos in driven systems. Europhys Lett 35:659
60. Claus I, Gaspard P (2000) Microscopic chaos and reaction-diffusion processes in the periodic Lorentz gas. J Stat Phys 101:161
61. Claus I, Gaspard P, van Beijeren H (2004) Fractals and dynamical chaos in a random 2D Lorentz gas with sinks. Physica D 187:146
62. Bunimovich LA, Demers MF (2005) Deterministic models of the simplest chemical reactions. J Stat Phys 120:239
63. van Beijeren H, Dorfman JR (1995) Lyapunov exponents and Kolmogorov–Sinai entropy for the Lorentz gas at low densities. Phys Rev Lett 74:4412, erratum 77:1974
64. van Beijeren, Latz A, Dorfman JR (2001) Chaotic Properties of dilute, two and three dimensional random Lorentz gases II: open systems. Phys Rev E 63:016312
65. van Zon R, van Beijeren H, Dorfman JR (2000) Kinetiic theory estimates for the Kolmogorov–Sinai entropy and the largest Lyapunov exponents for dilute, hard ball gases and for dilute, random Lorentz gases. In: Szasz D (ed) Hard Ball Systems and the Lorentz Gas. Springer, Berlin
66. Evans DJ, Hoover WG, Failor BH, Moran B, Ladd AJC (1983) Nonequilibrium molecular dynamics via Gauss' principle of least constraint. Phys Rev A 28:1016
67. Posch HA, Hoover WG (1988) Lyapunov instability of dense Lennard–Jones fluids. Phys Rev A 38:473
68. Posch HA, Hoover WG (1989) Equilibrium and non-equilibrium Lyapunov spectra for dense fluids and solids. Phys Rev A 39:2175
69. Chernov NI, Eyink GL, Lebowitz JL, Sinai YG (1993) Steady state electrical conduction in the periodic Lorentz gas. Commun Math Phys 154:569
70. Baranyi A, Evans DJ, Cohen EGD (1993) Field-dependent conductivity and diffusion in a two-dimensional Lorentz gas. J Stat Phys 70:1085
71. Evans DJ, Cohen EGD, Searles DJ, Bonetto F (2000) Note on the Kaplan–Yorke dimension and linear transport coefficients. J Stat Phys 101:17
72. Dellago C, Glatz L, Posch H (1995) Lyapunov spectrum of the driven Lorentz gas. Phys Rev E 52:4817
73. Dellago C Posch HA, Hoover WG (1996) Lyapunov instability in a system of hard disks in equilibrium and nonequilibrium steady states. Phys Rev E 53:1485
74. Dettmann CP (2000) The Lorentz Gas: A paradigm for nonequilibrium steady states. In: Szasz D (ed) Hardball Systems and the Lorentz Gas. Springer, Berlin
75. Posch HA, Hirshl R (2000) Simulation of billiards and hard body fluids. In: Szasz D (ed) Hard Ball Systems and the Lorentz Gas. Springer, Berlin

76. Dettmann CP, Morriss GP (1996) Proof of Lyapunov exponent pairing for systems at constant kinetic energy. Phys Rev E 53:R5545
77. Wojtkowski M, Liverani C (1998) Conformally Symplectic Dynamics and the symmetry of the Lyapunov spectrum. Commun Math Phys 194:7
78. Bohm A, Gadella M (1990) Dirac Kets, Gamow Vectors and Gelfand Triplets: The Rigged Hilbert Space Formulation of Quantum Mechanics. Springer, Berlin
79. Pollicott M (1985) On the rate of mixing of Axiom-A flows. Inventiones Mathematicae 81:413
80. Pollicott M (1986) Meromorphic extensions of generalized zeta functions. Inventiones Mathematicae 85:147
81. Ruelle D (1986) Resonances of chaotic dynamical systems. Phys Rev Lett 56:405
82. Ruelle D (1986) Locating Resonances for Axiom-A dynamical systems. J Stat Phys 44:281
83. Dörfle M (1985) Spectrum and eigenfunctions of the Frobenius–Perron operator for the tent map. J Stat Phys 40:93
84. Gaspard P (1992) r-adic one dimensional maps and the Euler summation formula. J Phys A 25:L483
85. Gaspard P (1992) Diffusion in uniformly hyperbolic one dimensional maps and Appell polynomials. Phys Lett A 168:13
86. Fox RF (1997) Construction of the Jordan basis for the baker map. Chaos 7:254
87. Gaspard P (1996) Hydrodynamic modes as singular eigenstates of Liouvillian dynamics: Deterministic diffusion. Phys Rev E 53:4379
88. Gilbert T, Dorfman JR, Gaspard P (2001) Fractal dimension of the hydrodynamic modes of diffusion. Nonlinearity 14:339
89. Gaspard P, Claus I, Gilbert T, Dorfman JR (2001) Fractality of hydrodynamic modes of diffusion. Phys Rev Lett 86:1506
90. Fox RF (1998) Entropy evolution for the baker map. Chaos 8:462
91. Goldstein S, Lebowitz JL, Sinai YG (1998) Remark on the (Non)convergence of ensemble densities in dynamical systems. Chaos 8:393
92. van Zon R, van Beijeren H, Dellago C (1998) Largest Lyapunov exponent for many-particle systems at low densities. Phys Rev Lett 80:2035
93. de Wijn A, van Beijeren H (2004) Goldstone modes in Lyapunov spectra of hard sphere systems. Phys Rev E 70:016207
94. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2960
95. Haake F (2001) Quantum Signatures of Chaos. Springer, Berlin
96. Stöckmann H-J (1999) Quantum Chaos: An Introduction. Cambridge University Press, Cambridge
97. Wojcik D (2006) Quantum maps with spatial extent: a paradigm for lattice quantum walks. Int J Mod Phys B 20:1969
98. Lazutkin VF (1993) KAM Theory and Semiclassical Approximations to Wave Functions. Springer, Berlin
99. Berry MV (1977) Regular and irregular wave functions. J Phys 10:2083
100. van Kampen N (1988) Ten theorems about quantum mechanical measurements. Physica A 153:97
101. Gutkin E (1996) Billiards in polygons: A survery of recent results. J Stat Phys 83:7
102. Tabachnikov S (2005) Billiards and Geometry. American Mathematical Society Press, Providence

Books and Reviews

Beck C, Schlögl F (1993) Thermodynamics of Chaotic Systems. Cambridge University Press, Cambridge

Tél T, Gaspard P, Nicolis G (eds) (1998) Focus Issue on Chaos and Irreversibility. Chaos 8(2):309–529

Rom-Kedar V, Zaslavsky G (eds) (2000) Focus Issue on Chaotic Kinetics and Transport. Chaos 10(1):1–288

Casati G, Chirikov B (eds) (1995) Quantum Chaos: Between Order and Disorder. Cambridge University Press, Cambridge

Dorfman JR (1998) Deterministic chaos and the foundation of the kinetic theory of gases. Phys Rep 301:151

Garbaczewski P, Olkiewicz R (eds) (2002) Dynamics of Dissipation. Lecture Notes in Physics, vol 597. Springer, Berlin

Charge-Based Solid-State Flying Qubits

ANDREA BERTONI
CNR-INFM National Research Center
on NanoStructures and BioSystems at Surfaces (S3),
Modena, Italy

Article Outline

Glossary

Bit Elementary unit of classical information represented by a binary digit.

Qubit (or quantum bit) Elementary unit of quantum information. The qubit refers also to the physical system whose state encodes the qubit of information.

Quantum gate Logical operation performed on one or a few qubits, that change their state according to a unitary transformation. Quantum gates are reversible by definition.

Quantum entanglement Property possessed by two or more quantum systems, when the state of the global system that includes all of them cannot be described by the simple composition of their states. Two entangled systems show quantum correlations between their states that have no classical analogue.

Bell's inequality Relation between two sets of measurements performed on two quantum systems spatially separated. Bell's inequality can only be violated if the two systems are entangled.

Quantum wire Metallic or semiconductor wire with nanometric thickness. While the longitudinal current carrying states form a continuum in the energy spectrum, the transverse component of the carriers wave functions originates a discrete spectrum.

Two-dimensional electron gas (2DEG) Gas of electrons that are quantum-confined in one dimension and free to move in the remaining two. In the confinement direction the single-particle states have a quantized energy spectrum. 2DEGs are usually obtained through the modulation of the material conduction band in a semiconductor heterostructure.

Surface acoustic wave (SAW) Elastic acoustic wave that propagates on the surface of a material. In piezoelectric materials SAWs couples with electrons through the SAW-induced piezoelectric field.

Definition of the Subject

The physical implementation of the *qubit*, the dicotomic unit of quantum information, has been, and still is, the starting point of any proposal for a quantum information processing device. Its proper definition is a necessary, although not sufficient, condition for a practical system to be able to exploit the quite large number of quantum algorithms that have been proven to outperform the corresponding classical ones. The first period of quantum information processing research saw many proposals for qubits based on a large spectrum of systems and approaches, but later, the focus was put on solid-state devices, mainly due to their promises for scalability and better integrability with present-day semiconductor electronics. Furthermore, a deep understanding of the quantum-based physics of semiconductor devices is available, due to the long-standing research on the field, mainly driven by the huge commercial success of semiconductor microprocessors. In fact, the application of quantum mechanics to solid-state systems has resulted in a large number of semiconductor devices with novel functionalities, and the quantum mechanical aspects of the device properties start to dominate their behavior, as device dimensions get smaller. While this may become a problem for traditional electronic devices, it is a prerequisite for the implementation of quantum technology in general and quantum information processing in particular. These possibilities have inspired a number of proposals for creating qubits, quantum gates and quantum registers from semiconductor devices and using them for implementing quantum algorithms. Theoretical studies of the device properties have been undertaken, but the challenging technological hurdles have slowed down experimental progress.

Introduction

In what follows, we will describe a class of solid-state qubits, known as *charge flying qubits*, whose characteristic is the evolution with time of their spatial localization. We will first briefly introduce the concept of quantum information processing and, more specifically, of quantum computing. Then we will focus on the definition of the qubit and of quantum entanglement, stemming from the non-separability of the multi-particle state. The logic transformations that must be applied to the qubit in order to implement quantum algorithms, namely the *quantum gates*, will be described and the requirements for a practical realization of quantum computing devices, listed by DiVincenzo in his famous checklist [41], will be reviewed. Then we will enter into the details of a specific proposal for the realization of quantum bits and quantum gates, based on the charge state of electrons in semiconductor quantum wires.

In the last sections, we will return to the concept of qubit entanglement and will present specific calculations for the entanglement creation in a quantum-wire system and, more general, in a carrier-carrier scattering. In fact quantum entanglement can be considered both a resource and a detriment for quantum computing. It is a resource since any quantum algorithm is based on the entanglement of two or more qubit states, it is a detriment when a qubit is entangled with a degree of freedom that is outside the computational space, i. e. the environment, giving rise to the phenomenon of decoherence. It is the aim of the theoretical modeling to identify the proper system parameters for the controlled production of the first kind of entanglement, while avoiding, or at leas reducing, the second one.

The idea of a solid-state qubit, whose state is encoded by the localization of one electron traveling in two coupled semiconductor quantum wires, is here anticipated. The basic qubit structure consists of two parallel quantum wires, close to each other and separated by a potential barrier. The state of the qubit is represented by the location of a single electron traveling in one (state $|0\rangle$) or the other (state $|1\rangle$) of the two wires with a controlled velocity. If a coupling window is introduced between the wires through the lowering of the potential barrier that separates them, the electronic wave function crossing the coupling region oscillates between the wires. When the electron reaches the end of the coupling window, the oscillation process terminates, and the wave function is eventually divided into two parts running along the left and right wires. The necessary superposition of quantum states is thus obtained. The introduction of a low potential barrier in one of the two wires, able to delay the propagation of the wave function with negligible reflection, delays the propagation of the part of the wave function in that particular wire and generates a phase shift between the two components of the qubit. Combinations of these two transformations allow for the implementation of a general single-qubit operation.

A two-qubit gate requires two pairs of wires. They are designed in such a way that only the electron running in the 1 component of the first qubit and the electron running in the 0 component of the second qubit feel their Coulomb interaction, thus generating the necessary entanglement. As a consequence, the electrons are slowed down, and the system undergoes a phase shift that originates from the spatial delay of the two interacting electrons. This last transformation and the two single-qubit transformations described above form a universal set of gates for quantum computing. Numerical simulation of the quantum dynamics of the proposed devices have been performed, in order to obtain the parameters for the experimental realization and to estimate the performance and decoherence times of these systems.

Prototypes of single-qubit devices have been realized experimentally, based on coupled electron wave guides obtained from two-dimensional electron gases (2DEGs) through split-gate or AFM nanolithography techniques. Their basic functioning is currently under investigation. The experiments and the different realization strategies will be mentioned.

Although the electrical properties of quantum wires are nowadays well established both theoretically and from the point of view of experimental realization, the feasibility of complex devices, with a network of coupled quantum wires, remains a challenge. Some considerations about the practical realizability of quantum computers based on solid-state flying qubits will be presented, and an example of a simple quantum-gate network will be given. A number of mechanisms of carrier injection into the QWRs, such as surface acoustic waves (SAW), electron pumps, resonant tunneling devices have been proposed in literature. We will describe, in particular, surface acoustic waves, that have demonstrated the ability to inject and drive single electrons in quantum wires. Numerical simulations of a simple model of such systems will be presented. In fact, surface acoustic waves are also expected to reduce decoherence effects due to the quantized energy spectrum induced on the embedded carriers. Concerning the detection of the final state of the charge qubit, it can be achieved by single-electron transistors placed at the end of each quantum wire. Different kinds of solid-state flying qubits have been proposed, as for example, the qubit encoding realized through the spin state of single electrons propagating in quantum wires [5]. We will not enter into the details of these proposals.

Quantum Information Processing and Quantum Computation

Over the few decades following the discovery of quantum mechanics, the tight link between information theory and physics has been recognized and, in 1961, Rolf Landauer pointed out that any irreversible process leading to erasure of information is necessarily a dissipative process. Landauer's statement [28] about the entropy cost of information, opened the search for understanding of how physics constrains the ability to use and manipulate information. Once recognized that "information is physical" [33], the next step is to consider that the universe is fundamentally quantum mechanical: the classical theories about information need to be revised according to quantum physics.

By taking into account the quantum nature of the information-storing devices it was clear that also the corresponding theory about information processing needed to include the distinctive feature of quantum physics: quantum entanglement. The first widespread proposal of a quantum information processing system was advanced by Richard Feynman, as a means of simulating a quantum system. In fact, an exact classical simulation of a quantum system needs to follow, separately, its possible quantum states, leading to a growth of the resources needed. On the other hand, a quantum simulation would follow a superposition state of the quantum system, with a corresponding superposition of the simulation tool.

However, quantum information processing burst onto the scene of active research in 1994, when Peter Shor proposed a quantum algorithm for the factorization of integer numbers that requires a computational time which is polynomial in $\log(N)$, with N the number to be factorized. Since no classical algorithm is known that can perform such a task more efficiently, the importance of the Shor algorithm was immediately realized. In fact, since many widely used public-key cryptographic methods are based on the practical impossibility of factorizing large integers, an "efficient" factorization algorithm could jeopardize their secrecy.

A number of branches of quantum information theory are nowadays fertile research topics, whose borders are, however, not sharp. We mention the fields of quantum algorithms, quantum cryptography, quantum communication, quantum teleportation and, finally, quantum computation. The starting point of all of the above tasks, is the *qubit*, i. e. the elementary particle of quantum information. A brief description of its basic properties, the formalism adopted to describe it, and the differences with classical bits, are given in the following. Then the specific case of a qubit based on coherent electron transport in coupled wires will be analyzed.

The Qubit

Classical computer systems store and process information in binary encoding. Any integer N is represented by an ordered sequence of n bits $a_n, a_{n-1}, \dots, a_1, a_0$ according to

$$N = \sum_{i=0}^{n-1} a_i 2^i \tag{1}$$

with $0 \leq N \leq (2^n - 1)$ and $a_i \in \{0, 1\}$. In conventional computers, single bits are usually stored as the charge or magnetic state of an electronic device, and transistors are used to realize logic gates, i. e. the transformation of a multiple bit state according to an established rule. Any bit can be in one and only one of two possible states: either 0 or 1. In the quantum case, a bit of information is also represented by one of two possible states, namely $|0\rangle$ and $|1\rangle$, of a given system, but now, since the latter is described by the laws of quantum mechanics, the bit can be in any superposition of the two states. In fact, the states are associated with two specific eigenstates ($|\psi_0\rangle$ and $|\psi_1\rangle$ respectively) of a suitable observable and now the system can be in any state

$$|\psi\rangle = c_0|\psi_0\rangle + c_1|\psi_1\rangle \tag{2}$$

with c_0 and c_1 complex numbers and $|c_0|^2 + |c_1|^2 = 1$. This means that the quantum analog of a bit, the qubit [14], can be simultaneously in both states 0 and 1. If we measure the qubit we will find it with probability $|c_0|^2$ with the value 0 and with probability $|c_1|^2$ with the value 1. It is important to underline that Eq. (2) describes a coherent superposition rather than a generic mixture between 0 and 1. The essential point here is that there is always a base in which the value of the qubit system is well defined, while an incoherent mixture is a mixture whatever base we choose to describe it. As anticipated, the two states of a qubit are usually indicated with $|0\rangle$ and $|1\rangle$, following the Dirac ket notation used in quantum mechanics. Also, it must be noted that the space spanned by a single qubit is isomorphic to the spin space of a spin 1/2 particle. This will allow us to use the formalism for spin states and spin rotations when dealing with single qubits.

Let us suppose now to process some information encoded in an array of n bits. Classically, an input state must be chosen from the 2^n possibilities given by the array, then the information is elaborated, finally an output is produced. To have the output corresponding to a different input, another elaboration must be performed. Things are different if the information is encoded in an array of qubits and if the information is elaborated by means of some kind of "quantum machine" able to preserve the quantum coherence of states. In this case it is possible to create an input state that is a linear superposition of q classical inputs (2^n possibilities *each*), then the output of the elaboration will be the same linear superposition of the q corresponding classical outputs.

In this case, quantum computation exhibits a kind of "natural" massive parallelism, but this peculiarity of quantum mechanics can not be easily exploited. In fact, in order to know the output of the elaboration, a *measure*, in the quantum mechanical sense, must be performed, this producing a collapse of the quantum state into a single component of the linear combination. In our simple example this means that only one random result, among the results of the q classical computations, can be revealed. To get benefit from quantum processing of information, specific algorithms must be implemented, as mentioned in the previous section, that exploit the main peculiarity of quantum physics i. e. quantum entanglement.

Quantum Entanglement

The deep difference between classical information and quantum information becomes evident when a state of two or more qubits is taken. In this case the whole many-qubit system can be in a superposition of states in which none of the qubits has a definite value but there is a fixed correlation between their values. In other terms, an entangled state is any state that cannot be written as a direct product of single-qubit states.

For example, in the two-qubit entangled state

$$\frac{1}{\sqrt{2}} \left(|0\rangle_1|1\rangle_2 + |1\rangle_1|0\rangle_2\right) , \tag{3}$$

where $|0\rangle_n$ ($|1\rangle_n$) indicate the zero (one) state of the qubit n, the probability to obtain the value 0 as the outcome of a measure on either the first or the second qubit is 50%. However, once a measure has been performed, the outcome of a measure on the remaining qubit is known with 100% confidence. The correlation between two systems in a joint quantum state is "stronger" than any classical correlation and leads to effects that are unknown in classical physics. This offers the possibility to encode information in a completely new way, in which the information is stored by the multi-qubit system as a whole and no information is carried by single qubits. In order to compact our notation, we will follow the usual convention of indicating the state of the qubits, lexicographically ordered, with a single ket vector and dropping their index. For example, $|01\rangle$ will be used to indicate the two-qubit state $|0\rangle_1|1\rangle_2$.

Entanglement is closely related to the issue of non-locality of quantum measurement. In fact it was evident since the early days of quantum physics that if two particles are in an entangled state, quantum theory predicts that a measurement process performed on a particle is able to change instantaneously the state of the other, even if they are spatially separated. It was indeed Schrödinger that introduced the word *entanglement* to indicate the superposition in a multi-particle system. Because of this "action at distance", quantum theory was strongly criticized by many physicists. Among them, Albert Einstein, Boris Podolsky and Nathan Rosen, in a famous paper [16] dated 1935, refused the non-locality implied by quantum mechanics and in the epistemological discussion that followed, addressed the new theory as incomplete. The discussion about incompleteness of quantum mechanics was considered merely philosophical until 1964. In that year John Bell showed [6] that, for entangled systems, measurements of correlated quantities should yield different results in the quantum mechanical case with respect to those expected if one assumes that the properties of the measured system are present prior to, and independent of, the observation (reality hypothesis).

In the last section, a proposal for an experimental setup able to violate Bell inequality, based on coherent electron transport in coupled quantum wires will be sketched and the generation of maximally entangled Bell's states by means of solid-state flying qubits will be analyzed. Some results of numerical simulations of such a system will be presented.

Bell's states, using the qubit formalism, are two-qubit states that do not carry any information on the state itself on the single qubits. In fact, the usual way is to encode a bit onto each system separately which gives the following four possibilities: $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$. By exploiting the features of entanglement it is possible to encode the information using four non-separable states of two qubits, the so-called Bell states:

$$\begin{aligned} |\Psi^+\rangle &= \frac{(|01\rangle + |10\rangle)}{\sqrt{2}}\,, \qquad |\Psi^-\rangle = \frac{(|01\rangle - |10\rangle)}{\sqrt{2}}\,, \\ |\Phi^+\rangle &= \frac{(|00\rangle + |11\rangle)}{\sqrt{2}}\,, \qquad |\Phi^-\rangle = \frac{(|00\rangle - |11\rangle)}{\sqrt{2}}\,. \end{aligned} \tag{4}$$

Note that in each of the four states, any single qubit has 50% of probability to give 0 or 1 when measured. As a consequence a single-qubit reading does not provide even partial information on the state. Only a measure on the joint state of two qubits allows to read the stored information.

We finally note that the search for an experimental violation of Bell's inequality started straight after its discovery but it was only in the 1980s that the violation was obtained experimentally [3] using polarization entangled photons from a spontaneous parametric down conversion. On the other hand, the violation of Bell's inequality in semiconductor-based devices has not been, so far, revealed.

Quantum Gates and the Universal Set

In this section, we introduce some of the simplest quantum logic gates that are at the basis of more complex transformations presented in the following and introduce the concept of a universal set of quantum gates.

For a (classical) digital computer the simplest logic gate is the single-bit NOT gate. It simply changes the state of the input bit from 0 to 1 and vice versa. In the quantum case a NOT gate will perform the transformation

$$|\psi_{\text{in}}\rangle = c_0|0\rangle + c_1|1\rangle \xrightarrow{\mathsf{NOT}} |\psi_{\text{out}}\rangle = c_0|1\rangle + c_1|0\rangle\,, \tag{5}$$

in which the complex coefficients of states $|0\rangle$ and $|1\rangle$ are exchanged.

Another useful single-qubit quantum gate that has no correspondence in classical information processing, is the Hadamard gate H, that transforms the two qubit states $|0\rangle$ and $|1\rangle$ according to:

$$|0\rangle \xrightarrow{\mathsf{H}} \frac{|0\rangle + |1\rangle}{\sqrt{2}} \quad \text{and} \quad |1\rangle \xrightarrow{\mathsf{H}} \frac{|0\rangle - |1\rangle}{\sqrt{2}}\,. \tag{6}$$

Although the idea of a quantum logic-gate operation is strictly related to that of classical logic operation, a crucial difference must be underlined. For many classical gates, like AND, OR, XOR, the number of input and output bits is different. In fact the basic logic gates used to implement digital computers are "many-to-one" operations and are not logically reversible. Once the input state is given, it is straightforward to compute the output state, but it is not possible to identify the input from the output. Thus, the logical transformations performed by these classical gates imply a loss of information. We will not enter here into the topic of reversible and non-reversible computation (see,

Charge-Based Solid-State Flying Qubits, Table 1
Truth table defining the controlled-NOT and Toffoli reversible gates

CNOT	
in	out
00	00
01	01
10	01
11	10

Toffoli			
in	out	in	out
000	000	100	000
001	001	101	101
010	010	110	111
011	010	111	110

for example, [7,33]), but simply report a fundamental result of quantum information theory: a quantum logic gate is a unitary transformation of the input state, therefore any quantum gate must be reversible. This means that it is not possible to start from the well-known classical logic gates and find quantum analogues (as for the NOT gate), but reversible transformations must be used instead, in order to identify a universal set of transformations.

A reversible classical gate of special interest, that will be described in the framework of solid-state flying qubit devices in the following section, is the controlled-NOT (CNOT). It is a two-bit to two-bit gate that has, as the first bit of output, a copy of the first bit of input, and as the second bit of output the XOR between the two input bits. The truth table for the controlled-NOT gate is given in Table 1.

It is easy to identify the quantum version of the CNOT gate. The operation performed on a generic two-qubit state $|\psi_{\text{in}}\rangle$ is:

$$\begin{aligned} & c_a|00\rangle + c_b|01\rangle + c_c|10\rangle + c_d|11\rangle \\ & \quad \xrightarrow{\text{CNOT}} c_a|00\rangle + c_b|01\rangle + c_c|11\rangle + c_d|10\rangle \,. \end{aligned} \tag{7}$$

Let us introduce here a last reversible classical gate, the Toffoli gate (sometimes called control-control-NOT). It is a three-bit to three-bit gate whose operation is: the first two bits are unchanged, the third bit undergoes a NOT operation if, and only if, first and second bits are both 1. The truth table for the Toffoli gate is given in Table 1. Even in this case it is trivial to find the quantum version of the Toffoli gate. We will omit the explicit expression for the sake of brevity.

In strict analogy with the usual quantum mechanics formalism, an n-qubit state can be written both in ket formalism and as a column vector, once a basis is chosen.

Here and in the following the matrix representation of operators is expressed on the lexicographically ordered basis, i.e., $|0\ldots00\rangle, |0\ldots01\rangle, |0\ldots10\rangle, \ldots, |1\ldots11\rangle$. For example for a 2-qubit system, the basis will be $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$.

With this notation, vector representation of the 2-qubit state representing the entangled state $a|00\rangle + b|01\rangle + c|11\rangle$ results to be

$$\begin{pmatrix} a \\ b \\ 0 \\ c \end{pmatrix} . \tag{8}$$

This formalism allows a fast calculation of the effect of a quantum gate on a qubit state, once the transformation matrix is given. The matrix representation for three quantum gates introduced previously, (NOT, controlled-NOT, Hadamard) is:

$$\text{NOT} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} ,$$

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} ,$$

$$\text{H} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} .$$

Let us now introduce a compact notation to represent the action performed by a transformation acting on a subset composed of n qubits of an m-qubit system (with $n < m$). The identity operation on the remaining $(m - n)$ qubits will be understood. Let us consider a generic basis ket of the m-qubit system $|b_1, b_2, \ldots, b_m\rangle$ and a generic basis ket of the n-qubit subsystem $|a_1, a_2, \ldots, a_n\rangle$ (with $a_i, b_i \in \{0, 1\}$). If $\mathbf{A}$ is a linear operator acting on $|a_1, a_2, \ldots, a_n\rangle$, we define the operator $\mathbf{A}^{(x_1, x_2, \ldots, x_n)}$ (with $x_i \leq m$ and $x_i \neq x_j$ iff $i \neq j, \forall i, j \leq n$) acting on the m-qubit state $|b_1, b_2, \ldots, b_m\rangle$ as the operator that transforms the n qubits state $|b_{x_1}, b_{x_2}, \ldots, b_{x_n}\rangle$, according the transformation $\mathbf{A}$ and leave the other qubits unchanged.

For example, the four-dimensional matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ that represent the action of a generic single-qubit transformation

$$\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

on a two-qubit system, and that will be used in the following sections, are:

$$\begin{aligned} \mathbf{M}^{(1)} &= \begin{pmatrix} a & 0 & b & 0 \\ 0 & a & 0 & b \\ c & 0 & d & 0 \\ 0 & c & 0 & d \end{pmatrix} , \\ \mathbf{M}^{(2)} &= \begin{pmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{pmatrix} . \end{aligned} \tag{9}$$

In fact, for a 2-qubit system the transformation $\mathbf{M}^{(1)}$ corresponds to the transformation $\mathbf{M}$ applied on the first qubit, while $\mathbf{M}^{(2)}$ corresponds to the transformation $\mathbf{M}$ applied on the second qubit.

To conclude this section, we introduce the concept of a universal set of gates, i. e. the set of simple transformations that can be combined in order to create any possible complex transformation of bits, (or qubits in the present case). We will show, in the following sections that, by using three types of coupled quantum wire devices, it is possible, at least in principle to implement such a universal set.

The concept is analogous to classical computation, where any given transformation can be obtained, for example, using AND and NOT gates. However, as mentioned above, the AND gate is not a reversible gate. This means that the transformations realized using the set {NOT, AND} are, in general, not reversible and we need a different choice for the quantum case.

In 1995 David DiVincenzo [40] showed that, for quantum computation, a universal set formed solely by one-qubit and two-qubit gates exists. A further simplification of the set of universal quantum gates was introduced by Barenco et al. [4]. They proved that a non-universal, classical two-bit gate can be found that, in conjunction with a generic quantum one-qubit gate, form a universal set for quantum computing. On the basis of this last result, we describe, in the following, the universal set implemented by the charge flying qubit.

Let us start by analyzing the one-qubit gate. The space of a two-component vector representing a single flying qubit is isomorphic to the spin space of a spin 1/2 particle. This means that a general transformation of one qubit is represented by a rotation matrix of the SU(2) group, i. e.

$$\mathbf{U}(\alpha,\beta,\theta) = \begin{pmatrix} \mathrm{e}^{i(\alpha/2+\beta/2)}\cos\left(\frac{\theta}{2}\right) & \mathrm{e}^{i(\alpha/2-\beta/2)}\sin\left(\frac{\theta}{2}\right) \\ -\mathrm{e}^{i(-\alpha/2+\beta/2)}\sin\left(\frac{\theta}{2}\right) & \mathrm{e}^{i(-\alpha/2-\beta/2)}\cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{10}$$

where α, β, θ are three real numbers representing three rotation angles around two orthogonal axes: α and β are rotations around Z, between them, a rotation around Y of γ is performed [39]. Any transformation $\mathbf{U}$ can be obtained using two of the three basic rotations:

$$\mathbf{R_x}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) \\ i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{11}$$

around the X axis,

$$\mathbf{R_y}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & \sin\left(\frac{\theta}{2}\right) \\ -\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{12}$$

around the Y axis, and

$$\mathbf{R_z}(\theta) = \begin{pmatrix} \mathrm{e}^{i\frac{\theta}{2}} & 0 \\ 0 & \mathrm{e}^{-i\frac{\theta}{2}} \end{pmatrix}, \tag{13}$$

around Z axis.

In particular, the generic transformation $\mathbf{U}$ in the form (10) can be obtained using only $\mathbf{R_y}$ and $\mathbf{R_z}$:

$$\mathbf{U}(\alpha,\beta,\theta) = \mathbf{R_z}(\alpha)\ \mathbf{R_y}(\theta)\ \mathbf{R_z}(\beta)\,. \tag{14}$$

Furthermore, one of the above three rotations can be obtained using the other two. In fact

$$\mathbf{R_x}(\theta) = \mathbf{R_z}(\pi/2)\ \mathbf{R_y}(\theta)\ \mathbf{R_z}(-\pi/2) \tag{15}$$

$$\mathbf{R_y}(\theta) = \mathbf{R_z}(-\pi/2)\ \mathbf{R_x}(\theta)\ \mathbf{R_z}(\pi/2) \tag{16}$$

$$\mathbf{R_z}(\theta) = \mathbf{R_x}(-\pi/2)\ \mathbf{R_y}(\theta)\ \mathbf{R_x}(\pi/2)\,. \tag{17}$$

This means that in order to obtain a generic single-qubit transformation (the first gate in our universal set) it is sufficient to have two of the rotations (11), (12), (13).

We now consider the two-qubit gate of the set. A typical choice for the latter is the CNOT introduced above. Another possibility is to use a two-qubit conditional phase shifter. This gate adds a phase factor to a given component of the qubit state. The usual form for the conditional phase shifter present in literature adds a phase $\mathrm{e}^{i\gamma}$ to the state $|11\rangle$:

$$|\psi_{\text{in}}\rangle = c_a|00\rangle + c_b|01\rangle + c_c|10\rangle + c_d|11\rangle \longrightarrow$$
$$\longrightarrow |\psi_{\text{out}}\rangle = c_a|00\rangle + c_b|01\rangle + c_c|10\rangle + \mathrm{e}^{i\gamma}c_d|11\rangle\,.$$

For any value of the angle γ, not equal to an integer multiple of π, we have a transformation that, together with the one-qubit gates $\mathbf{U}$ of Eq. (10), form a universal set.

In the following section, the conditional phase shifter used is slightly different from the one described above. In fact it adds a phase to the $|10\rangle$ component of the state. Its matrix representation is:

$$\mathbf{T}(\gamma) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \mathrm{e}^{i\gamma} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{18}$$

and forms, together with the one-qubit gates $\mathbf{U}$, a universal set.

The DiVincenzo Criteria

The concrete realization of quantum-computing capable hardware must face a number of general tasks and requirements that have been listed by David DiVincenzo [41] in the form of a checklist. Five criteria were listed, that must be met by physical systems that are to form the basis of a useful quantum computer.

For the sake of completeness, we report in the following the checklist, as given in [2]:

1. A scalable physical system of well-characterized qubits.
2. The ability to initialize the state of the qubits to a simple fiducial state.
3. Long relative decoherence times, much longer than the gate-operation time.
4. A universal set of quantum gates.
5. A qubit-specific measurement capability.

The above criteria are largely qualitative, but more quantitative bounds have been proposed [27,34]. A universally accepted criterion is the number of gate operations that can be executed before decoherence processes have degraded the quantum information significantly. This number is often estimated as the ratio of the decoherence time by the single gate operation time. However, additional mechanisms leading to the loss of quantum information must also be considered, such as the accuracy of the logic operation performed by the quantum gate. In fact, a problem that is often present in the design of quantum information devices is that optimization of one of the relevant parameters conflicts with the requirements for another operation. As an example, good readout sensitivity and fast gate operations require strong coupling to the classical apparatus that drives the device. However, strong interactions with the environment increases the speed of decoherence. Optimizing performance of the full system, i. e. from initialization through computation and readout, therefore forces one to make choices and to accept some additional decoherence if this results in improved readout sensitivity.

In addition to the above points, two additional points have been added in order to consider the issues related to the transfer of quantum information between different devices implementing the quantum operations:

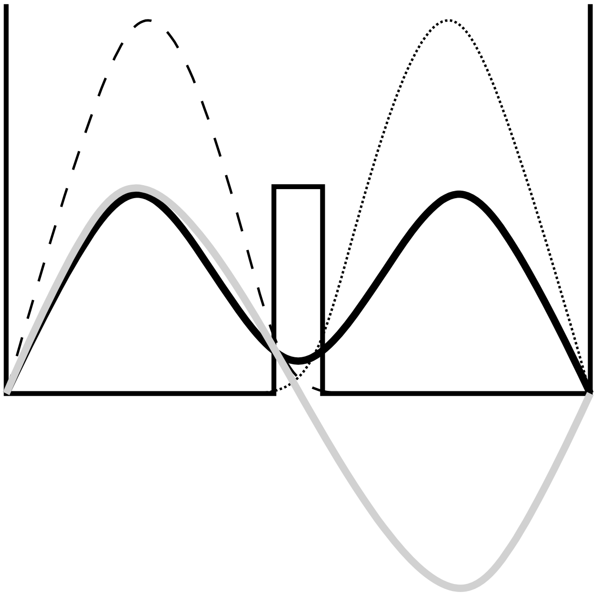
6. The ability to interconvert stationary and flying qubits.
7. The ability to faithfully transmit flying qubits between specified locations.

The latter criteria, related to quantum computer networkability, are automatically met by proposals whose qubit definition is based on flying qubits. In fact, in order to perform a quantum computation the qubits must undergo a sequence of logical operations, represented by unitary transformations on the multi-qubit state. Two classes of proposals can be identified, according to the kind of qubit encoding [8]. The first one is characterized by qubits that are defined by the quantum state of given physical systems, fixed into space, with quantum gates implemented by changing in time the coupling between the qubits. Quantum computation based on charge and spin degrees of freedom of semiconductor quantum dots are an example of these implementations, where a logic operation is performed by a proper time-dependent tuning of interdot coupling. The second class is represented by the flying qubits, where, contrary to the previous case, quantum gates are fixed, while the position of the qubits changes in time. At each stage, the outputs of a quantum transformation, i. e. the state of a qubit register, must be moved to the inputs of the following quantum gate. This process, overlooked in the early times of quantum information processing research, results to be critical in viable practical realizations and strictly connected to the problem of quantum communication [41].

Flying Qubits Based on Coherent Electron Transport in Coupled Quantum Wires

In this section, a physical system able to perform the basic operations needed for quantum computation is proposed and investigated theoretically. As anticipated in the Introduction, the elementary qubit is defined as the state of an electron running along a couple of quantum wires. We will not enter into the details of the modeling of a realistic device, for which we refer to recent literature [1,29,32,38,42]. We will use hard-wall confining potentials and simple models in order to focus on the functioning principles of the quantum gates. However, we will use GaAs-AlGaAs heterostructure parameters in the simulations presented in the following. We will show that a proper design of the system, together with the action of Coulomb interaction between electrons, allows the implementation of basic one-qubit and two-qubit quantum logic gates.

Let us consider the quantum well formed by a modulation-doped heterostructure. If we assume low free carrier density, less than $10^{10}/\mathrm{cm}^2$, in the so-formed two-dimensional electron gas (2DEG), the confining potential profile along the direction orthogonal to the layers can be considered as a narrow triangular well (see, for example, [19]). Along this direction the wave function is quantized and has, for the lower, bound states, a discrete energy spectrum. The motion of the carriers in the plane parallel to the interfaces is free, so that, for any transverse energy, there exist a continuum of two-dimensional states called



Charge-Based Solid-State Flying Qubits, Figure 1
Schematic transverse potential profile of two weakly coupled quantum wires (double well). The *dark* and *light solid lines* show the two lower eigenfunctions: $|\psi_e\rangle$ and $|\psi_o\rangle$, with even and odd parity, respectively. The *dashed* and *dotted lines* represent their sum ($|0\rangle$ state, localized in the left well) and difference ($|1\rangle$ state, localized in the right well), respectively

subbands. Such a 2DEG system is the basis from which quantum wire structures can be realized by lithographic process or electrostatic confinement. In the following, the confined direction, i. e. the material growth direction, is ignored since it is supposed that a single electron, injected in the system, will remain in the lowest subband.

The basic device consists of two parallel quantum wires [10]. It is supposed to operate at very low temperature in order to have a negligible number of electrons in the conduction band and to minimize the decoherence due to the interactions of the electrons with lattice vibrations. The transverse potential profile of the two-wire system is shown in Fig. 1 together with its ground eigenfunction $|\psi_e\rangle$, that has even parity, and the first excited state $|\psi_o\rangle$, that has odd parity.

Figure 1 shows also that the sum and difference of the two lower eigenfunctions are almost fully localized in the left and right well, respectively. The widths of the two wires are equal and constant along the whole device. This avoids any effect due to unalignment of subbands in the two quantum wires [24].

Let us consider a single electron injected into one of the two wires and assume that the injection process is able to keep the electron in the transverse ground state, and that, in the longitudinal direction along the wire, it is well described by a minimum uncertainty wave packet.

$$\psi(x,y) = \sqrt{\frac{2}{L}}\cos\left[\frac{\pi}{L}(x-x_0)\right] \frac{1}{\sqrt{\sigma\sqrt{2\pi}}}\mathrm{e}^{-\left(\frac{y-y_0}{2\sigma}\right)^2}\mathrm{e}^{ik_0 y}\,, \tag{19}$$

where, L is the width of the quantum wire, whose center is x_0, σ is the standard deviations of the Gaussian function and gives an estimate of the wave function extension in space along the wire direction, y_0 is the coordinate of the center of the wave function, k_0 is the wave number, representing the kinetic energy of the electron along the y direction. This is the initial condition for the following time-dependent numerical simulations.

The state of the qubit is defined as $|0\rangle$ if the electron is in the left wire, $|1\rangle$ if it is in the right wire. This means, in terms of transverse wave functions:

$$\begin{aligned}\langle x|0\rangle &= \frac{1}{\sqrt{2}}(\langle x|\psi_e\rangle + \langle x|\psi_o\rangle) = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x-x_0)\right)\\ \langle x|1\rangle &= \frac{1}{\sqrt{2}}(\langle x|\psi_e\rangle - \langle x|\psi_o\rangle) = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x-x_1)\right)\end{aligned} \tag{20}$$

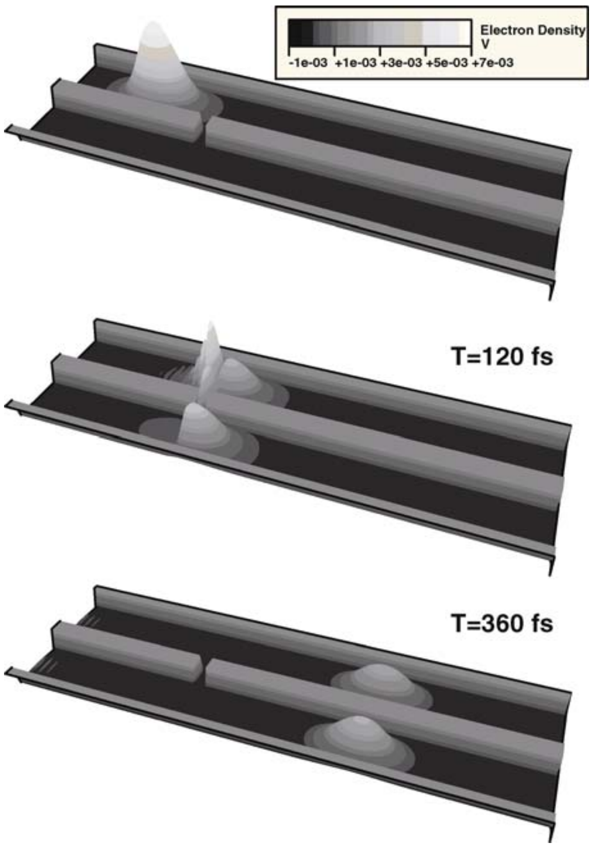
where $x_0(x_1)$ is the central point of the left (right) well formed by the potential in the direction orthogonal to the wires.

The time evolution of the states defined by Eq. (20), as regards the x component, is simply given by a multiplying factor $\mathrm{e}^{-i\omega t}$ with ω frequency of the two degenerate eigenstates.

One-Qubit Gates

If a coupling window is introduced between the wires, a significant lowering of the symmetric transverse state energy ω_e is produced. It can be easily shown that, with good approximation the energy of the anti-symmetric state is unchanged ($\omega_o = \omega$). It is well known (and it has been verified experimentally, see for example [18]) that the electronic wave function crossing this coupling region, oscillates between the wires with a period $\tau = 2\pi/(\omega - \omega_e)$. When the electron reaches the end of the coupling window, the oscillation process terminates and the wave function is eventually separated in two parts running along the left and the right wires, as shown in Fig. 2.

If T is the time that the electron spends in the coupling region, the transverse states in Eq. (20) evolve according



Charge-Based Solid-State Flying Qubits, Figure 2
Numerical solution of the time-dependent Schrödinger equation for one electron injected in a system of two coupled quantum wires [10]. The model potential profile and the square modulus of the wave function are shown at three different time steps. The width of the wires is $L = 6$ nm and the coupling window is 10 nm long. Note that different scales have been used for the two axes. The initial condition (in the upper image) is the wave function in Eq. (19) with $\sigma = 20$ nm and k_0 corresponding to an energy of 50 meV. The effective mass for GaAs has been used

to:

$$\begin{aligned}|0\rangle &\rightarrow \frac{e^{-i\omega T}}{\sqrt{2}}\left(e^{i(\omega-\omega_e)T}|\psi_e\rangle + |\psi_o\rangle\right)\\ &= e^{-i\omega T}\, e^{i\frac{\theta}{2}}\left(\cos\frac{\theta}{2}|0\rangle + i\sin\frac{\theta}{2}|1\rangle\right)\\ |1\rangle &\rightarrow \frac{e^{-i\omega T}}{\sqrt{2}}\left(e^{i(\omega-\omega_e)T}|\psi_e\rangle - |\psi_o\rangle\right)\\ &= e^{-i\omega T}\, e^{i\frac{\theta}{2}}\left(i\sin\frac{\theta}{2}|0\rangle + \cos\frac{\theta}{2}|1\rangle\right),\end{aligned} \tag{21}$$

where $\theta = (\omega - \omega_e)T$. The length of the coupling window, the height of the barrier in it and the velocity of the electron, can be properly chosen to perform a given transfer of the electronic wave function between the wires. It is easy to see that, using the matrix formalism introduced in the previous section, the transformation **S** performed is represented by the rotation matrix $\mathbf{R_x}$ multiplied by a phase factor:

$$\mathbf{S}(\theta) = e^{i\frac{\theta}{2}}\begin{pmatrix}\cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right)\\ i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right)\end{pmatrix} = e^{i\frac{\theta}{2}}\mathbf{R_x}(\theta)\,. \tag{22}$$

The global phase factor $e^{-i\omega T}$, due to the time evolution of the state, is ignored for simplicity: it is present in each of the transformations proposed and also in the free propagation of the electron, therefore it can be considered a global multiplying factor for the many-qubit system that gives no contribution.

Similarly, the exponential $e^{i\frac{\theta}{2}}$ of Eq. (22) results to be a global phase factor and can be neglected. Thus, in the following, the operation performed by the geometry described in this section (coupling window), will be indicated by either $\mathbf{S}(\theta)$ or $\mathbf{R_x}(\theta)$.

A second kind of transformation on a single qubit can be obtained with the introduction of a potential barrier able to delay the propagation of the wave function in one of the two wires (but low enough to avoid reflections [9]). This induces a phase shift between the two components of the qubit.

Let us suppose to have a qubit in a state $a|0\rangle + b|1\rangle$ (with a and b complex numbers and $|a^2 + b^2| = 1$). This means that the initial wave function $\psi_i(x, y)$: (a) in the longitudinal direction along the wires still has the form of a minimum uncertainty wave packet, (b) in the transverse x direction it is split in two parts, one in each wire, i. e., according to Eq. (20):

$$\begin{aligned}\psi_i(x, y) &= \left[a\cos\left(\frac{\pi}{L}(x - x_0)\right) + b\cos\left(\frac{\pi}{L}(x - x_1)\right)\right]\\ &\quad\times\sqrt{\frac{2}{L}}\frac{1}{\sqrt{\sigma\sqrt{2\pi}}}e^{-\left(\frac{y-y_0}{2\sigma}\right)^2}e^{ik_0 y}\\ &= [a\langle x|0\rangle + b\langle x|1\rangle]\frac{1}{\sqrt{\sigma\sqrt{2\pi}}}e^{-\left(\frac{y-y_0}{2\sigma}\right)^2}e^{ik_0 y}\,,\end{aligned}$$

or, as a column vector,

$$\psi_i(x, y) = \frac{1}{\sqrt{\sigma\sqrt{2\pi}}}\begin{pmatrix}a\, e^{-(\frac{y-y_0}{2\sigma})^2}e^{ik_0 y}\\ b\, e^{-(\frac{y-y_0}{2\sigma})^2}e^{ik_0 y}\end{pmatrix}. \tag{23}$$

It is known, from elementary quantum mechanics [30], that the effect of a potential barrier on a propagating wave function (like the one described in Eq. (19)) can be reduced, with good approximation, to a spatial delay ΔS, provided either that (i) the mean kinetic energy of the wave packet is much higher than the barrier, or that (ii) $k_0 L$ is a multiple of π and the broadening in momentum space is small (σ is large).

If the delaying potential barrier is inserted on the 0 wire, only the $|0\rangle$ component of the wave function will undergo the spatial delay ΔS. The state of Eq. (23) will be transformed in

$$\psi_f(x,y) = \frac{1}{\sqrt{\sigma\sqrt{2\pi}}} \begin{pmatrix} a\,\mathrm{e}^{-\left(\frac{y+\Delta S-y_0}{2\sigma}\right)^2} \mathrm{e}^{ik_0(y+\Delta S)} \\ b\,\mathrm{e}^{-\left(\frac{y-y_0}{2\sigma}\right)^2} \mathrm{e}^{ik_0 y} \end{pmatrix} . \tag{24}$$

The delay ΔS is small compared to the parameter σ of the enveloping Gaussian function, thus the difference induced by the substitution $y \rightarrow y + \Delta y$ in the real exponential, is small compared to the variation of the imaginary exponential representing the enveloped plane wave. As a consequence it is possible to make the approximation

$$\mathrm{e}^{-\left(\frac{y+\Delta S-y_0}{2\sigma}\right)^2} \mathrm{e}^{ik_0(y+\Delta S)} \approx \mathrm{e}^{-\left(\frac{y-y_0}{2\sigma}\right)^2} \mathrm{e}^{ik_0(y+\Delta S)} \tag{25}$$

and to rewrite the state of Eq. (24) as

$$\psi_f(x,y) \approx \frac{1}{\sqrt{\sigma\sqrt{2\pi}}} \mathrm{e}^{-\left(\frac{y-y_0}{2\sigma}\right)^2} \mathrm{e}^{ik_0 y} \begin{pmatrix} a\,\mathrm{e}^{ik_0\Delta S} \\ b \end{pmatrix} . \tag{26}$$

We introduce, now, an angle ϕ defined as

$$\phi = k_0 \Delta S\,. \tag{27}$$

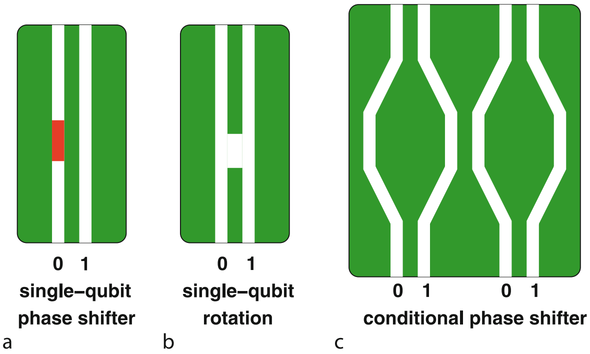
It is easy to verify that the transformation $\psi_i \longrightarrow \psi_f$ on the qubit state is given by

$$\mathbf{R_0}(\phi) = \begin{pmatrix} \mathrm{e}^{i\phi} & 0 \\ 0 & 1 \end{pmatrix} , \quad \text{or} \quad \mathbf{R_1}(\phi) = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{e}^{i\phi} \end{pmatrix} , \tag{28}$$

if the delay is realized on the wire representing the 0 or 1 component, respectively. The value of ϕ depends on the delay ΔS that depends, in turn, on the height and length of the potential barrier.

The transformations $\mathbf{R_0}$ and $\mathbf{R_1}$ can be combined to obtain a transformation $\mathbf{R_z}$, i. e. the rotation around the $\hat{z}$ axis defined in Eq. (13), that is rewritten here for convenience:

$$\mathbf{R_0}(\alpha/2)\mathbf{R_1}(-\alpha/2) = \mathbf{R_z}(\alpha) = \begin{pmatrix} \mathrm{e}^{i\frac{\alpha}{2}} & 0 \\ 0 & \mathrm{e}^{-i\frac{\alpha}{2}} \end{pmatrix} . \tag{29}$$



Charge-Based Solid-State Flying Qubits, Figure 3
Schematic representation of the three devices that constitute the universal set of quantum gates: a a delaying potential barrier (*red*) along the 0 wire realizes the R_0 transformation; b a coupling window between the wires realizes the R_x transformation; c a two-qubit Coulomb coupler is able to entangle the two qubit states and to realize a T transformation (see Eq. (18))

As described in the previous section, $\mathbf{R_z}$ together with $\mathbf{R_x}$, gives any rotation of the group SU(2). Thus, the two basic transformations proposed are able to realize the general rotation $\mathbf{U}$ of Eq. (10) that is the single-qubit gate of the universal set for the proposed qubit system.

A schematic representation of the two devices implementing the $\mathbf{R_0}$ and $\mathbf{R_x}$ transformations are reported in Fig. 3 (a) and (b), respectively.

It is straightforward to note that the transformations $\mathbf{R_x}$ and $\mathbf{R_z}$ for the electronic wave function are analogues to the transformations induced, in a two-path interferometer, by a photon beam splitter and a photon phase shifter. Furthermore, a simple calculation, that we omit for brevity, shows that the Hadamard gate of Eq. (6) can be obtained by using the above transformations.

Two-Qubit Gate

In order to implement the conditional phase shifter gate T, whose operation is described in Eq. (18), two qubits must be considered. Figure 3c represents the geometry of the two-qubit device. In the following we will call "first" the qubit represented by the left couple of wires, "second" the one represented by the right couple. In the two-qubit binary encoding of the state of the system, the second qubit is the less significant bit.

The device is designed as follows [10]. The four wires of the two qubits run parallel to each other and, in a small region, the wire 1 of the first qubit and the wire 0 of the second qubit get close to each other, enough to give rise to a significant Coulomb coupling between two electrons running along them. In fact, when the two electrons approach each other, part of their kinetic energy is transformed into a repulsive coupling potential and the velocity along the wires is reduced. Then, when the distance between them increases again to the original value, the potential energy is transformed back into kinetic energy and the initial velocity is restored. As a result of this process, the two electrons running along the right wire (state $|1\rangle$) of the first qubit and the left wire (state $|0\rangle$) of the second qubit, will suffer a delay in their propagation compared to the case in which either the first or the second qubit are not in one of the two central wires. This delay depends on the length of the coupling region and on the distance between the two central wires, and corresponds to a phase factor in front of the $|10\rangle$ component of the two-qubit system.

The transformation T induced by the described geometry (shown in Fig. 3c) can be also understood by analyzing the effect on the two electrons separately. If the first electron is in state $|1\rangle$ (the electron is in the wire near the second qubit), the slowing down caused on the second qubit is similar to the slowing down caused by a delaying potential barrier and the transformation on the second qubit is of the kind $\mathbf{R_0}$; similarly, if the second qubit is $|0\rangle$ the transformation on the first qubit is $\mathbf{R_1}$. As can be easily seen by direct inspection, if the whole two-qubits system is considered, the behavior described above gives rise to the two-qubits transformation T.

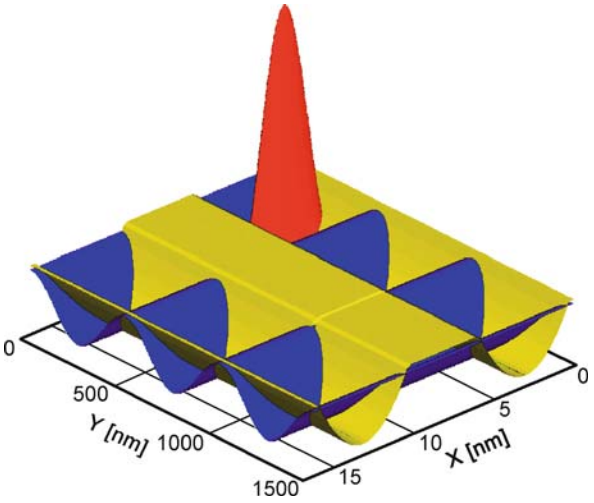
This proposal for the physical realization of a two-qubit gate is alternative to the one advanced in [25], where a CNOT gate was proposed (instead of the T gate proposed here) exploiting the difference in the oscillation period, induced by the Coulomb interaction between two electrons. It must be noted that, together with single qubit rotations, the conditional phase shifter is the basic transformation for the quantum discrete Fourier transform, needed to implement Shor's algorithm [17]. Starting from the CNOT gate, a quantum network is needed to implement a $\mathbf{T}$ transformation:

$$\mathbf{T}(\gamma) = \mathbf{R_1}^{(1)}\left(\frac{\gamma}{2}\right)\mathbf{R_1}^{(2)}\left(-\frac{\gamma}{2}\right)\text{CNOT}\;\;\mathbf{R_1}^{(2)}\left(\frac{\gamma}{2}\right)\text{CNOT}\,. \tag{30}$$

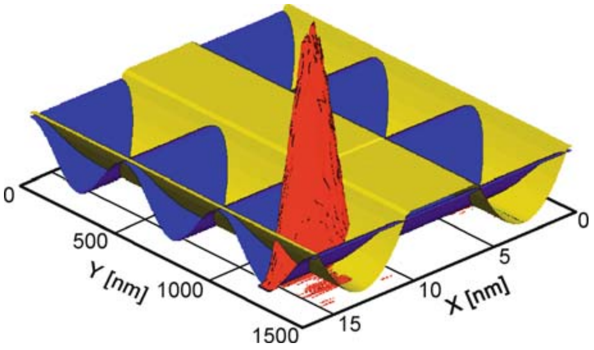
This network is simple, but the present proposal can avoid this step and directly realize the conditional phase shifter.

Electrons Driven by Surface Acoustic Waves

As an alternative to the ballistic propagation of electrons along the quantum wires, the use of surface acoustic waves (SAW) has been proposed. The SAW technology presents the drawback of some local heating near the SAW transducer. However, Barnes et al. [5] have successfully applied the SAW technique for the electron injection and transport along quantum wires. Within this approach a number of electrons are captured from a 2DEG and placed into the minima of a sinusoidal acoustic wave that propagates along the device. The 2DEG region is connected to the quantum-wire region in which a single electron must be injected. When the SAW minimum reaches the 1D channel, the trapped electrons undergo a further confinement due to the lateral potential that constitutes the wire, so that a moving quantum dot is formed. With a suitable choice of the SAW parameters it is possible to create a dot that carries a single electron. The electron moving inside the 1D channel is embedded in the minimum of the dot and, like in the free-propagation case, it experiences the cascade of



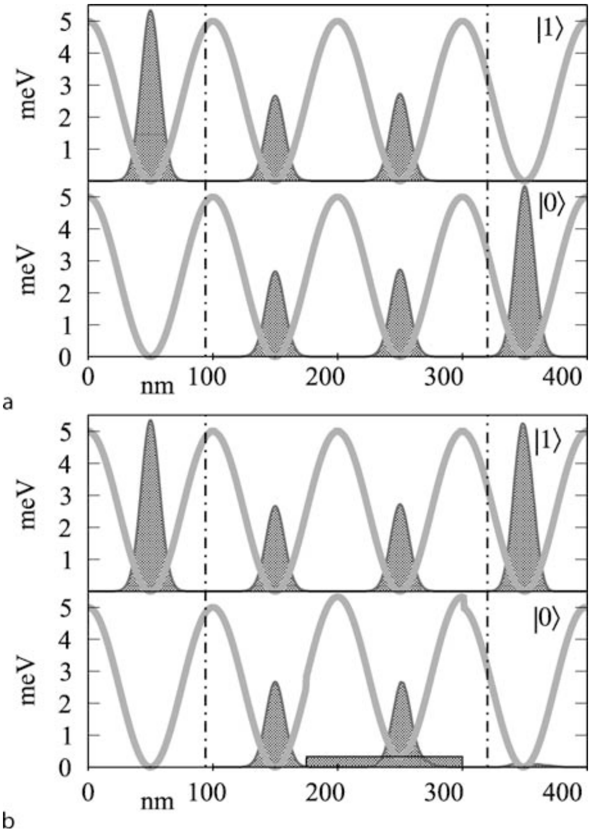
Charge-Based Solid-State Flying Qubits, Figure 4
Electron wave function (*red*) driven along a double wire device by the time-dependent sinusoidal potential of the SAW (*blue*) that are propagating from left to right



Charge-Based Solid-State Flying Qubits, Figure 5
Electron wave function (*red*) of Fig. 4 after the R_x transformation induced by the inter-wire coupling window at $y = 1000$ nm

quantum gates. In this way, the wave-packet could be more immune to decoherence effects, as it is confined along the three dimensions.

Figures 4 and 5 show the results, at initial and final time, respectively, of a time-dependent numerical simulation of a $\mathbf{R_x}$ gate device with SAW. Because of the long wave length of the SAWs (about 500 nm here), the spatial spreading of the electron wave function is large. Furthermore, the coupling window, located at $y = 1000$ nm, is small and hardly visible. However, the system parameters have been tuned in order to obtain a NOT operation and the electron wave function embedded in the SAW minimum is transferred from the wire 1 to the wire 0, thus confirming the proper functioning of the device.



Charge-Based Solid-State Flying Qubits, Figure 6

One-dimensional time-dependent simulation of an electron wave function trapped inside a minimum of a surface acoustic wave (*sinusoidal solid line*), propagating from the left to the right. The square modulus of the wave function is shown at four different time steps (*shaded regions*). The coupled-wire device has two single-qubit rotations $R_x(\pi/2)$ (represented by *dashed-dotted lines*). In the *lower graph* a phase shifter $R_z(\pi)$ (small potential around 200–300 nm) is inserted in the wire 0 between the rotations, leading to a different final state of the qubit

While the action of the coupling window ($\mathbf{R_x}$ transformation) is not altered in the SAW approach, the functioning of the gates $\mathbf{R_z}$ and $\mathbf{T}$ need to be revised [12]. In fact, the confinement along the wire, due to the SAW, prevents the spatial delay of the wave packet, since the velocity of the electron is fixed by the SAW velocity. However, also in this case the $\mathbf{R_z}$ transformation can be realized by means of a potential barrier along the wire. Now the phase shift originates from the change of the energy levels of the moving dot, caused by the additional potential of the barrier. The $\mathbf{R_z}$ and $\mathbf{T}$ transformations in the frame of the SAW approach have been validated by numerical simulations and an example is shown in Fig. 6, where two 1D simulations for a SAW-driven electron are reported [8]. The wave function, shown at four different time steps, is initialized in the ground state of the sinusoidal SAW potential, inside the $|1\rangle$-state wire. Two quantum gate networks are simulated, namely $\mathbf{R_x}(\pi/2)\mathbf{R_x}(\pi/2)$ (with two coupling windows) and $\mathbf{R_x}(\pi/2)\mathbf{R_z}(\pi)\mathbf{R_x}(\pi/2)$ (with a coupling window, a potential barrier and a second coupling window) in the two uppermost and the two lowermost plots, respectively. In the second case the presence of a small potential barrier between the two splitting windows is able to change the final state of the qubit, as in the case of free propagation of the carrier. The use of SAW-induced electron interferometry in QWRs is being actively investigated both theoretically [13,23,37] and experimentally [15].

Entangled States and Estimation of Quantum Entanglement

The simulation presented in the previous section, is the result of a time-dependent Schrödinger solver that evolves the electrons in a single particle approximation. This is necessary due to the high computational effort that is needed for a ballistic evolution, with Coulomb interaction, in two dimensions. In the above approach, each electron wave-function $\psi(x, y)$ is represented, at a given time step, by a two-dimensional array. This means that no entanglement between electrons can be simulated. Nevertheless, the simulations performed validated the devices using a base of factorizable qubit states: the effect on entangled states can be easily deduced once the behavior on the given basis is known, thanks to the superposition principle. For example, the result of a transformation $\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)$ on a state $(|00\rangle + |10\rangle)$ is an entangled state:

$$\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)(|00\rangle + |10\rangle) = -(|01\rangle + i|10\rangle)\,. \quad (31)$$

The condition of separability imposed on the wave function will be dropped in this section. In fact, the possibility to simulate an entangled two-electron state allows the study of the physical system proposed in [26] for the test of Bell inequality using quantum-wire semiconductor devices.

Since the straightforward numerical simulation, in two dimensions, of the time evolution of $\psi(x_1, y_1, x_2, y_2, t)$ results to be too expensive from a computational point of view, a simpler semi-1D model can be used to study the proposed quantum wire system [11]. Let us consider, for the moment, a single qubit, i. e. an electron running along a couple of QWRs, with its wave function $\psi(x_1, y_1)$ eventually split between the wires. If the transverse component (orthogonal to the wires) of the electron wave function remains in the ground state of the transverse potential, i. e. the inter-subband scattering is negligible, then, along the device, the transverse dynamics can be neglected. This is obviously false in the regions of the $\mathbf{R_x}$ coupling windows, which will be considered in the following.

Two possible states of the qubit can be used to label the two possible forms for the transverse wave function, a cosine centered either in the wire 0 or 1:

$$\begin{aligned} x_1 = 0 &\longrightarrow \langle x|0\rangle = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x - X_0)\right) \\ x_1 = 1 &\longrightarrow \langle x|1\rangle = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x - X_1)\right) , \end{aligned} \tag{32}$$

where X_0 and X_1 are the central points of wires 0 and 1, respectively, and L is their width.

Concerning the $\mathbf{R_x}(\theta)$ transformation, within this semi-1D model, it is not possible to simulate directly the wave function splitting induced by the coupling window. In fact the wave function evolution that originates the transformation comes from the transverse dynamics whose detailed description is now lost. It is, however, possible to directly use the analytical expression of $\mathbf{R_x}$ in order to account for the coupling window, once it has been validated through numerical simulations [11]. The above considerations allow one to include the effect of an $\mathbf{R_x}(\theta)$ transformation, directly by the application of the proper transformation matrix. A similar approach has been used to include the two beam-splitters in the simulations of Fig. 6.

Let us consider now the third gate, namely the two-qubit conditional phase shifter sketched in Fig. 3c. As for the single-qubit case, it is possible to associate to each wire of the first or second qubit, a given value (either 0 or 1) of the x coordinate, considered now as a parameter. This is done for both the first and second qubit. With this approach, the model is strongly simplified: from a time-dependent Schrödinger equation for the five-variable wave function $\psi(x_1, y_1, x_2, y_2, t)$, to four equations:

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\psi_{x_1,x_2}(y_1, y_2, t) = &-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial y_2^2}\right) \\ &\cdot \psi_{x_1,x_2}(y_1, y_2, t) \\ &+ V_{x_1,x_2}(y_1, y_2)\psi_{x_1,x_2}(y_1, y_2, t) \end{aligned} \tag{33}$$

with $x_1, x_2 \in \{0, 1\}$. During the time evolution, the four different components of the wave function

$$\psi_{x_1,x_2}(y_1, y_2, t) \quad \text{with} \quad x_1, x_2 \in \{0, 1\} , \tag{34}$$

are coupled by the $\mathbf{R_x}^{(1)}$ and $\mathbf{R_x}^{(2)}$ transformations, given by (see Eq. (9) for the notation):

$$\mathbf{R_x}^{(1)}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & 0 & i\sin\left(\frac{\theta}{2}\right) & 0 \\ 0 & \cos\left(\frac{\theta}{2}\right) & 0 & i\sin\left(\frac{\theta}{2}\right) \\ i\sin\left(\frac{\theta}{2}\right) & 0 & \cos\left(\frac{\theta}{2}\right) & 0 \\ 0 & i\sin\left(\frac{\theta}{2}\right) & 0 & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} , \tag{35}$$

$$\mathbf{R_x}^{(2)}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) & 0 & 0 \\ i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) & 0 & 0 \\ 0 & 0 & \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) \\ 0 & 0 & i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} . \tag{36}$$

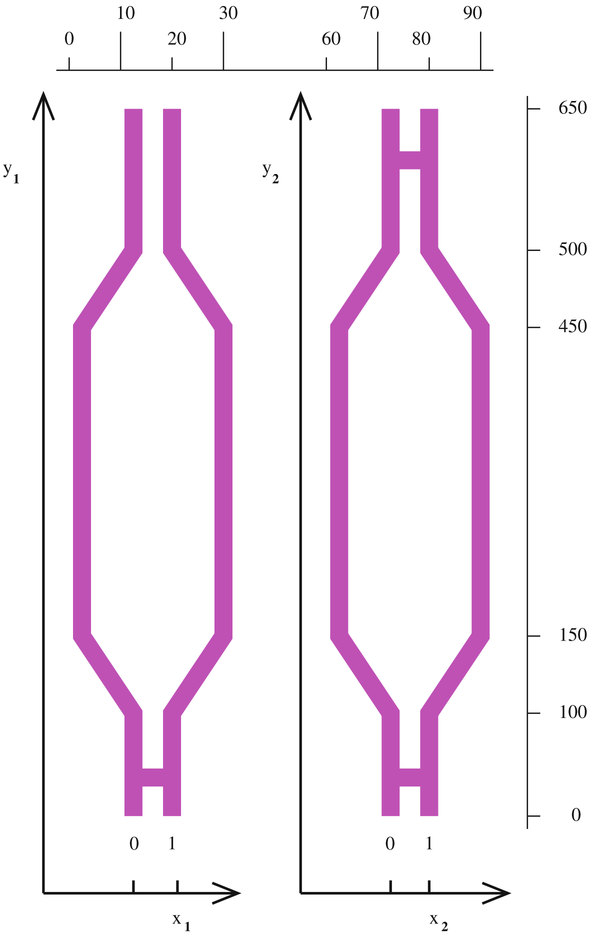
The geometry of the system is contained in the two-particle potential $V_{x_1,x_2}(y_1, y_2)$. It consists of three terms: the two structure potentials along the wires 0 and 1 of each qubit, and the Coulomb interaction between the electrons:

$$V_{x_1,x_2}(y_1, y_2) = U_{x_1}(y_1) + V_{x_2}(y_2) + \frac{e^2}{\epsilon D_{x_1,x_2}(y_1, y_2)} . \tag{37}$$

where $D_{x_1,x_2}(y_1, y_2)$ represents the distance between point y_1 in x_1 wire of the first qubit and point y_2 in x_2 wire of the second qubit and is given explicitly by

$$D_{x_1,x_2}(y_1, y_2) = \sqrt{[p_{x_1}(y_1) - q_{x_2}(y_2)]^2 + [y_1 - y_2]^2} \tag{38}$$

where $p_{x_1}(y_1)$ is the x coordinate of a point that has y coordinate y_1 on the wire representing the x_1 state of the first qubit.



Charge-Based Solid-State Flying Qubits, Figure 7
QWRs network that realizes the transformation $R_x^{(2)}(\pi/2)T(\pi)$ $R_x^{(2)}(\pi/2)R_x^{(1)}(\pi/2)$. The geometry shown is able to produce the four maximally entangled two-qubit Bell states [11]

Since the phase shift $\mathbf{R_0}$ of the first qubit is obtained with a delaying potential barrier inserted along a wire, its functioning is not altered with respect to the full-2D model. The potential barrier is accounted for in potential $U_0(y_1)$. Similarly, for the second qubit the potential is inserted in $V_0(y_1)$. We stress that, within the proposed universal set of gates, the conditional phase shifter is the only two-qubit gate and, as a consequence, the only one able to produce an entangled state.

In conclusion, let us present the result of a numerical simulation of a simple gate network able to produce a maximally entangled Bell's state. The transformation is

$$\mathbf{R_x}^{(1)}(\pi/2)\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)\,, \tag{39}$$

and the network obtained is schematized in Fig. 7. Two electrons are injected in the device. One electron is injected in the wire 1 of the first qubit and one in the wire 0 of the second qubit. The initial state is $|10\rangle$. It is easy to verify that the network described will perform the transformation:

$$\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)\mathbf{R_x}^{(1)}(\pi/2)|10\rangle \\ = \frac{1-i}{2}(|01\rangle + |10\rangle)\,. \tag{40}$$

The numerical simulations performed, and shown in Fig. 8, confirm the good functioning of the device and the production of the entangled Bell's state

$$|\Psi^+\rangle = \frac{(|01\rangle + |10\rangle)}{\sqrt{2}}\,. \tag{41}$$

Future Directions

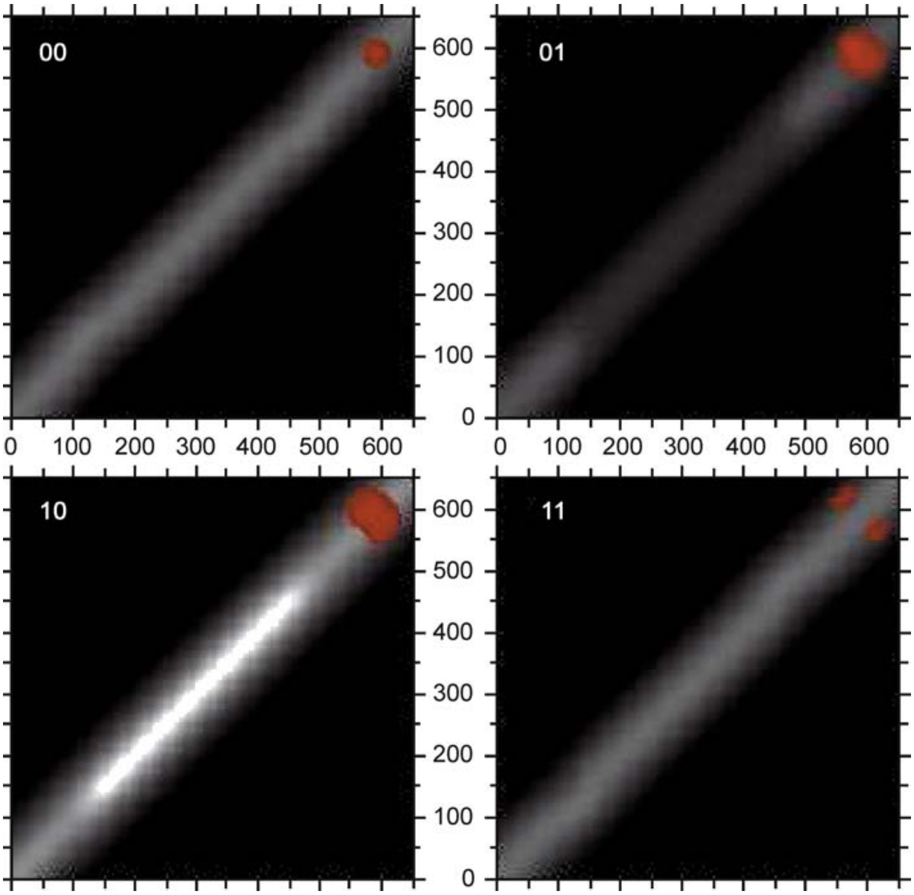
The devices presented in this work have, in principle, the full potentiality to create entangled states of electrons and to perform the logic operations of the universal quantum set of gates. For the modeling of realistic structures, we refer to the literature [1,29,32,38,42].

Concerning experimental realizations of prototypes of the single-qubit gates, we note that Pingue et al. [31] obtained the evidence of switching capabilities of the coupled-wire device, but the presence of occupied localized states in a quantum wire near the coupling window, also predicted in [29], and the Coulomb blockade regime of the experiment, made the device not suitable as a quantum gate.

Both vertical [20,22] and planar [35,36] double-channel structures are currently under investigation. In particular, Ramamoorty et al. [36] obtained switching characteristics whose temperature dependence show a clear signature of coherent behavior of a coupling-window device. It is also worth noting that Fischer et al. demonstrated that the control of the coupling between two modes of two 1D channels is feasible. The two modes were obtained exploiting the two minima of the conduction-band edge in the growth direction of a GaAs 2DEG [21] of two vertically coupled 2DEGs [22]: in both cases clear signatures of coherent coupling have been observed.

Acknowledgment

I would like to express my sincere thanks to the groups of Carlo Jacoboni and Massimo Rudan, that performed, together with the author, a large part of the work here reported. Furthermore, I thank, for most fruitful discussions, Marco Fanciulli and Peter Vogl.



Charge-Based Solid-State Flying Qubits, Figure 8
Square modulus of the two-particle wave function (*red*) at the final time step of the simulation represented in Eq. (40). The four graphs represent the cases: $x_1 = 0, x_2 = 0; x_1 = 0, x_2 = 1; x_1 = 1, x_2 = 0; x_1 = 1, x_2 = 1$, as indicated in the *upper left corners* (see Eq. (33)). The *vertical and horizontal axes* of each graph represent the position y_1 of the first electron and y_2 of the second electron, respectively. The *white region along the diagonals* represent the electron–electron potential given by Eq. (37) [11]

Bibliography

Primary Literature

1. Akguc GB, Reichl LE, Shaji A, Snyder MG (2004) Bell states in a resonant quantum waveguide network. Phys Rev A 69:42303
2. ARDA (2004) Quantum Computation Roadmap. http://qist.lanl.gov. Accessed 1 June 2008
3. Aspect A, Grangier P, Roger G (1981) Experimental tests of realistic local theories via Bell's theorem. Phys Rev Lett 47:460
4. Barenco A, Bennett CH, Cleve R, DiVincenzo DP, Margolus N, Shor P, Sleator T, Smolin JA, Weinfurter H (1995) Elementary gates for quantum computation. Phys Rev A 52:3457
5. Barnes CHW, Shilton JM, Robinson AM (2000) Quantum computation using electrons trapped by surface accoustic waves. Phys Rev B 62:8410
6. Bell JS (1964) On Einstein Podolsky Rosen paradox Physics 1:195
7. Bennett CH (1982) The thermodynamics of computation – a review. Int J Theor Phys 21:905
8. Bertoni A (2007) Perspectives on solid-state flying qubits. J Comp Electr 6:67
9. Bertoni A, Bordone P, Brunetti R, Jacoboni C, Reggiani S (2000) Quantum logic gates based on coherent electron transport in quantum wires. Phys Rev Lett 84:5912
10. Bertoni A, Bordone P, Brunetti R, Jacoboni C, Reggiani S (2002) Numerical simulation of coherent transport in quantum wires for quantum computing. J Mod Opt 49:1219
11. Bertoni A, Ionicioiu R, Zanardi P, Rossi F, Jacoboni C (2002) Simulation of entangled electronic states in semiconductor quantum wires. Physica B 314:10
12. Bertoni A, Reggiani S (2004) Entanglement and quantum computing with ballistic electrons. Semicond Sci Technol 19:S113
13. Bordone P, Bertoni A, Rosini M, Reggiani S, Jacoboni C (2004) Coherent transport in coupled quantum wires assisted by surface acoustic waves. Semicond Sci Technol 19:412
14. Braunstein SL, Mann A, Revzen M (1992) Maximal violation of bell inequalities for mixed states. Phys Rev Lett 68:3259
15. Cunningham J, Pepper M, Talyanskii VI, Ritchie DA (2005)

Acoustic transport of electrons in parallel quantum wires. Acta Phys Polonica A 107:38

16. Einstein A, Podolsky B, Rosen N (1935) Can quantum-mechanical description of physical reality be considered complete? Phys Rev 47:777
17. Ekert A, Jozsa R (1996) Quantum computation and shor's factoring algorithm. Rev Mod Phys 68:733
18. Eugster CC, delAlamo JA, Rooks MJ, Melloch MR (1994) One-dimensional to onedimensional tunneling between electron waveguides. Appl Phys Lett 64:3157
19. Ferry DK (1997) Transport in Nanostructures. Cambridge University Press, Cambridge
20. Fischer SF (2007) Magnetotransport spectroscopy of dual one-dimensional electron systems. Int J Mod Phys B 21:1326
21. Fischer SF, Apetrii G, Kunze U, Schuh D, Abstreiter G (2005) Tunnel-coupled one-dimensional electron systems with large subband separations. Phys Rev B 74:115324
22. Fischer SF, Apetrii G, Kunze U, Schuh D, Abstreiter G (2006) Energy spectroscopy of controlled coupled quantum-wire states. Nature Phys 2:91
23. Furuta S, Barnes CH, Doran CJL(2004) Single-qubit gates and measurements in the surface acoustic wave quantum computer. Phys Rev B 70:205320
24. Governale M, Macucci M, Pellegrini P (2000) Shape of the tunneling conductance peaks for coupled electron waveguides. Phys Rev B 62:4557
25. Ionicioiu R, Amaratunga, Udrea F (2001) Quantum computation with balistic electrons. Int J Mod Phys B 15:125
26. Ionicioiu R, Zanardi P, Rossi F (2001) Testing bell's inequality with ballistic electrons in semiconductors. Phys Rev A 63: 50101
27. Knill E, Laflamme R, Martinez R, Tseng C-H (2000) An algorithmic benchmark for quantum information processing. Nature 404:368
28. Landauer R (1961) Irreversibility and heat generation in the computing process. IBM J Res Dev 5:183
29. Marchi A, Bertoni A, Reggiani S, Rudan M (2004) Investigation on single-electron dynamics in coupled gaas-algaas quantum wires. IEEE Trans Nanotech 3:129
30. Messiah A (1961) Quantum Mechanics. North-Holland, Amsterdam
31. Pingue P, Piazza V, Beltram F, Farrer I, Ritchie DA, Pepper M (2005) Coulomb blockade directional coupler. Appl Phys Lett 86:52102
32. Polizzi E, Ben Abdallah N (2002) Self-consistent three-dimensional models for quantum ballistic transport in open systems. Phys Rev B 66:245301
33. Preskill J (1998) Caltech lecture notes. http://www.theory.caltech.edu/~preskill/ph229. Accessed 1 June 2008
34. Preskill J (1998) Reliable quantum computers. Proc R Soc Lond A 454:3851
35. Ramamoorthy A, Bird JP, Reno JL (2006) Quantum asymmetry of switching in laterally coupled quantum wires with tunable coupling strength. Appl Phys Lett 89:153128
36. Ramamoorthy A, Bird JP, Reno JL (2006) Switching characteristics of coupled quantum wires with tunable coupling strength. Appl Phys Lett 89:13118
37. Rodriquez R, Oi DK, Kataoka M, Barnes CH, Ohshima T, Ekert AK (2005) Surface-acoustic-wave single-electron interferometry. Phys Rev B 72:85329
38. Sabathil M, Mamaluy D, Vogl P (2004) Prediction of a realistic quantum logic gate using the contact block reduction method. Semicond Sci Technol 19:S137
39. Tinkham M (1964) Group Theory and Quantum Mechanics. McGraw-Hill, New York
40. DiVincenzo DP (1995) Two-bit gates are universal for quantum computation. Phys Rev A 50:1015
41. DiVincenzo D (2000) The physical implementation of quantum computation. quant-ph/0002077
42. Zibold T, Vogl P, Bertoni A (2007) Theory of semiconductor quantum-wire based single- and two-qubit gates. Phys Rev B 76:195301

Books and Reviews

Benenti G, Casati G, Strini G (2004) Principles of Quantum Computation and Information, vol 1. World Scientific, Singapore. ISBN 9-812-38858-3

DiVincenzo DP (1995) Quantum Computation. Science 270:255

Harrison P (2005) Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures. Wiley-Interscience, Chichester. ISBN 0-470-01080-0

Hurt NE (2000) Mathematical Physics of Quantum Wires and Devices: From Spectral Resonances to Anderson Localization. Kluwer, Dordrecht. ISBN 0-792-36288-8

Nielsen M, Chuang I (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge. ISBN 0-521-63503-9

Chronological Calculus in Systems and Control Theory

MATTHIAS KAWSKI
Department of Mathematics, Arizona State University, Tempe, USA

Article Outline

Glossary

Controllability A control system is controllable if for every pair of points (*states*) p and q there exists an admissible control such that the corresponding solution curve that starts at p ends at q. Local controllability about a point means that all states in some open neighborhood can be reached.

Pontryagin Maximum Principle of optimal control Optimality of a control-trajectory pair geometrically is a property *dual* to local controllability in the sense that an optimal trajectory (endpoint) lies on the boundary of the reachable sets (after possibly augmenting the state of the system by the running cost). The maximum principle is a necessary condition for optimality. Geometrically it is based on analyzing the effect of families of control variations on the endpoint map. The chronological calculus much facilitates this analysis.

$\mathcal{E}(M) = C^\infty(M)$ The algebra of smooth functions on a finite dimensional manifold M, endowed with the topology of uniform convergence of derivatives of all orders on compact sets.

$\Gamma^\infty(M)$ The space of smooth vector fields on the manifold M.

Chronological calculus An approach to systems theory based on a functional analytic operator calculus, that replaces nonlinear objects such as smooth manifolds by infinite dimensional linear ones, by commutative algebras of smooth functions.

Chronological algebra A linear space with a bilinear product ★ that satisfies the identity $a \star (b \star c) - b \star (a \star c) = (a \star b) \star c - (b \star a) \star c$. This structure arises naturally via the product $(f \star g)_t = \int_0^t [f_s, g_s']\mathrm{d}s$ of time-varying vector fields f and g in the chronological calculus. Here [,] denotes the Lie bracket.

Zinbiel algebra A linear space with a bilinear product that satisfies the identity $a * (b * c) = (a * b) * c + (b * a) * c$. This structure arises naturally in the special case of affine control systems for the product $(U * V)(t) = \int_0^t U(s)V'(s)\mathrm{d}s$ of absolutely continuous scalar valued functions U and V.

The name Zinbiel is Leibniz read backwards, reflecting the duality with Leibniz algebras, a form of noncommutative Lie algebras. There has been some confusion in the literature with Zinbiel algebras incorrectly been called chronological algebras.

$\mathcal{IIF}(\mathcal{U}^Z)$ For a suitable space $\mathcal{U}$ of *time-varying scalars*, e. g. the space of locally absolutely continuous real-valued functions defined on a fixed time interval, and an indexing set Z, $\mathcal{IIF}(\mathcal{U}^Z)$ denotes the space of iterated integral functionals from the space of Z-tuples with values in $\mathcal{U}$ to the space $\mathcal{U}$.

Definition of the Subject

The chronological calculus is a functional analytic operator calculus tool for nonlinear systems theory. The central idea is to replace nonlinear objects by linear ones, in particular, smooth manifolds by commutative algebras of smooth functions. Aside from its elegance, its main virtue is to provide tools for problems that otherwise would effectively be untractable, and to provide new avenues to investigate the underlying geometry. Originally conceived to investigate problems in optimization and control, specifically for extending Pontryagin's Maximum Principle, the chronological calculus continues to establish itself as the preferred language of geometric control theory, and it is spawning new sets of problems, including its own set of algebraic structures that are now studied in their own right.

Introduction, History, and Background

This section starts with a brief historical survey of some landmarks that locate the chronological calculus at the interface of systems and control theory with functional analysis. It is understood that such brief survey cannot possibly do justice to the many contributors. Selected references given are meant to merely serve as starting points for the interested reader.

Many problems in modern systems and control theory are inherently nonlinear and e. g., due to conserved quantities or symmetries, naturally *live* on manifolds rather than on Euclidean spaces. A simple example is the problem of stabilizing the attitude of a satellite via feedback controls. In this case the natural state space is the tangent bundle $TSO(3)$ of a rotation group. The controlled dynamics are described by generally nonautonomous nonlinear differential equations. A key characteristic of their flows is their general lack of commutativity. Solutions of nonlinear differential equations generally do not admit closed form expressions in terms of the traditional sets of elementary functions and symbols. The chronological calculus circumvents such difficulties by reformulating systems and control problems in a different setting which is infinite dimensional, but linear. Building on well established tools and theories from functional analysis, it develops a new formalism and a precise language designed to facilitate studies of nonlinear systems and control theory.

The basic plan of replacing nonlinear objects by linear ones, in particular smooth manifolds by commutative algebras of smooth functions, has a long history. While its roots go back even further, arguably this approach gained much of its momentum with the path-breaking innovations by John von Neumann's work on the *"Mathematical Foundations of Quantum Mechanics"* [104] and Marshall Stone's seminal work on *"Linear Transformations in Hilbert Space"* [89], quickly followed by Israel Gelfand's dissertation on *"Abstract Functions and Linear Operators"* [34] (published in 1938). The fundamental concept of maximal ideal justifies this approach of identifying manifolds with commutative normed rings (algebras), and vice versa. Gelfand's work is described as uniting previously uncoordinated facts and revealing the close connections between classical analysis and abstract functional analysis [102]. In the seventy years since, the study of Banach algebras, C^*-algebras and their brethren has continued to develop into a flourishing research area.

In the different arena of systems and control theory, major innovations at formalizing the subject were made in the 1950s. The Pontryagin Maximum Principle [8] of optimal control theory went far beyond the classical calculus of variations. At roughly the same time Kalman and his peers introduced the Kalman filter [51] for extracting signals from noisy observations, and pioneered state-space approaches in linear systems theory, developing the fundamental concepts of controllability and observability. Linear systems theory has bloomed and grown into a vast array of subdisciplines with ubiquitous applications. Via well-established transform techniques, linear systems lend themselves to be studied in the *frequency domain*. In such settings, systems are represented by linear operators on spaces of functions of a complex variable. Starting in the 1970s new efforts concentrated on rigorously extending linear systems and control theory to nonlinear settings. Two complementary major threads emerged that rely on differential geometric methods, and on operators represented by formal power series, respectively. The first is exemplified in the pioneering work of Brockett [10,11], Haynes and Hermes [43], Hermann [44], Hermann and Krener [45], Jurdjevic and Sussmann [50], Lobry [65], and many others, which focuses on state-space representations of nonlinear systems. These are defined by collections of vector fields on manifolds, and are analyzed using, in particular, Lie algebraic techniques. On the other side, the input-output approach is extended to nonlinear settings primarily through a formal power series approach as initiated by Fliess [31]. The interplay between these approaches has been the subject of many successful studies, in which a prominent role is played by Volterra series and the problem of *realizing* such input-output descriptions as a state-space system, see e. g. Brockett [11], Crouch [22], Gray and Wang [37], Jakubczyk [49], Krener and Lesiak [61], and Sontag and Wang [87].

In the late 1970s Agrachëv and Gamkrelidze introduced into nonlinear control theory the aforementioned abstract functional analytic approach, that is rooted in the work of Gelfand. Following traditions from the physics community, they adopted the name chronological calculus. Again, this abstract approach may be seen as much unifying what formerly were disparate and isolated pieces of knowledge and tools in nonlinear systems theory. While originally conceived as a tool for extending Pontryagin's Maximum Principle in optimal control theory [1], the chronological calculus continues to yield a stream of new results in optimal control and geometry, see e. g. Serres [85], Sigalotti [86], Zelenko [105] for very recent results utilizing the chronological calculus for studying the geometry.

The chronological calculus has led to a very different way of thinking about control systems, epitomized in e. g. the monograph on geometric control theory [5] based on the chronological calculus, or in such forceful advocacies for this approach for e. g. nonholonomic path finding by Sussmann [95]. Closely related are studies of nonlinear controllability, e. g. Agrachëv and Gamkrelidze [3,4], Tretyak [96,97,98], and Vakhrameev [99], including applications to controllability of the Navier–Stokes equation by Agrachëv and Sarychev [6]. The chronological calculus also lends itself most naturally to obtaining new results in averaging theory as in Sarychev [83], while Cortes and Martinez [20] used it in motion control of mechanical systems with symmetry and Bullo [13,68] for vibrational control of mechanical systems. Noteworthy applications include locomotion of robots by Burdick and Vela [100], and even robotic fish [71]. Instrumental is its interplay with series expansion as in Bullo [12,21] that utilize affine connections of mechanical systems. There are further applications to stability and stabilization Caiado and Sarychev [15,82], while Monaco et.al. [70] extended this approach to discrete-time dynamics and Komleva and Plotnikov used it in differential game theory [60].

Complementing such *applications* are new subjects of study such as the abstract algebraic structures that underlie the chronological calculus. The chronological algebra itself has been the subject of study as early as Agrachëv and Gamkrelidze [2]. The closely related structure of Zinbiel structures has recently found more attention, see work by Dzhumadil'daev [25,26], Kawski and Sussmann [57] and Kawski [54,55]. Zinbiel algebras arise in the special case when the dynamics ("*time-varying vector fields*") splits into a sum of products of time-varying coefficients and autonomous vector fields. There is an unfortunate confusion of terms in the literature as originally Zinbiel algebras had also been called chronological algebras. Recent usage disentangles these closely related, but distinct, structures and reflects the primacy of the latter term coined by Loday [66,67] who studies Leibniz algebras (which appear in cyclic homology). *Zinbiel* is simply *Leibniz* spelled backwards, a choice which reflect that Leibniz and Zinbiel are dual operands in the sense of Koszul duality as investigated by Ginzburg and Kapranov [36].

Fundamental Notions of the Chronological Calculus

From a Manifold to a Commutative Algebra

This and the next sections very closely follow the introductory exposition of Chapter 2 of Agrachëv and Sachkov [5], which is also recommended for the full technical and analytical details regarding the topology and convergence.

The objective is to develop the basic tools and formalism that facilitate the analysis of generally nonlinear systems that are defined on smooth manifolds. Rather than primarily considering points on a smooth manifold M, the key idea is to instead focus on the commutative algebra of $\mathcal{E}(M) = C^\infty(M, \mathbb{R})$ of real-valued smooth functions on M. Note that $\mathcal{E}(M)$ not only has the structure of a vector space over the field $\mathbb{R}$, but it also inherits the structure of a commutative ring under pointwise addition and multiplication from the codomain $\mathbb{R}$.

Every point $p \in M$ gives rise to a functional $\hat{p}\colon \mathcal{E}(M) \mapsto \mathbb{R}$ defined by $\hat{p}(\varphi) = \varphi(p)$. This functional is linear and multiplicative, and is a homomorphism of the algebras $\mathcal{E}(M)$ and $\mathbb{R}$: For every $p \in M$, for $\varphi, \psi \in \mathcal{E}(M)$ and $t \in \mathbb{R}$ the following hold

$$\hat{p}(\varphi + \psi) = \hat{p}\varphi + \hat{p}\varphi, \quad \hat{p}(\varphi \cdot \psi) = (\hat{p}\varphi) \cdot (\hat{p}\psi)\,, \quad \text{and} \quad \hat{p}(t\varphi) = t \cdot (\hat{p}\varphi)\,.$$

Much of the power of this approach derives from the fact that this correspondence is invertible: For a nontrivial multiplicative linear functional $\theta\colon \mathcal{E}(M) \mapsto \mathbb{R}$ consider its kernel $\ker\theta = \{\varphi \in \mathcal{E}(M)\colon \theta\varphi = 0\}$. A critical observation is that this is a maximal ideal, and that it must be of the form $\{\varphi \in \mathcal{E}(M)\colon \varphi(p) = 0\}$ for some, uniquely defined, $p \in M$. For the details of a proof see appendix A.1. of [5].

Proposition 1 *For every nontrivial multiplicative linear functional $\theta\colon \mathcal{E}(M) \mapsto \mathbb{R}$ there exists $p \in M$ such that $\theta = \hat{p}$.*

Note on the side, that there may be maximal ideals in the space of all multiplicative linear functionals on $\mathcal{E}(M)$ that do not correspond to any point on M – e. g. the ideal of all linear functionals that vanish on every function with compact support. But this does not contradict the stated proposition.

Not only can one *recover* the manifold M as a set from the commutative ring $\mathcal{E}(M)$, but using the weak topology on the space of linear functionals on $\mathcal{E}(M)$ one also recovers the topology on M

$$p_n \longrightarrow p \ \text{ if and only if } \ \forall f \in \mathcal{E}(M),\ \widehat{p_n} f \longrightarrow \hat{p} f\,. \quad (1)$$

The smooth structure on M is recovered from $\mathcal{E}(M)$ in a trivial way: A function g on the space of multiplicative linear functionals $\hat{p}\colon p \in M$ is smooth if and only if there exists $f \in \mathcal{E}(M)$ such that for every $\hat{p}$, $g(\hat{p}) = \hat{p}f$.

In modern differential geometry it is routine to identify tangent vectors to a smooth manifold with either equivalence classes of smooth curves, or with first order partial differential operators. In this context, tangent vectors at a point $q \in M$ are derivations of $\mathcal{E}(M)$, that is, linear functionals $\hat{f}_q$ on $\mathcal{E}(M)$ that satisfy the Leibniz rule. Using $\hat{q}$, this means for every $\varphi, \psi \in \mathcal{E}(M)$,

$$\hat{X}_q(\varphi\psi) = (\hat{X}_q\varphi)(\hat{q}\psi) + (\hat{q}\varphi)(\hat{X}_q\psi)\,. \quad (2)$$

Smooth vector fields on M correspond to linear functionals $\hat{f}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ that satisfy for all $\varphi, \psi \in \mathcal{E}(M)$

$$\hat{f}(\varphi\psi) = (\hat{f}\varphi) \cdot \psi + \varphi \cdot (\hat{f}\psi)\,. \quad (3)$$

Again, the correspondence between tangent vectors and vector fields and the linear functionals as above is invertible. Write $\Gamma^\infty(M)$ for the space of smooth vector fields on M.

Finally, there is a one-to-one correspondence between smooth diffeomorphisms $\Phi\colon M \mapsto M$ and automorphisms of $\mathcal{E}(M)$. The map $\hat{\Phi}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ defined for $p \in M$ and $\varphi \in \mathcal{E}(M)$ by $\hat{\Phi}(\varphi)(p) = \varphi(\Phi(p))$ clearly has the desired properties. For the reverse direction, suppose $\Psi\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ is an automorphism. Then for every $p \in M$ the map $\hat{p} \circ \Psi\colon \mathcal{E}(M) \mapsto \mathbb{R}$ is a nontrivial linear multiplicative functional, and hence equals $\hat{q}$ for some $q \in M$. It is easy to see that the map $\Phi\colon M \mapsto M$ is indeed a diffeomorphism, and $\Psi = \hat{\Phi}$.

In the sequel we shall omit the *hats*, and simply write, say, p for the linear functional $\hat{p}$. The role of each object is usually clear from the order. For example, for a point p, a smooth function φ, smooth vector fields f, g, h, with flow e^{th} and its tangent map $\left(e^{th}\right)_*$, what in traditional format might be expressed as $(\left(e^{th}\right)_* g)\varphi(e^{\tau f}(p))$ is simply written as $pe^{\tau f} g e^{th}\varphi$, not requiring any parentheses.

Frechet Space and Convergence

The usual topology on the space $\mathcal{E}(M)$ is the one of uniform convergence of all derivatives on compact sets, i. e., a sequence of functions $\{\varphi_k\}_{k=1}^\infty \subseteq \mathcal{E}(M)$ converges to $\phi \in \mathcal{E}(M)$ if for every finite sequence $f_1, f_2, \ldots, f_s$ of smooth vector fields on M and every compact set $K \subseteq M$ the sequence $\{f_s \ldots f_2 f_1 \varphi_k\}_{k=1}^\infty$ converges uniformly on K to $f_s \ldots f_2 f_1 \varphi$.

This topology is also obtained by a countable family of semi-norms $\|\cdot\|_{s,K}$ defined by

$$\|\varphi\|_{s,K} = \sup\{\, |p f_s \ldots f_2 f_1 \varphi| : \\ p \in K,\ f_i \in \Gamma^\infty(M),\ s \in \mathbb{Z}^+ \,\} \quad (4)$$

where K ranges over a countable collection of compact subsets whose union is all of M. In an analogous way define semi-norms of smooth vector fields $f \in \Gamma^\infty(M)$

$$\|f\|_{s,K} = \sup\{\ \|f\varphi\|_{s,K} : \ \|f\varphi\|_{s+1,K} = 1\ \}\,. \tag{5}$$

Finally, for every smooth diffeomorphism Φ of M, $s \in \mathbb{Z}^+$ and $K \subseteq M$ compact there exist $C_{s,K,\Phi} \in \mathbb{R}$ such that for all $\varphi \in \mathcal{E}(M)$

$$\|\Phi\varphi\|_{s,K} \ \leq\ C_{s,K,\Phi}\|\varphi\|_{s,\varphi(K)}\,. \tag{6}$$

Regularity properties of one-parameter families of vector fields and diffeomorphisms (*"time-varying vector fields and diffeomorphisms"*) are understood in the weak sense. In particular, for a family of smooth vector fields $f_t \in \Gamma^\infty(M)$, its integral and derivative (if they exist) are defined as the operators that satisfy for every $\varphi \in \mathcal{E}(M)$

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}f_t\right)\varphi = \frac{\mathrm{d}}{\mathrm{d}t}\left(f_t\varphi\right) \quad \text{and} \quad \left(\int_a^b f_t \mathrm{d}t\right)\varphi = \int_a^b \left(f_t\varphi\right)\mathrm{d}t\,. \tag{7}$$

The convergence of series expansions of vector fields and of diffeomorphisms encountered in the sequel are to be interpreted in the analogous weak sense.

The Chronological Exponential

This section continues to closely follow [5] which contains full details and complete proofs. On a manifold M consider generally time-varying differential equations of the form $\frac{\mathrm{d}}{\mathrm{d}t}q_t = f_t(q_t)$. To assure existence and uniqueness of solutions to initial value problems, make the typical regularity assumptions, namely that in every coordinate chart $U \subseteq M$, $x \colon U \mapsto \mathbb{R}^n$ the vector field $(x_* f_t)$ is (i) measurable and locally bounded with respect to t for every fixed x and (ii) smooth with locally bounded partial derivatives with respect to x for every fixed t. For the purposes of this article and for clarity of exposition, also assume that vector fields are complete. This means that solutions to initial value problems are defined for all times $t \in \mathbb{R}$. This is guaranteed if, for example, all vector fields considered vanish identically outside a common compact subset of M.

As in the previous sections, for each fixed t interpret $q_t \colon \mathcal{E}(M) \mapsto \mathbb{R}$ as a linear functional on $\mathcal{E}(M)$, and note that this family satisfies the time-varying, but linear differential equation, suggestively written as

$$\dot{q}_t = q_t \circ f_t\,. \tag{8}$$

on the space of linear functionals on $\mathcal{E}(M)$. It may be shown that under the above assumptions it has a unique solution, called the *right chronological exponential* of the vector field f_t as the corresponding flow. Formally, it satisfies for almost all $t \in \mathbb{R}$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right) = \left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right) \circ f_t\,. \tag{9}$$

Analogously the *left chronological exponential* satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right) = f_t \circ \left(\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right)\,. \tag{10}$$

Formally, one obtains a series expansion for the chronological exponentials by rewriting the differential equation as an integral equation and solving it by iteration

$$q_t = q_0 + \int_0^t q(\tau)\circ f_\tau\,\mathrm{d}\tau \tag{11}$$

$$= q_0 + \int_0^t \left(q_0 + \int_0^\tau q(\sigma)\circ f_\sigma\, d\sigma\right)\circ f_\tau\,\mathrm{d}\tau \tag{12}$$

to eventually, formally, obtain the expansion

$$\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \sim \mathrm{Id} + \sum_{k=1}^{\infty}\int_0^t\int_0^{t_k}\cdots\int_0^{t_2} \cdot f_{\tau_k}\circ\cdots\circ f_{\tau_2}\circ f_{\tau_1}\,\mathrm{d}\tau_k\ldots\mathrm{d}\tau_2\,\mathrm{d}\tau_1 \tag{13}$$

and analogously for the left chronological exponential

$$\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \sim \mathrm{Id} + \sum_{k=1}^{\infty}\int_0^t\int_0^{t_k}\cdots\int_0^{t_2} \cdot f_{\tau_1}\circ f_{\tau_2}\circ\cdots\circ f_{\tau_k}\,\mathrm{d}\tau_k\ldots\mathrm{d}\tau_2\,\mathrm{d}\tau_1\,. \tag{14}$$

While this series never converges, not even in a weak sense, it nonetheless has an interpretation as an asymptotic expansion. In particular, for any fixed function $\phi \in \mathcal{E}(M)$ and any semi-norm as in the previous section, on any compact set one obtains an error estimate for the remainder after truncating the series at any order N which establishes that the truncation error is of order $O(t^N)$ as $t \to 0$. When one considers the restrictions to any f_t-invariant normed linear subspace $L \subseteq \mathcal{E}(M)$, i. e. for all t, $f_t(L) \subseteq L$, on which f_t is bounded, then the asymptotic series converges to the chronological exponential and it satisfies for every $\varphi \in L$

$$\left\|\left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right)\varphi\right\| \ \leq\ \mathrm{e}^{\int_0^t\|f_\tau\|\,\mathrm{d}\tau}\ \|\varphi\|\,. \tag{15}$$

In the case of analytic vector fields f_τ and analytic functions φ one obtains convergence of the series for sufficiently small t.

While in general there is no reason why the vector fields f_t should commute at different times, the chronological exponentials nonetheless still share some of the usual properties of exponential and flows, for example the composition of chronological exponentials satisfies for all $t_i \in \mathbb{R}$

$$\left(\overrightarrow{\exp}\int_{t_1}^{t_2} f_\tau \, \mathrm{d}\tau\right) \circ \left(\overrightarrow{\exp}\int_{t_2}^{t_3} f_\tau \, \mathrm{d}\tau\right) = \left(\overrightarrow{\exp}\int_{t_1}^{t_3} f_\tau \, \mathrm{d}\tau\right) . \tag{16}$$

Moreover, the left and right chronological exponentials are inverses of each other in the sense that for all $t_0, t_1 \in \mathbb{R}$

$$\begin{aligned}\left(\overrightarrow{\exp}\int_{t_0}^{t_1} f_\tau \, \mathrm{d}\tau\right)^{-1} &= \overrightarrow{\exp}\int_{t_1}^{t_0} f_\tau \, \mathrm{d}\tau \\ &= \overleftarrow{\exp}\int_{t_0}^{t_1} (-f_\tau) \, \mathrm{d}\tau \, .\end{aligned} \tag{17}$$

Variation of Parameters and the Chronological Logarithm

In control one rarely considers only a single vector field, and, instead, for example, is interested in the interaction of a perturbation or control vector field with a reference or drift vector field. In particular, in the nonlinear setting, this is where the chronological calculus very much facilitates the analysis. For simplicity consider a differential equation defined by two, generally time varying, vector fields (with the usual regularity assumptions)

$$\dot{q} = f_t(q) + g_t(q) \, . \tag{18}$$

The objective is to obtain a formula for the flow of the field $(f_t + g_t)$ as a *perturbation* of the flow of f_t. Writing

$$\Phi_t = \overrightarrow{\exp}\int_0^t f_\tau \, \mathrm{d}\tau \quad \text{and} \quad \Theta_t = \overrightarrow{\exp}\int_0^t (f_\tau + g_\tau) \, \mathrm{d}\tau , \tag{19}$$

this means one is looking for a family of operators Ψ_t such that for all t

$$\Theta_t = \Psi_t \circ \Phi_t \, . \tag{20}$$

Differentiating and using the invertibility of the flows one obtains a differential equation in the standard form

$$\begin{aligned}\dot{\Psi}(t) \;=\; \Psi_t \circ \Phi_t \circ g_t \circ \Phi_t^{-1} \;&=\; \Psi_t \circ (\mathrm{Ad}\Phi_t) \; g_t \\ &= \Psi_t \circ \left(\overrightarrow{\exp}\int_0^t \mathrm{ad} f_\tau \, \mathrm{d}\tau\right) \; g_t \, ,\end{aligned} \tag{21}$$

which has the unique solution

$$\Psi_t = \overrightarrow{\exp}\int_0^t \left(\overrightarrow{\exp}\int_0^\tau \mathrm{ad} f_\sigma \, \mathrm{d}\sigma\right) \; g_\tau \, \mathrm{d}\tau \, , \tag{22}$$

and consequently one has the *variations formula*

$$\begin{aligned}\overrightarrow{\exp}\int_0^t (f_\tau + g_\tau) \, \mathrm{d}\tau = \Bigg(\overrightarrow{\exp}\int_0^t \left(\overrightarrow{\exp}\int_0^\tau \mathrm{ad} f_\sigma \, \mathrm{d}\sigma\right) \\ \cdot \; g_\tau \, \mathrm{d}\tau\Bigg) \circ \left(\overrightarrow{\exp}\int_0^t f_\tau \, \mathrm{d}\tau\right) \, .\end{aligned} \tag{23}$$

This formula is of fundamental importance and used ubiquitously for analyzing controllability and optimality: One typically considers f_t as defining a reference dynamical system and then considers the effect of adding a control term g_t. A special application is in the theory of averaging where g_t is considered a perturbation of the field f_t, which in the most simple form may be assumed to be time-periodic. The monograph [81] by Sanders and Verhulst is the classical reference for the theory and applications of averaging using traditional language. The chronological calculus, in particular the above variations formula, have been used successfully by Sarychev [83] to investigate in particular higher order averaging, and by Bullo [13] for averaging in the context of mechanical systems and its interplay with mechanical connections. Instrumental to the work [83] is the notion of the chronological logarithm, which is motivated by rewriting the defining Eqs. (9) and (10) [2,80] as

$$\begin{aligned}f_t &= \left(\overrightarrow{\exp}\int_0^t f_\tau \, \mathrm{d}\tau\right)^{-1} \circ \frac{\mathrm{d}}{\mathrm{d}t}\left(\overrightarrow{\exp}\int_0^t f_\tau \, \mathrm{d}\tau\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\overleftarrow{\exp}\int_0^t f_\tau \, \mathrm{d}\tau\right) \circ \left(\overleftarrow{\exp}\int_0^t f_\tau \, \mathrm{d}\tau\right)^{-1} \, .\end{aligned} \tag{24}$$

While in general the existence of the logarithm is a delicate question [62,83], under suitable hypotheses it is instrumental for the derivation of the nonlinear Floquet theorem in [83].

Theorem 1 (Sarychev [83]) *Suppose $f_t = f_{t+1}$ is a period-one smooth vector field generating the flow Φ_t and $\Lambda = \log \Phi_1$, then there exists a period-one family of diffeomorphisms $\Psi_t = \Psi_{t+1}$ such that for all t $\Phi_t = e^{t\Lambda} \circ \Psi_t$.*

The logarithm of the chronological exponential of a time-varying vector field εf_t may be expanded as a formal chronological series

$$\log \overrightarrow{\exp}\left(\int_0^1 \varepsilon f_\tau \, \mathrm{d}\tau\right) \;=\; \sum_{j=1}^{\infty} \varepsilon^j \Lambda^j \, . \tag{25}$$

The first three terms of which are given explicitly [83] as

$$\Lambda^1 = \int_0^1 f_\tau \ \mathrm{d}\tau \,, \quad \Lambda^2 = \frac{1}{2}\int_0^1 \left[\int_0^\tau f_\sigma \, \mathrm{d}\sigma \ , \ f_\tau \right] \ \mathrm{d}\tau,$$
$$\text{and} \quad \Lambda^3 = -\frac{1}{2}[\Lambda^1, \Lambda^2] + \frac{1}{3}\int_0^1 \mathrm{ad}^2 \left(\int_0^\tau f_\sigma \, \mathrm{d}\sigma \right) f_\tau \ \mathrm{d}\tau \,. \qquad (26)$$

Aside from such applications as averaging, the main use of the chronological calculus has been for studying the dual problems of optimality and controllability, in particular controllability of families of diffeomorphisms [4]. In particular, this approach circumvents the limitations encountered by more traditional approaches that use parametrized families of control variations that are based on a finite number of switching times. However, since [52] it is known that a general theory must allow also, e. g., for increasing numbers of switchings and other more general families of control variations. The chronological calculus is instrumental to developing such general theory [4].

Chronological Algebra

When using the chronological calculus for studying controllability, especially when differentiating with respect to a parameter, and also in the aforementioned averaging theory, the following *chronological product* appears almost everywhere: For two time-varying vector fields f_t and g_t that are absolutely continuous with respect to t, define their chronological product as

$$(f \star g)_t = \int_0^t [f_s, g_s']\mathrm{d}s \qquad (27)$$

where $[\cdot, \cdot]$ denotes the Lie bracket. It is easily verified that this product is generally neither associative nor satisfies the Jacobi identity, but instead it satisfies the chronological identity

$$x \star (y \star z) - y \star (x \star z) = (x \star y) \star z - (y \star x) \star z \,. \quad (28)$$

One may define an abstract chronological algebra (over a field $\mathbf{k}$) as a linear space that is equipped with a bilinear product that satisfies the chronological identity (28). More compactly, this property may be written in terms of the left translation λ that associates with any element $x \in \mathcal{A}$ of a not-necessarily associative algebra $\mathcal{A}$ the map $\lambda_x \colon y \mapsto xy$. Using this λ, the chronological identity (28) is simply

$$\lambda_{[x,y]} = [\lambda_x, \lambda_y] \,, \qquad (29)$$

i. e., the requirement that the map $x \mapsto \lambda_x$ is a homomorphism of the algebra $\mathcal{A}$ with the commutator product $(x, y) \mapsto xy - yx$ into the Lie algebra of linear maps from $\mathcal{A}$ into $\mathcal{A}$ under its commutator product.

Basic general properties of chronological algebras have been studied in [2], including a discussion of free chronological algebras (over a fixed generating set S). Of particular interest are bases of such free chronological algebras as they allow one to minimize the number of terms in series expansions by avoiding redundant terms. Using the graded structure, it is shown in [2] that an ordered basis of a free chronological algebra over a set S may be obtained recursively from products of the form $\lambda_{b_1}\lambda_{b_2} \cdots \lambda_{b_k} s$ where $s \in S$ is a generator and $b_1 \preccurlyeq b_2 \preccurlyeq \cdots \preccurlyeq b_k$ are previously constructed basis elements. According to [2], the first elements of such a basis over the single-element generating set $S = \{s\}$ may be chosen as (using juxtaposition for multiplication)

$$\begin{aligned} &b_1 = s \,, \quad b_2 = \lambda_{b_1} s = s^2 \,, \quad b_3 = \lambda_{b_2} s = s^2 s \,, \\ &b_4 = \lambda_{b_1}\lambda_{b_1} s = ss^2 \,, \quad b_5 = \lambda_{b_3} s = (s^2 s)s \,, \\ &b_6 = \lambda_{b_4} s = (ss^2)s \,, \quad b_7 = \lambda_{b_1}\lambda_{b_2} s = s(s^2 s) \,, \\ &\qquad b_8 = \lambda_{b_2}\lambda_{b_1} s = s(ss^2) \,, \ldots \qquad (30) \end{aligned}$$

Using only this algebraic and combinatorial structure, one may define exponentials (and logarithms) and analyze the group of formal flows [2]. Of particular interest systems and control is an explicit formula that allows one to express products (compositions) of two or more exponentials (flows) as a single exponential. This is tantamount to a formula for the logarithm of a product of exponentials of two noncommuting indeterminates X and Y. The classical version is known as the Campbell–Baker–Hausdorff formula [16], the first few terms being

$$\mathrm{e}^X \cdot \mathrm{e}^Y = \mathrm{e}^{X+Y+\frac{1}{2}[X,Y]+\frac{1}{12}[X-Y,[X,Y]]-\frac{1}{24}[X,[Y,[X,Y]]]-\ldots} \,. \qquad (31)$$

A well-known short-coming of the classical formula is that the higher-order iterated Lie brackets appearing in this formula are always linearly dependent and hence the coefficients are never well-defined. Hence one naturally looks for more compact, minimal formulas which avoid such redundancies. The chronological algebra provides one such alternative and has been developed and utilized in [2]. The interested reader is referred to the original reference for technical details.

Systems That Are Affine in the Control

Separating Time-Varying Control and Geometry

The chronological calculus was originally designed for work with nonstationary (time-varying) families of vector fields, and it arguably reaps the biggest benefits in that setting where there are few other powerful tools available. Nonetheless, the chronological calculus also much streamlines the analysis of autonomous vector fields. The best studied case is that of affine control systems in which the time-varying control can be separated from the geometry determined by autonomous vector fields. In classical notation such control systems are written in the form

$$\dot{x} = u_1(t) f_1(x) + \dots + u_m(t) f_m(x) \,. \tag{32}$$

where f_i are vector fields on a manifold M and u is a measurable function of time taking values typically in a compact subset $U \subseteq \mathbb{R}^m$. This description allows one to also accommodate physical systems that have an *uncontrolled drift term* by simply fixing $u_1 \equiv 1$. For the sake of clarity, and for best illustrating the kind of results available, assume that the fields f_i are smooth and complete. Later we specialize further to real analytic vector fields. This set-up does not necessarily require an interpretation of the u_i as controls: they may equally well be disturbances, or it may simply be a dynamical system which splits in this convenient way.

As a starting point consider families of piecewise constant control functions $u\colon [0, T] \mapsto U \subseteq \mathbb{R}^m$. On each interval $[t_j, t_{j+1}]$ on which the control is constant $u_{[t_j, t_{j+1}]} = u^{(j)}$, the right hand side of (32) is a fixed vector field $g_j = u_1^{(j)} f_1 + \dots u_m^{(j)} f_m$. The endpoint of the solution curve starting from $x(0) = p$ is computed as a directed product of ordinary exponentials (flows) of vector fields

$$p\mathrm{e}^{(t_1 - t_0) g_1} \circ \mathrm{e}^{(t_2 - t_1) g_2} \circ \dots \mathrm{e}^{(T - t_{s-1}) g_s} \,. \tag{33}$$

But even in this case of autonomous vector fields the chronological calculus brings substantial simplifications. Basically this means that rather than considering the expression (33) as a point on M, consider this product as a functional on the space $\mathcal{E}(M)$ of smooth functions on M. To reiterate the benefit of this approach [5,57] consider the simple example of a tangent vector to a smooth curve $\gamma\colon (-\varepsilon, \varepsilon) \mapsto M$ which one might want to define as

$$\dot{\gamma}(0) = \lim_{t \to 0} \frac{1}{t} (\gamma(t) - \gamma(0)) \,. \tag{34}$$

Due to the lack of a linear structure on a general manifold, with a classical interpretation this does not make any sense at all. Nonetheless, when interpreting $\gamma(t)$, $\gamma(0)$, $\gamma'(0)$, and $\frac{1}{t}(\gamma(t) - \gamma(0))$ as linear functionals on the space $\mathcal{E}(M)$ of smooth functions on M this is perfectly meaningful. The meaning of the limit can be rigorously justified as indicated in the preceding sections, compare also [1].

A typical example to illustrate how this formalism almost trivializes important calculations involving Lie brackets is the following, compare [57]. Suppose f and g are smooth vector fields on the manifold M, $p \in M$, and $t \in R$ is sufficiently small in magnitude. Using that $\frac{\mathrm{d}}{\mathrm{d}t} p\mathrm{e}^{tf} = p\mathrm{e}^{tf} f$ and so on, one immediately calculates

$$\begin{aligned} &(\mathrm{d}/\mathrm{d}t)(p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg}) \\ &\quad = p\mathrm{e}^{tf} f \mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg} g \mathrm{e}^{-tf}\mathrm{e}^{-tg} \\ &\quad - p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} f \mathrm{e}^{-tg} - p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} g \,. \end{aligned} \tag{35}$$

In particular, at $t = 0$ this expression simplifies to $pf + pg - pf - pg = 0$. Analogously the second derivative is calculated as

$$\begin{aligned} &\frac{\mathrm{d}^2}{\mathrm{d}t^2}(p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg}) = p\mathrm{e}^{tf} f^2 \mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} + p\mathrm{e}^{tf} f \\ &\cdot \mathrm{e}^{tg} g \mathrm{e}^{-tf}\mathrm{e}^{-tg} - p\mathrm{e}^{tf} f \mathrm{e}^{tg}\mathrm{e}^{-tf} f \mathrm{e}^{-tg} - p\mathrm{e}^{tf} f \mathrm{e}^{tg}\mathrm{e}^{-tf} \\ &\cdot \mathrm{e}^{-tg} g + p\mathrm{e}^{tf} f \mathrm{e}^{tg} g \mathrm{e}^{-tf}\mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg} g^2 \mathrm{e}^{-tf}\mathrm{e}^{-tg} \\ &- p\mathrm{e}^{tf}\mathrm{e}^{tg} g \mathrm{e}^{-tf} f \mathrm{e}^{-tg} - p\mathrm{e}^{tf}\mathrm{e}^{tg} g \mathrm{e}^{-tf}\mathrm{e}^{-tg} g - p\mathrm{e}^{tf} f \\ &\cdot \mathrm{e}^{tg}\mathrm{e}^{-tf} f \mathrm{e}^{-tg} - p\mathrm{e}^{tf}\mathrm{e}^{tg} g \mathrm{e}^{-tf} f \mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} \\ &\cdot f^2 \mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} f \mathrm{e}^{-tg} g - p\mathrm{e}^{tf} f \mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} g \\ &- p\mathrm{e}^{tf}\mathrm{e}^{tg} g \mathrm{e}^{-tf}\mathrm{e}^{-tg} g + p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} f \mathrm{e}^{-tg} g + p\mathrm{e}^{tf} \\ &\cdot \mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} g^2 \end{aligned} \tag{36}$$

which at $t = 0$ evaluates to

$$\begin{aligned} &pf^2 + pfg - pf^2 - pfg + pfg + pg^2 - pgf - pg^2 \\ &- pf^2 - pgf + pf^2 + pfg - pfg - pg^2 + pfg + pg^2 \\ &= 2pfg - 2pgf = 2p[f, g] \,. \end{aligned}$$

While this simple calculation may not make sense on a manifold with a classical interpretation in terms of points and vectors on a manifold, it does so in the context of the chronological calculus, and establishes the familiar formula for the Lie bracket as the limit of a composition of flows

$$p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} = p + t^2 p[f, g] + O(t^2) \;\text{ as }\; t \to 0 \,. \tag{37}$$

Similar examples may be found in [5] (e. g. p. 36), as well as analogous simplifications for the variations formula for autonomous vector fields ([5] p. 43).

Asymptotic Expansions for Affine Systems

Piecewise constant controls are a starting point. But a general theory requires being able to take appropriate limits and demands completeness. Typically, one considers measurable controls taking values in a compact subspace $U \subseteq \mathbb{R}^m$, and uses $L^1([0,t],U)$ as the space of admissible controls. Correspondingly the finite composition of exponentials as in (33) is replaced by continuous formulae. These shall still separate the effects of the time-varying controls from the underlying geometry determined by the autonomous vector fields. This objective is achieved by the Chen–Fliess series which may be interpreted in a number of different ways. Its origins go back to the 1950s when K. T. Chen [18], studying geometric invariants of curves in $\mathbb{R}^n$, associated to each smooth curve a formal noncommutative power series. In the early 1970s, Fliess [30,31] recognized the utility of this series for studying control systems. Using a set $X_1, X_2, \ldots, X_m$ of indeterminates, the Chen–Fliess series of a measurable control $u \in L^1([0,t],U)$ as above is the formal power series

$$S_{\mathrm{CF}}(T,u) = \sum_I \int_0^T u_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} u_{i_2}(t_2) \cdot \int_0^{t_2} u_{i_1}(t_1)\,\mathrm{d}t_1 \ldots \mathrm{d}t_s \cdot X_{i_1} \ldots X_{i_s} \quad (38)$$

where the sum ranges over all multi-indices $I = (i_1, \ldots, i_s)$, $s \geq 0$ with $i_j \in \{1,2\ldots m\}$. This series gives rise to an asymptotic series for solution curves of the system (32). For example, in the analytic setting, following [91], one has:

Theorem 2 *Suppose f_i are analytic vector fields on $\mathbb{R}^n$ $\phi\colon \mathbb{R}^n \mapsto \mathbb{R}$ is analytic and $U \subset \mathbb{R}^m$ is compact. Then for every compact set $K \subseteq \mathbb{R}^n$, there exists $T > 0$ such that the series*

$$S_{\mathrm{CF},f}(T,u,p)(\varphi) = \sum_I \int_0^T u_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} u_{i_2}(t_2) \cdot \int_0^{t_2} u_{i_1}(t_1)\,dt_1 \ldots dt_s \cdot p f_{i_1} \ldots f_{i_s} \varphi \quad (39)$$

converges uniformly to the solution of (32) for initial conditions $x(0) = p \in K$ and $u\colon [0,T] \mapsto U$ measurable.

Here the series $S_{\mathrm{CF},f}(T,u,p)$ is, for every fixed triple (f,u,p), interpreted as an element of the space of linear functionals on $\mathcal{E}(M)$. But one may abstract further. In particular, for each fixed multi-index I as above, the iterated integral coefficient is itself a functional that takes as input a control function $u \in \mathcal{U}^m$ and maps it to the corresponding iterated integral. It is convenient to work with the primitives $U_j\colon t \mapsto \int_0^t u_j(s)\,ds$ of the control functions u_j rather than the controls themselves. More specifically, if e. g. $\mathcal{U} = AC([0,T],[-1,1])$ then for every multi-index $I = (i_1, \ldots, i_s) \in \{1, \ldots, m\}^s$ as above one obtains the iterated integral functional $\Upsilon_I\colon \mathcal{U}^m \mapsto \mathcal{U}$ defined by

$$\Upsilon_I\colon U \mapsto \int_0^T U'_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} U'_{i_2}(t_2) \cdot \int_0^{t_2} U'_{i_1}(t_1)\,\mathrm{d}t_1 \mathrm{d}t_2 \ldots \mathrm{d}t_s\,. \quad (40)$$

Denoting by $\mathcal{IIF}(\mathcal{U}^Z)$ the linear space of iterated integral functionals spanned by the functionals of the above form, the map Υ maps multi-indices to $\mathcal{IIF}(\mathcal{U}^Z)$. The algebraic and combinatorial nature of these spaces and this map will be explored in the next section.

On the other side, there is the map $\mathcal{F}$ that maps a multi-index $(i_1, i_2, \ldots, i_s)$ to the partial differential operator $f_{i_1} \ldots f_{i_s}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$, considered as a linear transformation of $\mathcal{E}(M)$. In the analytic case one may go further, since as a basic principle, the relations between the iterated Lie brackets of the vector fields f_j completely determine the geometry and dynamical behavior [90,91]. Instead of considering actions on $\mathcal{E}(M)$, it suffices to consider the action of these partial differential operators, and of the dynamical system (32) on a set of polynomial functions. In a certain technical sense, an associative algebra (of polynomial functions) is dual to the Lie algebra (generated by the analytic vector fields f_j). However, much further simplifications and deeper insights are gained by appropriately accounting for the intrinsic noncommutative structure of the flows of the system. Using a different product structure, the next section will *lift* the system (32) to a *universal* system that has no nontrivial relations between the iterated Lie brackets of the formal vector fields and which is modeled on an algebra of noncommutative polynomials (and noncommutative power series) rather than the space $\mathcal{E}(M)$ of all smooth functions on M. An important feature of this approach is that it does not rely on polynomial functions with respect to an a-priori chosen set of local coordinates, but rather uses polynomials which map to intrinsically geometric objects.

Zinbiel Algebra and Combinatorics

Just as the chronological calculus of time-varying vector fields is intimately connected to chronological algebras, the calculus of affine systems gives rise to its own algebraic and combinatorial structure. This, and the subsequent section, give a brief look into this algebra structure and demonstrate how, together with the chronological calculus, it leads to effective solution formulas for important control problems and beyond.

It is traditional and convenient to slightly change the notation and nomenclature. For further details see

[78] or [57]. Rather than using small integers $1, 2, \dots, m$ to index the components of the control and the vector fields in (32), use an arbitrary index set Z whose elements will be called *letters* and denoted usually by $a, b, c, \dots$ For the purposes of this note the *alphabet* Z will be assumed to be finite, but much of the theory can be developed for infinite alphabets as well. Multi-indices, i. e. finite sequences with values in Z are called *words*, and they have a natural associative product structure denoted by juxtaposition. For example, if $w = a_1a_2 \dots a_r$ and $z = b_1b_2 \dots b_s$ then $wz = a_1a_2 \dots a_rb_1b_2 \dots b_s$. The empty word is denoted e or 1. The set of all words of length n is Z^n, $Z^* = \cup_{n=0}^{\infty} Z^n$ and $Z^+ = \cup_{n=1}^{\infty} Z^n$ are the sets of all words and all nonempty words. Endow the associative algebra $A(Z)$, generated by the alphabet Z with coefficients in the field $\mathbf{k} = \mathbb{R}$, with an inner product $\langle \cdot, \cdot \rangle$ so that Z^* is an orthonormal basis. Write $\hat{A}(Z)$ for the completion of $A(Z)$ with respect to the uniform structure in which a fundamental system of basic neighborhoods of a polynomial $p \in A(Z)$ is the family of subsets $\{q \in A(Z) \colon \forall w \in W,\ \langle w, p - q\rangle = 0\}$ indexed by all finite subsets $W \subset Z^*$.

The smallest linear subspace of the associative algebra $A(Z)$ that contains the set Z and is closed under the commutator product $(w, z) \mapsto [w, z] = wz - zw$ is isomorphic to the free Lie algebra $L(Z)$ over the set Z. On the other side, one may equip $A^+(Z)$ with a Zinbiel algebra structure by defining the product $*\colon A^+(Z) \times A^+(Z \mapsto A^+(Z)$ for $a \in Z$ and $w, z \in Z^+$ by

$$w * a = wa \quad \text{and} \quad w * (za) = (w * z + z * w)a \quad (41)$$

and extending bilinearly to $A^+(Z) \times A^+(Z)$. One easily verifies that this product satisfies the Zinbiel identity for all $r, s, t \in A^+(Z)$

$$r * (s * t) = (r * s) * t + (s * r) * t\,. \quad (42)$$

The symmetrization of this product is the better known associative *shuffle product* $w \sqcup\!\sqcup z = w * z + z * w$ which may be defined on all of $A(Z) \times A(Z)$. Algebraically, the shuffle product is characterized as the transpose of the coproduct $\Delta\colon A(Z) \mapsto A(Z) \otimes A(Z)$ which is the *concatenation* product homomorphism that on letters $a \in Z$ is defined as

$$\Delta\colon a \mapsto a \otimes 1 + 1 \otimes a\,. \quad (43)$$

This relation between this coproduct and the shuffle is that for all $u, v, w \in A(Z)$

$$\langle \Delta(w), u \otimes v\rangle = \langle w, u \sqcup\!\sqcup v\rangle\,. \quad (44)$$

After this brief side-trip into algebraic and combinatorial objects we return to the control systems. The shuffle product $\sqcup\!\sqcup$ has been utilized for a long time in this and related contexts. But the Zinbiel product structure was only recently recognized as encoding important information. Its most immediate role is seen in the fact that the aforementioned map Υ from multi-indices to iterated integral functionals, now easily extends to a map (using the same name) $\Upsilon\colon A(Z) \mapsto \mathcal{IIF}(\mathcal{U}^Z)$ is a Zinbiel algebra homomorphism when the latter is equipped with a *pointwise* product induced by the product $*$ on $\mathcal{U} = AC([0, T], [-1, 1])$ defined by

$$(U * V)(t) = \int_0^t U(s)V'(s)\mathrm{d}s\,. \quad (45)$$

It is straightforward to verify that this product indeed equips $\mathcal{U}$ with a Zinbiel structure, and hence also $\mathcal{IIF}(\mathcal{U}^Z)$. On the side, note that the map Υ maps the shuffle product of words (noncommuting polynomials) to the associative pointwise product of scalar functions. The connection of the Zinbiel product with differential equations and the Chen–Fliess series is immediate. First *lift* the system (32) from the manifold M to the algebra $A(Z)$, roughly corresponding to the linear space of polynomial functions in an algebra of iterated integral functionals. Formally consider curves $S\colon [0, T] \mapsto A(Z)$ that satisfy

$$\dot{S} = S \sum_{a \in Z} a u_a(t)\,. \quad (46)$$

Using the Zinbiel product (45), and writing $U_a(t) = \int_0^t u_a(\tau)\,\mathrm{d}\tau$ for the primitives of the controls, the integrated form of the *universal control system* (46)

$$S(t) = 1 + \int_0^t S(\tau)F'(\tau)\,\mathrm{d}\tau \quad \text{with} \quad F = \sum_{a \in Z} U_a\,, \quad (47)$$

is most compactly written as

$$S = 1 + S * F\,. \quad (48)$$

Iteration yields the explicit series expansion

$$\begin{aligned} S &= 1 + (1 + S * F) * F \\ &= 1 + F + ((1 + S * F) * F) * F \\ &= 1 + F + (F * F) + (((1 + S * F) * F) * F) * F \\ &= 1 + F + (F * F) + ((F * F) * F) \\ &\qquad + ((((1 + S * F) * F) * F) * F) * F \\ &\ \vdots \\ &= 1 + F + (F * F) + ((F * F) * F) \\ &\qquad + (((F * F) * F) * F) \dots \end{aligned}$$

Using intuitive notation for Zinbiel powers this solution formula in the form of an infinite series is compactly written as

$$S = \sum_{n=0}^{\infty} F^{*n} = 1 + F + F^{*2} + F^{*3} + F^{*4} + F^{*5} + F^{*6} + \cdots \tag{49}$$

The reader is encouraged to expand all of the terms and match each step to the usual computations involved in obtaining Volterra series expansions. Indeed this formula (49) is nothing more than the Chen–Fliess series (38) in very compact notation. For complete technical details and further discussion we refer the interested reader to the original articles [56,57] and the many references therein.

Application: Product Expansions and Normal Forms

While the Chen–Fliess series as above has been extremely useful to obtain precise estimates that led to most known necessary conditions and sufficient conditions for local controllability and optimality, compare e. g. [53,88,94], this series expansion has substantial shortcomings. For example, it is clear from Ree's theorem [77] that the series is an exponential Lie series. But this is not at all obvious from the series as presented (38). Most detrimental for practical purposes is that finite truncations of the series *never* correspond to any approximating systems. Much more convenient, in particular for path planning algorithms [47,48,63,64,75,95], but also in applications for numerical integration [17,46] are expansions as directed infinite products of exponentials or as an exponential of a Lie series.

In terms of the map $\mathcal{F}$ that substitutes for each letter $a \in Z$ the corresponding vector field f_a, one finds that the Chen–Fliess series (39) is simply the image of a natural object under the map $\Upsilon \otimes \mathcal{F}$. Indeed, under the usual identification of the space $\mathrm{Hom}(V, W)$ of linear maps between vector spaces V and W with the product $V^* \otimes W$, and noting that Z^* is an orthonormal basis for $A(Z)$, the identity map from $A(Z)$ to $A(Z)$ is identified with the series

$$\mathrm{Id}_{A(Z)} \equiv \sum_{w \in Z^*} w \otimes w \quad \in \; A(Z) \otimes \hat{A}(Z)\,. \tag{50}$$

Thus, any rewriting of the combinatorial object on the right hand side will, via the map $\Upsilon \otimes \mathcal{F}$, give an alternative presentation of the Chen–Fliess series. In particular, from elementary consideration, using also the Hopf-algebraic structures of $A(Z)$, it is a-priori clear that there exist expansions in the forms

$$\begin{aligned} \sum_{w \in Z^*} w \otimes w &= \exp\left(\sum_{h \in \mathcal{H}} \zeta_h \otimes [h] \right) \\ &= \overleftarrow{\prod_{h \in \mathcal{H}}} \exp\left(\xi_h \otimes [h]\right) \end{aligned} \tag{51}$$

where $\mathcal{H}$ indexes an ordered basis $\{[h] \colon H \in \mathcal{H}\}$ of the free Lie algebra $L(Z)$ and for each $h \in \mathcal{H}, \xi_h$ and ζ_h are polynomials in $A(Z)$ that are mapped by Υ to corresponding iterated integral functionals. The usefulness of such expression depends on the availability of simple formulas for $\mathcal{H}, \xi_h$ and ζ_h.

Bases for free Lie algebras are well-known since Hall [42], and have been unified by Viennot [101]. Specifically, a Hall set over a set Z is any strictly ordered subset $\tilde{\mathcal{H}} \subseteq \mathcal{T}(\mathbb{Z})$ from the set $\mathcal{M}(Z)$ labeled binary trees (with leaves labeled by Z) that satisfies

(i) $Z \subseteq \tilde{\mathcal{H}}$
(ii) Suppose $a \in Z$. Then $(t, a) \in \tilde{\mathcal{H}}$ iff $t' \in \tilde{\mathcal{H}}$, $t' < a$ and $a < (t', a)$.
(iii) Suppose $u, v, w, (u, v) \in \tilde{\mathcal{H}}$. Then $(t', (t''', t'''')) \in \tilde{\mathcal{H}}$ iff $t''' \leq t' \leq (t''', t'''')$ and $t' < (t', (t''', t''''))$.

There are natural mappings $\theta\colon \mathcal{T}(Z) \mapsto L(Z) \subseteq A(Z)$ and $\vartheta\colon \mathcal{T}(Z) \mapsto A(Z)$ (the *foliage map*) that map labeled binary trees to Lie polynomials and to words, and which are defined for $a \in Z$, $(t', t'') \in \mathcal{T}(Z)$ by $\theta(a) = \vartheta(a) = a$ and recursively $\theta((t', t'')) = [\theta(t'), \theta(t'')]$ and $\vartheta((t', t'')) = \vartheta(t')\vartheta(t'')$. The image of a Hall set under the map θ is a basis for the free Lie algebra $L(Z)$. Moreover, the restriction of the map ϑ to a Hall-set is one-to-one which leads to a fundamental unique factorization property of words into products of Hall-words, and a unique way of recovering Hall trees from the Hall words that are their images under ϑ [101]. This latter property is very convenient as it allows one to use Hall words rather than Hall trees for indexing various objects.

The construction of Hall sets, as well as the proof that they give rise to bases of free Lie algebras, rely in an essential way on the process of *Lazard elimination* [9] which is intimately related to the solution of differential Eqs. (32) or (46) by a successive variation of parameters [93]. As a result, using Hall sets it is possible to obtain an extremely simple and elegant recursive formula for the coefficients ξ_h in (51), whereas formulas for the ζ_h are less immediate, although they may be computed by straightforward formulas in terms of natural maps of the Hopf algebra structure on $A(Z)$ [33]. Up to a normalization factor μ in terms of multi-factorials, the formula for the ξ_h is extremely simple in terms of the Zinbiel product on $A(Z)$. For letters $a \in Z$ and Hall words $h, k, hk \in \mathcal{H}$ one has

$$\xi_a = a \quad \text{and} \quad \xi_{hk} = \mu_{hk} \cdot \xi_h * \xi_k \,. \tag{52}$$

Using completely different notation, this formula appeared originally in the work of Schützenberger [84], and was later rediscovered in various settings by Grayson and Grossman [38], Melancon and Reutenauer [69] and Sussmann [93]. In the context of control, it appeared first in [93].

To illustrate the simplicity of the recursive formula, which is based on the factorization of Hall words, consider the following *normal form* for a free nilpotent system (of rank $r = 5$) using a typical Hall set over the alphabet $Z = \{0, 1\}$. This is the closest nonlinear analogue to the Kalman controller normal form of a linear system, which is determined by Kronecker indices that classify the lengths of chains of integrators. For practical computations nilpotent systems are commonly used as approximations of general nonlinear systems – and every nilpotent system (that is, a system of form (32) whose vector fields f_a generate a nilpotent Lie algebra) is the image of a free nilpotent system as below under some projection map. For convenience, the example again uses the notation x_h instead of ξ_h.

$$\begin{aligned}
\dot{x}_0 &= u_0\\
\dot{x}_1 &= u_1\\
\dot{x}_{01} &= x_0 \cdot \dot{x}_1 = x_0\, u_1\\
\dot{x}_{001} &= x_0 \cdot \dot{x}_{01} = x_0^2\, u_1\\
&\quad \text{using } \vartheta^{-1}(001) = (0(01))\\
\dot{x}_{101} &= x_1 \cdot \dot{x}_{01} = x_1 x_0\, u_1\\
&\quad \text{using } \vartheta^{-1}(101) = (1(01))\\
\dot{x}_{0001} &= x_0 \cdot \dot{x}_{001} = x_0^3\, u_1\\
&\quad \text{using } \vartheta^{-1}(0001) = (0(0(01)))\\
\dot{x}_{1001} &= x_1 \cdot \dot{x}_{001} = x_1 x_0^2\, u_1\\
&\quad \text{using } \vartheta^{-1}(1001) = (1(0(01)))\\
\dot{x}_{1101} &= x_1 \cdot \dot{x}_{101} = x_1^2 x_0\, u_1\\
&\quad \text{using } \vartheta^{-1}(1101) = (1(1(01)))\\
\dot{x}_{00001} &= x_0 \cdot \dot{x}_{001} = x_0^4\, u_1\\
&\quad \text{using } \vartheta^{-1}(00001) = (0(0(0(01))))\\
\dot{x}_{10001} &= x_1 \cdot \dot{x}_{0001} = x_1 x_0^3\, u_1\\
&\quad \text{using } \vartheta^{-1}(10001) = (1(0(0(01))))\\
\dot{x}_{11001} &= x_1 \cdot \dot{x}_{1001} = x_1^2 x_0^2\, u_1\\
&\quad \text{using } \vartheta^{-1}(11001) = (1(1(0(01))))\\
\dot{x}_{01001} &= x_{01} \cdot \dot{x}_{001} = x_{01} x_0^3\, u_1\\
&\quad \text{using } \vartheta^{-1}(01001) = ((01)(0(01)))\\
\dot{x}_{01101} &= x_{01} \cdot \dot{x}_{101} = x_{01} x_1^2 x_0\, u_1\\
&\quad \text{using } \vartheta^{-1}(01101) = ((01)(1(01)))\,.
\end{aligned}$$

This example may be interpreted as a system of form (32) with the x_h being a set of coordinate functions on the manifold M, or as the truncation of the lifted system (46) with the x_h being a basis for a finite dimensional subspace of the linear space $\mathcal{E}(M)$. For more details and the technical background see [56,57].

Future Directions

The chronological calculus is still a young methodology. Its intrinsically infinite dimensional character demands a certain sophistication from its users, and thus it may take some time until it reaches its full potential. From a different point of view this also means that there are many opportunities to explore its advantages in ever new areas of applications. A relatively straightforward approach simply checks classical topics of systems and control for whether they are amenable to this approach, and whether this will be superior to classical techniques. The example of [70] which connects the chronological calculus with discrete time systems is a model example for such efforts. There are many further opportunities, some more speculative than others.

Control The chronological calculus appears ideally suited for the investigation of fully nonlinear systems – but in this area there is still much less known than in the case of nonlinear control systems that are affine in the control, or the case of linear systems. Specific conditions for the existence of solutions, controllability, stabilizability, and normal forms are just a few subjects that, while studies have been initiated [1,2,4], deserve further attention.

Computation In recent years much effort has been devoted to understanding the foundations of numerical integration algorithms in nonlinear settings, and to eventually design more efficient algorithms and eventually prove their superiority. The underlying mathematical structures, such as the compositions of noncommuting flows are very similar to those studies in control, as observed in e. g. [23], one of the earlier publications in this area. Some more recent work in this direction is [17,46,72,73,74]. This is also very closely related to the studies of the underlying combinatorial and algebraic structures – arguably [27,28,29,76] are some of the ones most closely related to the subject of this article. There appears to be much potential for a two-way exchange of ideas and results.

Geometry Arguably one of the main thrusts will continue to be the use of the chronological calculus for studying the differential geometric underpinnings of systems and control theory. But it is conceivable that such studies may eventually abstract further – just like after the 1930s – to studies of generalizations of algebras that stem from systems on classical manifolds, much in the spirit of modern noncommutative geometry with [19] being the best-known proponent.

Algebra and combinatorics In general very little is known about possible ideal and subalgebra structures of chronological and Zinbiel algebras. This work was initiated in [2], but relatively little progress has been made since. (Non)existence of finite dimensional (nilpotent) Zinbiel algebras has been established, but only over a complex field [25,26]. Bases of free chronological algebras are discussed in [2], but many other aspects of this algebraic structure remain unexplored. This also intersects with the aforementioned efforts in foundations of computation – on one side there are Hopf-algebraic approaches to free Lie algebras [78] and nonlinear control [33,40,41], while the most recent breakthroughs involve dendriform algebras and related subjects [27,28,29]. This is an open field for uncovering the connections between combinatorial and algebraic objects on one side and geometric and systems objects on the other side. The chronological calculus may well serve as a vehicle to elucidate the correspondences.

Acknowledgments

This work was partially supported by the National Science Foundation through the grant DMS 05-09030

Bibliography

1. Agrachëv A, Gamkrelidze R (1978) Exponential representation of flows and chronological calculus. Math Sbornik USSR (Russian) 107(N4):487–532. Math USSR Sbornik (English translation) 35:727–786
2. Agrachëv A, Gamkrelidze R (1979) Chronological algebras and nonstationary vector fields. J Soviet Math 17:1650–1675
3. Agrachëv A, Gamkrelidze R, Sarychev A (1989) Local invariants of smooth control systems. Acta Appl Math 14:191–237
4. Agrachëv A, Sachkov YU (1993) Local controllability and semigroups of diffeomorphisms. Acta Appl Math 32:1–57
5. Agrachëv A, Sachkov YU (2004) Control Theory from a Geometric Viewpoint. Springer, Berlin
6. Agrachëv A, Sarychev A (2005) Navier Stokes equations: controllability by means of low modes forcing. J Math Fluid Mech 7:108–152
7. Agrachëv A, Vakhrameev S (1983) Chronological series and the Cauchy–Kowalevski theorem. J Math Sci 21:231–250
8. Boltyanski V, Gamkrelidze R, Pontryagin L (1956) On the theory of optimal processes (in Russian). Doklady Akad Nauk SSSR, vol.10, pp 7–10
9. Bourbaki N (1989) Lie Groups and Lie algebras. Springer, Berlin
10. Brockett R (1971) Differential geometric methods in system theory. In: Proc. 11th IEEE Conf. Dec. Cntrl., Berlin, pp 176–180
11. Brockett R (1976) Volterra series and geometric control theory. Autom 12:167–176
12. Bullo F (2001) Series expansions for the evolution of mechanical control systems. SIAM J Control Optim 40:166–190
13. Bullo F (2002) Averaging and vibrational control of mechanical systems. SIAM J Control Optim 41:542–562
14. Bullo F, Lewis A (2005) Geometric control of mechanical systems: Modeling, analysis, and design for simple mechanical control systems. Texts Appl Math 49 IEEE
15. Caiado MI, Sarychev AV () On stability and stabilization of elastic systems by time-variant feedback. ArXiv:math.AP/0507123
16. Campbell J (1897) Proc London Math Soc 28:381–390
17. Casas F, Iserles A (2006) Explicit Magnus expansions for nonlinear equations. J Phys A: Math General 39:5445–5461
18. Chen KT (1957) Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. Ann Math 65:163–178
19. Connes A (1994) Noncommutative geometry. Academic Press, San Diego
20. Cortés J, Martinez S (2003) Motion control algorithms for simple mechanical systems with symmetry. Acta Appl Math 76:221–264
21. Cortés J, Martinez S, Bullo F (2002) On nonlinear controllability and series expansions for Lagrangian systems with dissipative forces. Trans IEEE Aut Control 47:1401–1405
22. Crouch P (1981) Dynamical realizations of finite Volterra series. SIAM J Control Optim 19:177–202
23. Crouch P, Grossman R (1993) Numerical integration of ordinary differential equations on manifolds. J Nonlinear Sci 3:1–33
24. Crouch P, Lamnabhi-Lagarrigue F (1989) Algebraic and multiple integral identities. Acta Appl Math 15:235–274
25. Dzhumadil'daev A (2007) Zinbiel algebras over a q-commutator. J Math Sci 144:3909–3925
26. Dzhumadil'daev A, Tulenbaev K (2005) Nilpotency of Zinbiel algebras. J Dyn Control Syst 11:195–213
27. Ebrahimi-Fard K, Guo L (2007) Rota–Baxter algebras and dendriform algebras. J Pure Appl Algebra 212:320–339
28. Ebrahimi-Fard K, Manchon D, Patras F (2007) A Magnus- and Fer-type formula in dendriform algebras. J Found Comput Math (to appear) http://springerlink.com/content/106038/
29. Ebrahimi-Fard K, Manchon D, Patras F (2008) New identities in dendriform algebras. J Algebr 320:708–727
30. Fliess M (1978) Développements fonctionelles en indéterminées non commutatives des solutions d'équations différentielles non linéaires forcées. CR Acad Sci France Ser A 287:1133–1135
31. Fliess M (1981) Fonctionelles causales nonlinéaires et indeterminées noncommutatives. Bull Soc Math France 109:3–40
32. Gamkrelidze RV, Agrachëv AA, Vakhrameev SA (1991) Ordinary differential equations on vector bundles and chronological calculus. J Sov Math 55:1777–1848

33. Gehrig E (2007) Hopf algebras, projections, and coordinates of the first kind in control theory. Ph D Dissertation, Arizona State University
34. Gelfand I (1938) Abstract functions and linear operators. Math Sbornik NS 4:235–284
35. Gelfand I, Raikov D, Shilov G (1964) Commutative normed rings. (Chelsea) New York (translated from the Russian, with a supplementary chapter), Chelsea Publishing, New York
36. Ginzburg V, Kapranov M (1994) Koszul duality for operads. Duke Math J 76:203–272
37. Gray W, Wang Y (2006) Noncausal fliess operators and their shuffle algebra. In: Proc MTNS 2006 (Mathematical Theory of Networks and Systems). MTNS, Kyoto, pp 2805–2813
38. Grayson M, Grossman R (1990) Models for free nilpotent algebras. J Algebra 135:177–191
39. Grayson M, Larson R (1991) The realization of input-output maps using bialgebras. Forum Math 4:109–121
40. Grossman R, Larson R (1989) Hopf-algebraic structure of combinatorial objects and differential operators. Israel J Math 72:109–117
41. Grossman R, Larson R (1989) Hopf-algebraic structure of families of trees. J Algebra 126:184–210
42. Hall M (1950) A basis for free Lie rings and higher commutators in free groups. Proc Amer Math Soc 1:575–581
43. Haynes G, Hermes H (1970) Nonlinear controllability via Lie theory. SIAM J Control 8:450–460
44. Herman R (1963) On the accessibility problem in control theory. In: Int. Symp. Nonlinear Diff. Eqns. Nonlinear Mechanics. Academic Press, New York, pp 325–332
45. Hermann R, Krener A (1977) Nonlinear controllability and observability. IEEE Trans Aut Control 22:728–740
46. Iserles A, Munthe-Kaas H, Nrsett S, Zanna A (2000) Lie-group methods. Acta numerica 9:215–365
47. Jacob G (1991) Lyndon discretization and exact motion planning. In: Proc. Europ. Control Conf., pp 1507–1512, ECC, Grenoble
48. Jacob G (1992) Motion planning by piecewise constant or polynomial inputs. In: Proc. IFAC NOLCOS. Int Fed Aut, Pergamon Press, Oxford
49. Jakubczyk B (1986) Local realizations of nonlinear causal operators. SIAM J Control Opt 24:231–242
50. Jurdjevic V, Sussmann H (1972) Controllability of nonlinear systems. J Diff Eqns 12:95–116
51. Kalman R (1960) A new approach to linear filtering and prediction problems. Trans ASME – J Basic Eng 82:35–45
52. Kawski M (1988) Control variations with an increasing number of switchings. Bull Amer Math Soc 18:149–152
53. Kawski M (1990) High-order small-time local controllability. In: Sussmann H (ed) Nonlinear Controllability and Optimal Control. Dekker, pp 441–477, New York
54. Kawski M (2000) Calculating the logarithm of the Chen Fliess series. In: Proc. MTNS 2000, CDROM. Perpignan, France
55. Kawski M (2000) Chronological algebras: combinatorics and control. Itogi Nauki i Techniki 68:144–178 (translation in J Math Sci)
56. Kawski M (2002) The combinatorics of nonlinear controllability and noncommuting flows. In: Abdus Salam ICTP Lect Notes 8. pp 223–312, Trieste
57. Kawski M, Sussmann HJ (1997) Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory. In: Helmke U, Prätzel–Wolters D, Zerz E (eds) Operators, Systems, and Linear Algebra. Teubner, pp 111–128 , Stuttgart
58. Kirov N, Krastanov M (2004) Higher order approximations of affinely controlled nonlinear systems. Lect Notes Comp Sci 2907:230–237
59. Kirov N, Krastanov M (2005) Volterra series and numerical approximation of ODEs. Lect Notes Comp Sci 2907:337–344. In: Li Z, Vulkov L, Was'niewski J (eds) Numerical Analysis and Its Applications. Springer, pp 337–344, Berlin
60. Komleva T, Plotnikov A (2000) On the completion of pursuit for a nonautonomous two-person game. Russ Neliniini Kolyvannya 3:469–473
61. Krener A, Lesiak C (1978) The existence and uniqueness of Volterra series for nonlinear systems. IEEE Trans Aut Control 23:1090–1095
62. Kriegl A, Michor P (1997) The convenient setting of global analysis. Math Surv Monogr 53. Amer Math Soc, Providence
63. Lafferiere G, Sussmann H (1991) Motion planning for controllable systems without drift. In: IEEE Conf. Robotics and Automation. pp 1148–1153, IEEE Publications, New York
64. Lafferiere G, Sussmann H (1993) A differential geometric approach to motion planning. In: Li Z, Canny J (eds) Nonholonomic Motion Planning. Kluwer, Boston, pp 235–270
65. Lobry C (1970) Controllabilit'e des systèmes non linéares. SIAM J Control 8:573–605
66. Loday JL (1993) Une version non commutative des algèbres de Lie: les algèbres de Leibniz. Enseign Math 39:269–293
67. Loday JL, Pirashvili T (1996) Leibniz representations of Lie algebras. J Algebra 181:414–425
68. Martinez S, Cortes J, Bullo F (2003) Analysis and design of oscillatory control systems. IEEE Trans Aut Control 48:1164–1177
69. Melançon G, Reutenauer C (1989) Lyndon words, free algebras and shuffles. Canadian J Math XLI:577–591
70. Monaco S, Normand-Cyrot D, Califano C (2007) From chronological calculus to exponential representations of continuous and discrete-time dynamics: a lie-algebraic approach. IEEE Trans Aut Control 52:2227–2241
71. Morgansen K, Vela P, Burdick J (2002) Trajectory stabilization for a planar carangiform robot fish. In: Proc. IEEE Conf. Robotics and Automation. pp 756–762, New York
72. Munthe-Kaas H, Owren B (1999) Computations in a free Lie algebra. Royal Soc Lond Philos Trans Ser A 357:957–981
73. Munthe-Kaas H, Wright W (2007) On the Hopf algebraic structure of lie group integrators. J Found Comput Math 8(2):227–257
74. Munthe-Kaas H, Zanna A (1997) Iterated commutators, lie's reduction method and ordinary differential equations on matrix lie groups. In: Cucker F (ed) Found. Computational Math. Springer, Berlin, pp 434–441
75. Murray R, Sastry S (1993) Nonholonomic path planning: steering with sinusoids. IEEE T Autom Control 38:700–716
76. Murua A (2006) The Hopf algebra of rooted trees, free Lie algebras, and Lie series. J Found Comput Math 6:387–426
77. Ree R (1958) Lie elements and an algebra associated with shuffles. Annals Math 68:210–220
78. Reutenauer C (1991) Free Lie Algebras. Oxford University Press, New York
79. Rocha E (2003) On computation of the logarithm of the Chen–Fliess series for nonlinear systems. In: Zinober I, Owens D (eds)

Nonlinear and adaptive Control, Lect Notes Control Inf Sci 281:317–326, Sprtinger, Berlin

80. Rocha E (2004) An algebraic approach to nonlinear control theory. Ph D Dissertation, University of Aveiro, Portugal
81. Sanders J, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Appl Math Sci 59. Springer, New York
82. Sarychev A (2001) Lie- and chronologico-algebraic tools for studying stability of time-varying systems. Syst Control Lett 43:59–76
83. Sarychev A (2001) Stability criteria for time-periodic systems via high-order averaging techniques. In: Lect. Notes Control Inform. Sci. 259. Springer, London, pp 365–377
84. Schützenberger M (1958) Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un probléme de mathématiques appliquées. In: Dubreil S (ed) Algèbres et Théorie des Nombres. Faculté des Sciences de Paris vol 12 no 1 (1958-1959), Exposé no 1 pp 1–23
85. Serres U (2006) On the curvature of two-dimensional optimal control systems and zermelos navigation problem. J Math Sci 135:3224–3243
86. Sigalotti M (2005) Local regularity of optimal trajectories for control problems with general boundary conditions. J Dyn Control Syst 11:91–123
87. Sontag E, Wang Y (1992) Generating series and nonlinear systems: analytic aspects, local realizability and i/o representations. Forum Math 4:299–322
88. Stefani G (1985) Polynomial approximations to control systems and local controllability. In: Proc. 25th IEEE Conf. Dec. Cntrl., pp 33–38, New York
89. Stone M (1932) Linear Transformations in Hilbert Space. Amer Math Soc New York
90. Sussmann H (1974) An extension of a theorem of Nagano on transitive Lie algebras. Proc Amer Math Soc 45:349–356
91. Sussmann H (1983) Lie brackets and local controllability: a sufficient condition for scalar-input systems. SIAM J Cntrl Opt 21:686–713
92. Sussmann H (1983) Lie brackets, real analyticity, and geometric control. In: Brockett RW, Millman RS, Sussmann HJ (eds) Differential Geometric Control. pp 1–116, Birkhauser
93. Sussmann H (1986) A product expansion of the Chen series. In: Byrnes C, Lindquist A (eds) Theory and Applications of Nonlinear Control Systems. Elsevier, North-Holland, pp 323–335
94. Sussmann H (1987) A general theorem on local controllability. SIAM J Control Opt 25:158–194
95. Sussmann H (1992) New differential geometric methods in nonholonomic path finding. In: Isidori A, Tarn T (eds) Progr Systems Control Theory 12. Birkhäuser, Boston, pp 365–384
96. Tretyak A (1997) Sufficient conditions for local controllability and high-order necessary conditions for optimality. A differential-geometric approach. J Math Sci 85:1899–2001
97. Tretyak A (1998) Chronological calculus, high-order necessary conditions for optimality, and perturbation methods. J Dyn Control Syst 4:77–126
98. Tretyak A (1998) Higher-order local approximations of smooth control systems and pointwise higher-order optimality conditions. J Math Sci 90:2150–2191
99. Vakhrameev A (1997) A bang-bang theorem with a finite number of switchings for nonlinear smooth control systems. Dynamic systems 4. J Math Sci 85:2002–2016
100. Vela P, Burdick J (2003) Control of biomimetic locomotion via averaging theory. In: Proc. IEEE Conf. Robotics and Automation. pp 1482–1489, IEEE Publications, New York
101. Viennot G (1978) Algèbres de Lie Libres et Monoïdes Libres. Lecture Notes in Mathematics, vol 692. Springer, Berlin
102. Visik M, Kolmogorov A, Fomin S, Shilov G (1964) Israil Moiseevich Gelfand, On his fiftieth birthday. Russ Math Surv 19:163–180
103. Volterra V (1887) Sopra le funzioni che dipendono de altre funzioni. In: Rend. R Academia dei Lincei. pp 97–105, 141–146, 153–158
104. von Neumann J (1932) Mathematische Grundlagen der Quantenmechanik. Grundlehren Math. Wissenschaften 38. Springer, Berlin
105. Zelenko I (2006) On variational approach to differential invariants of rank two distributions. Diff Geom Appl 24:235–259

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies

MICHAEL BATTY
Center for Advanced Spatial Analysis,
University College London, London, UK

Article Outline

Glossary

Agent-based models Systems composed of individuals who act purposely in making locational/spatial decisions.

Bifurcation A process whereby divergent paths are generated in a trajectory of change in an urban system.

City size distribution A set of cities ordered by size, usually population, often in rank order.

Emergent patterns Land uses or economic activities which follow some spatial order.

Entropy maximizing The process of generating a spatial model by maximizing a measure of system complexity subject to constraints.

Equilibrium A state of the urban system which is balanced and unchanging.

Exponential growth The process whereby an activity changes through positive feedback on itself.

Fast dynamics A process of frequent movement between locations, often daily.

Feedback The process whereby a system variable influences another variable, either positively or negatively.

Fractal structure A pattern or arrangement of system elements that are self-similar at different spatial scales.

Land use transport model A model linking urban activities to transport interactions.

Life cycle effects Changes in spatial location which are motivated by aging of urban activities and populations.

Local neighborhood The space immediately around a zone or cell.

Logistic growth Exponential growth capacitated so that some density limit is not exceeded.

Lognormal distribution A distribution which has fat and long tails which is normal when examined on a logarithmic scale.

Microsimulation The process of generating synthetic populations from data which is collated from several sources.

Model validation The process of calibrating and testing a model against data so that its goodness of fit is optimized.

Multipliers Relationships which embody *n*th order effects of one variable on another.

Network scaling The in-degrees and out-degrees of a graph whose nodal link volumes follow a power law.

Population density profile A distribution of populations which typically follows an exponential profile when arrayed against distance from some nodal point.

Power laws Scaling laws that order a set of objects according to their size raised to some power.

Rank size rule A power law that rank orders a set of objects.

Reaction-diffusion The process of generating changes as a consequence of a reaction to an existing state and interactions between states.

Scale-free networks Networks whose nodal volumes follow a power law.

Segregation model A model which generates extreme global segregation from weak assumptions about local segregation.

Simulation The process of generating locational distributions according to a series of sub-model equations or rules.

Slow dynamics Changes in the urban system that take place over years or decades.

Social physics The application of classical physical principles involving distance, force and mass to social situations, particularly to cities and their transport.

Spatial interaction The movement of activities between different locations ranging from traffic distributions to migration patterns.

Trip distribution The pattern of movement relating to trips made by the population, usually from home to work but also to other activities such as shopping.

Urban hierarchy A set of entities physically or spatially scaled in terms of their size and areal extent.

Urban morphology Patterns of urban structure based on the way activities are ordered with respect to their locations.

Urban system A city represented as a set of interacting subsystems or their elements.

Definition of the Subject

Cities have been treated as systems for fifty years but only in the last two decades has the focus changed from aggregate equilibrium systems to more evolving systems whose structure emerges from the bottom up. We first outline the rudiments of the traditional approach focusing on equilibrium and then discuss how the paradigm has changed to one which treats cities as emergent phenomena generated through a combination of hierarchical levels of decision, driven in decentralized fashion. This is consistent with the complexity sciences which dominate the simulation of urban form and function. We begin however with a review of equilibrium models, particularly those based on spatial interaction, and we then explore how simple dynamic frameworks can be fashioned to generate more realistic models. In exploring dynamics, nonlinear systems which admit chaos and bifurcation have relevance but recently more pragmatic schemes of structuring urban models based on cellular automata and agent-based modeling principles have come to the fore. Most urban models deal with the city in terms of the location of its economic and demographic activities but there is also a move to link such models to urban morphologies which are clearly fractal in structure. Throughout this chapter, we show how key concepts in complexity such as scaling, self-similarity and far-from-equilibrium structures dominate our current treatment of cities, how we might simulate their functioning and how we might predict their futures. We conclude with the key problems that dominate the field and suggest how these might be tackled in future research.

Cities were first conceived as complex systems in the 1960s when architects and urban planners began to change their perceptions that cities be treated as ‘works of art’ to something much more akin to a functioning economic system that required social engineering. Since the time of Newton, social scientists had speculated that social systems could be described using concepts from classical physics and the notion that interrelationships between the component parts of such systems might be articulated using concepts of mass, force and energy established a rudimentary framework that came to be known as social physics. Together with various macro-economic theories of how economies function based on Keynesian ideas of input, output and economic multipliers, cities were assumed to be equilibrium systems in which their interactions such as traffic, trade flows, and demographic migration could be modeled in analogy to gravitation. These flows were seen as complementary to the location of employment and population which reflected costs of travel and land rent, in turn the product of micro-economic theories of how agents resolved their demand for and the supply of space through the land market. Operational models for policy analysis were initially built on this basis and used for testing the impact of different plans for making cities more efficient, locationally and in terms of their movement/traffic patterns.

This early work did not emphasize the dynamics of urban change or the morphology of cities. Thus theories and models were limited in their ability to predict patterns of urban growth as reflected in sprawl and the regeneration of urban areas. The first wave of models which treated the city system in aggregate, static and top down fashion, fell into disrepute due to these limitations. To address them, the focus moved in the 1980s to more theoretical considerations in which static social physics types of model were embedded in nonlinear dynamic frameworks built around ideas in chaos, and bifurcation theory. A parallel development in simulating urban morphology was built around treating cities as fractals which evolved from the bottom up and operational models in which such morphologies were governed using cellular automata were developed. Developments in moving from aggregate to disaggregate or bottom-up individual-based models were spawned from these developments with agent-based modeling providing a new focus to the field. There is now a momentum for treating urban aggregates through new ideas about growth and form through scaling while new forms of representing cities through networks linking these to fractal morphologies are being developed through network science at different scales. These will ultimately lead to operational urban and transport models built from the bottom up which simulate spatial processes operating on networks and other morphologies. These have the potential to address key problems of urban growth and evolution, congestion, and inequality.

Introduction

Cities were first treated formally as systems when General System Theory and Cybernetics came to be applied to the softer social sciences in the 1950s. Ludwig von Bertalanffy [67] in biology and Norbert Wiener [73] in engineering gave enormous impetus to this emerging interdisciplinary field that thrust upon us the idea that phenomena of interest in many disciplines could be articulated in generic terms as 'systems'. Moreover the prospect that the systems approach could yield generic policy, control and management procedures applicable to many different areas, appeared enticing. The idea of a general systems theory was gradually fashioned from reflections on the way distinct entities which were clearly collections of lower order elements, organized into a coherent whole, displaying pattern and order which in the jargon of the mid-twentieth century was encapsulated in the phrase that "the whole is greater than the sum of the parts". The movement began in biology in the 1920s, gradually eclipsing parts of engineering in the 1950s and spreading to the management and social sciences, particularly sociology and political science in the 1960s. It was part of a wave of change in the social sciences which began in the late 19th century as these fields began to emulate the physical sciences, espousing positivist methods which had appeared so successful in building applicable and robust theory.

The focus then was on ways in which the elements comprising the system interacted with one another through structures that embodied feedbacks keeping the system sustainable within bounded limits. The notion that such systems have controllers to 'steer' them to meet certain goals or targets is central to this early paradigm and the science of "…control and communication in the animal and the machine" was the definition taken up by Norbert Wiener [73] in his exposition of the science of cybernetics. General system theory provided the generic logic for both the structure and behavior of such systems through various forms of feedback and hierarchical organization while cybernetics represented the 'science of steersmanship' which would enable such systems to move towards explicit goals or targets. Cities fit this characterization admirably and in the 1950s and 1960s, the traditional approach that articulated cities as structures that required physical and aesthetic organization, quickly gave way to deeper notions that cities needed to be understood as general systems. Their control and planning thus required much more subtle interventions than anything that had occurred hitherto in the name of urban planning.

Developments in several disciplines supported these early developments. Spatial analysis, as it is now called, began to develop within quantitative geography, linked to the emerging field of regional science which represented a synthesis of urban and regional economics in which location theory was central. In this sense, the economic structure of cities and regions was consistent with classical macro and micro economics and the various techniques and models that were developed within these domains had immediate applicability. Applications of physical analogies to social and city systems, particularly ideas about gravitation and potential, had been explored since the mid 19th century under the banner of 'social physics' and as transportation planning formally began in the 1950s, these ideas were quickly adopted as a basis for transport modeling. Softer approaches in sociology and political science also provided support for the idea of cities as organizational systems while the notion of cybernetics as the basis for management, policy and control of cities was adopted as an important analogy in their planning [26,53].

The key ideas defined cities as sets of elements or components tied together through sets of interactions. The archetypal structure was fashioned around land use activities with economic and functional linkages between them represented initially in terms of physical movement, traffic. The key idea of feedback, which is the dynamic that holds a general system together, was largely represented in terms of the volume and pattern of these interactions, at a single point in time. Longer term evolution of urban structure was not central to these early conceptions for the focus was largely on how cities functioned as equilibrium structures. The prime imperative was improving how interactions between component land uses might be made more efficient while also meeting goals involving social and spatial equity. Transportation and housing were of central importance in adopting the argument that cities should be treated as examples of general systems and steered according to the principles of cybernetics.

Typical examples of such systemic principles in action involve transportation in large cities and these early ideas about systems theory hold as much sway in helping make sense of current patterns as they did when they were first mooted fifty or more years ago. Different types of land use with different economic foci interact spatially with respect to how employees are linked to their housing locations, how goods are shipped between different locations to service the production and consumption that define these activities, how consumers purchase these economic activities which are channeled through retail and commercial centers, how information flows tie all these economies together, and so on: the list of linkages is endless. These activities are capacitated by upper limits on density and capacity. In Greater London for example, the traffic has reached saturation limits in the central city and with few new roads being constructed over the last 40 years, the focus has shifted to improving public transport and to road pricing.

The essence of using a systems model of spatial interaction to test the impact of such changes on city structure is twofold: first such a model can show how people might shift mode of transport from road to rail and bus, even to walking and cycling, if differential pricing is applied to the road system. The congestion charge in central London imposed in 2003 led to a 30 percent reduction in the use of vehicles and this charge is set to increase massively for certain categories of polluting vehicles in the near future. Second the slightly longer term effects of reducing traffic are to increase densities of living, thus decreasing the length and cost of local work journeys, also enabling land use to respond by changing their locations to lower cost areas. All these effects ripple through the system with the city system models presented here designed to track and predict such *n*th order effects which are rarely obvious. Our focus in this chapter is to sketch the state-of-the-art in these complex systems models showing how new developments in the methods of the complexity sciences are building on a basis that was established half century ago.

Since early applications of general systems theory, the paradigm has changed fundamentally from a world where systems were viewed as being centrally organized, from the top down, and notions about hierarchy were predominant, to one where we now consider systems to be structured from the bottom up. The idea that one or the other – the centralized or the decentralized view – are mutually exclusive of each other is not entirely tenable of course but the balance has certainly changed. Theories have moved from structures and behaviors being organized according to some central control to theories about how systems retain their own integrity from the bottom up, endorsing what Adam Smith over 300 years ago, called "the hidden hand". This shift has brought onto the agenda the notion of equilibrium and dynamics which is now much more central to systems theory than it ever was hitherto. Systems such as cities are no longer considered to be equilibrium structures, notwithstanding that many systems models built around equilibrium are still eminently useful. The notion that city systems are more likely to be in disequilibrium, all the time, or even classed as far-from-equilibrium continually reinforcing the move away from equilibrium, is comparatively new but consistent with the speed of change and volatility in cities observed during the last fifty years.

The notion to that change is nowhere smooth but discontinuous, often chaotic, has become significant. Equilibrium structures are renewed from within as unanticipated innovations, many technological but some social, change the way people make decisions about how they locate and move within cities. Historical change is important in that historical accidents often force the system onto a less than optimal path with such path dependence being crucial to an understanding of any current equilibria and the dynamic that is evolving. Part of this newly emerging paradigm is the idea that new structures and behaviors that emerge are often unanticipated and surprising. As we will show in this chapter, when we look at urban morphologies, they are messy but ordered, self-similar across many scales, but growing organically from the bottom up. Planned cities are always the exception rather than the rule and when directly planned, they only remain so for very short periods of time.

The new complexity sciences are rewriting the theory of general systems but they are still founded on the rudiments of structures composed of elements, now often called actors or agents, linked through interactions which determine the processes of behavior which keep the system in equilibrium and/or move it to new states. Feedback is still central but recently has been more strongly focused on how system elements react to one another through time. The notion of an unchanging equilibrium supported by such feedbacks is no longer central; feedback is now largely seen as the way in which these structures are evolved to new states. In short, system theory has shifted to consider such feedbacks in positive rather than negative terms although both are essential. Relationships between the system elements in terms of their interactions are being enriched using new ideas from networks and their dynamics [56]. Key notions of how the elements of systems scale relative to one another and relative to their system hierarchies have become useful in showing how local actions and interactions lead to global patterns which can only be predicted from the bottom up [54]. This new view is about how emergent patterns can be generated using models that grow the city from the bottom up [37], and we will discuss all these ideas in the catalog of models that we present below.

We begin by looking at models of cities in equilibrium where we illustrate how interactions between key system elements located in space follow certain scaling laws reflecting agglomeration economies and spatial competition. The network paradigm is closely linked to these ideas in structural terms. None of these models, still important for operational simulation modeling in a policy context, have an internal dynamic and thus we turn to examine dynamics in the next section. We then start with simple exponential growth, showing how it can be capacitated as logistic growth from which nonlinear behaviors can result as chaos and bifurcation. We show how these models might be linked to a faster dynamics built around equilibrium spatial interaction models but to progress these developments, we present much more disaggregate models based on agent simulation and cellular automata principles. These dynamics are then generalized as reaction-diffusion models.

Our third section deals with how we assemble more integrated models built from these various equilibrium and dynamic components or sub-models. We look at large-scale land use transport models which are equilibrium in focus. We then move to cellular automata models of land development, concluding our discussion with reference to the current development of fine scale agent-based models where each individual and trip maker in the city system is simulated. We sprinkle our presentation with various empirical applications, many based on data for Greater London showing how employment and population densities scale, how movement patterns are consistent with the underling infrastructure networks that support them, and how the city has grown through time. We show how the city can be modeled in terms of its structure and the way changes to it can be visualized. We then link these more abstract notions about how cities are structured in spatial-locational terms to their physical or fractal morphology which is a direct expression of their scaling and complexity. We conclude with future directions, focusing on how such models can be validated and used in practical policy-making.

Cities in Equilibrium

Arrangements of Urban Activities

Cities can usually be represented as a series of n locations, each identified by i, and ordered from $i = 1, 2, \ldots, n$. These locations might be points or areas where urban activity takes place, pertaining either to the inter-urban scale where locations are places not necessarily adjacent to one another or at the intra-urban scale where a city is exhaustively partitioned into a set of areas. We will use both representations here but begin with a generic formulation which does not depend on these differences per se.

It is useful to consider the distribution of locations as places where differing amounts of urban activity can take place, using a framework which shows how different arrangements of activity can be consistently derived. Different arrangements of course imply different physical forms of city. Assume there is N amount of activity to be distributed in n locations as $N_1, N_2, \ldots$ Beginning with N_1, there are $N!/[N_1!(N - N_1)!$ allocations of

N_1, $(N-N_1)!/[N_2!(N-N_1-N_2)!$ allocations of N_2, $(N-N_1-N_2)!/[N_3!(N-N_1-N_2-N_3)!$ of N_3 and so on. To find the total number of arrangements W, we multiply each of these quantities together where the product is

$$W = \frac{N!}{\prod_i N_i!} \,. \tag{1}$$

This might be considered a measure of complexity of the system in that it clearly varies systematically for different allocations. If all N activity were to be allocated to the first location, then $W = 1$ while if an equal amount of activity were to be allocated to each location, then W would vary according to the size of N and the number of locations n. It can be argued that the most likely arrangement of activities would be the one which would give the greatest possibility of distinct individual activities being allocated to locations and such an arrangement could be found by maximizing W (or the logarithm of W which leads to the same). Such maximizations however might be subject to different constraints on the arrangements which imply different conservation laws that the system must meet. This would enable different types of urban form to be examined under different conditions related to density, compactness, sprawl and so on, all of which might be formalized in this way.

To show how this is possible, consider the case where we now maximize the logarithm of W subject to meaningful constraints. The logarithm of Eq. (1) is

$$\ln W = \ln(N!) - \sum_i \ln(N_i!) \,, \tag{2}$$

which using Stirling's formula, simplifies to

$$\ln W \approx N + \ln(N!) - \sum_i N_i \ln N_i \,. \tag{3}$$

N_i which is the number of units of urban activity allocated to location i, is a frequency that can be normalized into a probability as $p_i = N_i/N$. Substituting for $N_i = Np_i$ in Eq. (3) and dropping the constant terms leads to

$$\ln W \propto -\sum_i p_i \ln p_i = H \,, \tag{4}$$

where it is now clear that the formula for the number of arrangements is proportional to Shannon's entropy H. Thus the process of maximizing $\ln W$ is the well-known process of maximizing entropy subject to relevant constraints and this leads to many standard probability distributions [66]. Analogies between city and other social systems with statistical thermodynamics and information theory were developed in the 1960s and represented one of the first formal approaches to the derivation of models for simulating the interaction between locations and the amount of activity attracted to different locations in city, regional and transport systems. As such, it has become a basis on which to build many different varieties of urban model [74].

Although information or entropy has been long regarded as a measure of system complexity, we will not take this any further here except to show how it is useful in deriving different probability distributions of urban activity. Readers are however referred to the mainstream literature for both philosophic and technical expositions of the relationship between entropy and complexity (for example see [42]). The measure H in Eq. (4) is at a maximum when the activity is distributed evenly across locations, that is when $p_i = 1/n$ and $H = \ln n$ while it is at a minimum when $p_i = 1$ and $p_j = 0, j = 1, 2, \ldots, n, i \neq j$, and $H = 0$. It is clear too that H varies with n; that is as the number of locations increases, the complexity or entropy of the system also increases. However what is of more import here is the kind of distribution that maximizing entropy generates when H is maximized subject to appropriate constraints. We demonstrate this as follows for a simple but relevant case where the key constraint is to ensure that the system reproduces the mean value of an attribute of interest. Let p_i be the probability of finding a place i which has P_i population residing there. Then we maximize the entropy

$$H = -\sum_i p_i \ln p_i \,, \tag{5}$$

subject to a normalization constraint on the probabilities

$$\sum_i p_i = 1 \,, \tag{6}$$

and a constraint on the mean population of places $\bar{P}$ in the system, that is

$$\sum_i p_i P_i = \bar{P} \,. \tag{7}$$

The standard method of maximizing Eq. (5) subject to constraint Eqs. (6) and (7) is to form a Langrangian L – a composite of the entropy and the constraints

$$L = -\sum_i p_i \ln p_i - \beta \left(\sum_i p_i - 1 \right) - \vartheta \left(\sum_i p_i P_i - \bar{P} \right) \tag{8}$$

where β and ϑ are multipliers designed to ensure that the constraints are met. Maximizing (8) with respect to p_i gives

$$\frac{\partial L}{\partial p_i} = \ln p_i - 1 - \beta - \vartheta P_i = 0 \,, \tag{9}$$

leading directly to a form for p_i which is

$$p_i = \exp(-\beta - 1)\exp(-\lambda P_i) = K \exp(-\vartheta P_i) \,. \quad (10)$$

K is the composite constant of proportionality which ensures that the probabilities sum to 1. Note also that the sign of the parameters is determined from data through the constraints. If we substitute the probability in Eq. (10) into the Shannon entropy, the measure of complexity of this system which is at a maximum for the given set of constraints, simplifies to $H = \beta + 1 + \vartheta \bar{P}$. There are various interpretations of this entropy with respect to dispersion of activities in the system although these represent a trade-off between the form of the distribution, in this case, the negative exponential, and the number of events or objects n which characterize the system.

Distributions and Densities of Population

The model we have derived can be regarded as an approximation to the distribution of population densities over a set of n spatial zones as long as each zone is the same size (area), that is, $A_i = A, \forall_i$ where nA is the total size (area) of the system. A more general form of entropy takes this area into account by maximizing the expected value of the logarithm of the density, not distribution, where the 'spatial' entropy is defined as

$$S = -\sum_i p_i \ln \frac{p_i}{A_i} \,, \quad (11)$$

with the probability density as p_i/A_i. Using this formula, the procedure simply generalizes the maximization to densities rather than distributions [10] and the model we have derived simply determines these densities with respect to an average population size $\bar{P}$. If we order populations over the zones of a city or even take their averages over many cities in a region or nation, then they are likely to be distributed in this fashion; that is, we would expect there to be many fewer zones or cities of high density than zones or cities of low density, due to competition through growth.

However the way this method of entropy-maximizing has been used to generate population densities in cities is to define rather more specific constraints that relate to space. Since the rise of the industrial city in the 19th century, we have known that population densities tend to decline monotonically with distance from the center of the city. More than 50 years ago, Clark [29] demonstrated quite clearly that population densities declined exponentially with distance from the center of large cities and in the 1960s with the application of micro-economic theory to urban location theory following von Thünen's [68] model, a range of urban attributes such as rents, land values, trip densities, and population densities were shown to be consistent with such negative exponential distributions [5]. Many of these models can also be generated using utility maximizing which under certain rather weak constraints can be seen as equivalent to entropy-maximizing [6]. However it is random utility theory that has been much more widely applied to generate spatial interaction models with a similar form to the models that we generate below using entropy-maximizing [20,46].

We will show how these typical micro-economic urban density distributions can be derived using entropy-maximizing in the following way. Maximizing S in Eq. (11) or H in Eq. (5) where we henceforth assume that the probability p_i is now the population density, we invoke the usual normalization constraint in Eq. (6) and a constraint on the average travel cost $\bar{C}$ incurred by the population given as $\sum_i p_i c_i = \bar{C}$ where c_i is the generalized travel cost/distance from the central business district (CBD) to a zone i. This maximization leads to

$$p_i = K \exp(-\mu c_i) \quad (12)$$

where μ is the parameter controlling the rate of decay of the exponential function, sometimes called the 'friction' of distance or travel cost.

Gravitational Models of Spatial Interaction

It is a simple matter to generalize this framework to generate arrangements of urban activities that deal with interaction patterns, that is movements or linkages between pairs of zones. This involves extending entropy to deal with two rather than one dimensional systems where the focus of interest is on the interaction between an origin zone called $i, i = 1, 2, \ldots, I$ and a destination zone $j, j = 1, 2, \ldots, J$ where there are now a total of IJ interactions in the system. These kinds of model can be used to simulate routine trips from home to work, for example, or to shop, longer term migrations in search of jobs, moves between residential locations in the housing market, as well as trade flows between countries and regions. The particular application depends on context as the generic framework is independent of scale.

Let us now define a two-dimensional entropy as

$$H = -\sum_i \sum_j p_{ij} \ln p_{ij} \,. \quad (13)$$

p_{ij} is the probability of interaction between origin i and destination j where the same distinctions between distribution and density noted above apply. Without loss of generality, we will assume in the sequel that these variables p_{ij} covary with density in that the origin and destination zones all have the same area. The most constrained system is where we assume that all the interactions originating from any zone i must sum to the probability p_i of originating in that zone, and all interactions destined for zone j must sum to the probability p_j of being attracted to that destination zone. There is an implicit constraint that these origin and destination probabilities sum to 1, that is

$$\sum_i \sum_j p_{ij} = \sum_i p_i = \sum_j p_j = 1 \,, \tag{14}$$

but Eq. (14) is redundant with respect to the origin and destination normalization constraints which are stated explicitly as

$$\left.\begin{array}{l} \sum\limits_j p_{ij} = p_i \\ \sum\limits_i p_{ij} = p_j \end{array}\right\} . \tag{15}$$

There is also a constraint on the average distance or cost traveled given as

$$\sum_i \sum_j p_{ij} c_{ij} = \bar{C} \,. \tag{16}$$

The model that is derived from the maximization of Eq. (13) subject to Eqs. (15) and (16) is

$$p_{ij} = K_i K_j p_i p_j \exp(-\gamma c_{ij}) \tag{17}$$

where K_i and K_j are normalization constants associated with Eq. (15), and γ is the parameter on the travel cost c_{ij} between zones i and j associated with Eq. (16). It is easy to compute K_i and K_j by substituting for p_{ij} from Eq. (17) in Eq. (15) respectively and simplifying. This yields

$$\left.\begin{array}{l} K_i = \dfrac{1}{\sum\limits_j K_j p_j \exp(-\gamma c_{ij})} \\ K_j = \dfrac{1}{\sum\limits_i K_i p_i \exp(-\gamma c_{ij})} \end{array}\right\} , \tag{18}$$

equations that need to be solved iteratively.

These models can be scaled to deal with real trips or population simply by multiplying these probabilities by the total volumes involved, T for total trips in a transport system, P for total population in a city system, Y for total income in a trading system and so on. This system however forms the basis for a family of interaction models which can be generated by relaxing the normalization constraints; for example by omitting the destination constraint, $K_j = 1, \forall_j$, or by omitting the origin constraint, $K_i = 1, \forall_i$ or by omitting both where we need an explicit normalization constraint of the form $\sum_{ij} p_{ij} = 1$ in Eq. (14) to provide an overall constant K. Wilson [74] refers to this set of four models as: *doubly-constrained* – the model in Eqs. (17) and (18), the next two as *singly-constrained*, first when $K_i = 1, \forall_i$, the model is origin constrained, and second when $K_j = 1, \forall_j$, the model is destination constrained; and when we have no constraints on origins or destinations, we need to invoke the global constant K and the model is called *unconstrained*. It is worth noting that these models can also be generated in nearly equivalent form using random utility theory where they are articulated at the level of the individual rather than the aggregate trip-maker and are known as discrete choice models [20].

Let us examine one of these models, a singly-constrained model where there are origin constraints. This might be a model where we are predicting interactions from work to home given we know the distribution of work at the origin zones. Then noting that $K_j = 1, \forall_j$, the model is

$$p_{ij} = K_i p_i p_j \exp(-\gamma c_{ij}) = p_i \frac{p_j \exp(-\gamma c_{ij})}{\sum\limits_j p_j \exp(-\gamma c_{ij})} \,. \tag{19}$$

The key issue with this sort of model is that not only are we predicting the interaction between zones i and j but we can predict the probability of locating in the destination zone p'_j, that is

$$p'_j = \sum_i p_{ij} = \sum_i p_i \frac{p_j \exp(-\gamma c_{ij})}{\sum\limits_j p_j \exp(-\gamma c_{ij})} \,. \tag{20}$$

If we were to drop both origin and destination constraints, the model becomes one which is analogous to the traditional gravity model from which it was originally derived prior to the development of these optimization frameworks. However to generate the usual standard gravitational form of model in which the 'mass' of each origin and destination zone appears, given by P_i and P_j respectively, then we need to modify the entropy formula, thus maximizing

$$H = -\sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{P_i P_j} \,, \tag{21}$$

subject to the normalization

$$\sum_i \sum_j p_{ij} = 1 \,, \tag{22}$$

and this time a constraint on the average 'logarithmic' travel cost $\overline{\ln C}$

$$\sum_i \sum_j p_{ij} \ln c_{ij} = \overline{\ln C} . \tag{23}$$

The model that is generated from this system can be written as

$$p_{ij} = K \frac{P_i P_j}{c_{ij}^{\eta}} , \tag{24}$$

where the effect of travel cost/distance is now in power law form with η the scaling parameter. Besides illustrating the fact that inverse power forms as well as negative exponential distributions can be generated in this way according to the form of the constraints, one is also able to predict both the probabilities of locating at the origins and the destinations from the traditional gravity model in Eq. (24).

Scaling, City Size, and Network Structure: Power Laws

Distance is a key organizing concept in city systems as we have already seen in the way various urban distributions have been generated. Distance is an attribute of nearness or proximity to the most accessible places and locations. Where there are the lowest distance or travel costs to other places, the more attractive or accessible are those locations. In this sense, distance or travel cost acts as an inferior good in that we wish to minimize the cost occurred in overcoming it. Spatial competition also suggests that the number of places that have the greatest accessibilities are few compared to the majority of places. If you consider that the most accessible place in a circular city is the center, then assuming each place is of similar size, as the number of places by accessibility increases, the lower the accessibility is. In short, there are many places with the same accessibility around the edge of the city compared to only one place in the center. The population density model in Eq. (12) implies such an ordering when we examine the frequency distribution of places according to their densities.

If we now forget distance for a moment, then it is likely that the distribution of places at whatever scale follows a distribution which declines in frequency with attributes based on size due to competition. If we look at all cities in a nation or even globally, there are far fewer big cities than small ones. Thus the entropy-maximizing framework that we have introduced to predict the probability (or frequency) of objects of a certain size occurring, is quite applicable in generating such distributions. We derived a negative exponential distribution in Eq. (10) but to generate a power law, all we need to do is to replace the constraint in Eq. (7) with its logarithmic equivalent, that is

$$\sum_i p_i \ln P_i = \overline{\ln P} , \tag{25}$$

and then maximize Eq. (5) subject to (6) and (25) to give

$$p_i = K P_i^{-\varphi} = \exp(-\beta - 1) \exp(-\varphi \ln P_i) , \tag{26}$$

where φ is the scaling parameter. Equation (26) gives the probability or frequency – the number of cities – for a zone (or city) with P_i population which is distributed according to an inverse power law. It is important to provide an interpretation of the constraint which generates this power law. Equation (25) implies that the system conserves the average of the logarithm of size which gives greater weight to smaller values of population than to larger, and as such, is recognition that the average size of the system is unbounded as a power function implies. With such distributions, it is unlikely that normality will prevail due to the way competition constrains the distribution in the long tail. Nevertheless in the last analysis, it is an empirical matter to determine the shape of such distributions from data, although early research on the empirical distributions of city sizes following Zipf's Law [81] by Curry [34] and Berry [21] introduced the entropy-maximizing framework to generate such size distributions.

The power law implied for the probability p_i of a certain size P_i of city or zone can be easily generalized to a two-dimensional equivalent which implies a network of interactions. We will maximize the two-dimensional entropy H in Eq. (13) subject to constraints on the mean logarithm of population sizes at origins and destinations which we now state as

$$\left.\begin{array}{l} \sum_i \sum_j p_{ij} \ln P_i = \sum_i p_i \ln P_i = \bar{P}_{\text{origins}} \\ \sum_i \sum_j p_{ij} \ln P_j = \sum_j p_j \ln P_j = \bar{P}_{\text{destinations}} \end{array}\right\} , \tag{27}$$

where $p_i = \sum_j p_{ij}$ and $p_j = \sum_i p_{ij}$. Note however that there are no constraints on these origins and destination probabilities p_i and p_j per se but the global constraints in Eq. (14) must hold. This maximization leads to the model

$$p_{ij} = K P_i^{-\lambda_i} P_j^{-\lambda_j} = \frac{P_i^{-\lambda_i}}{\sum_i P_i^{-\lambda_i}} \frac{P_j^{-\lambda_j}}{\sum_j P_j^{-\lambda_j}} , \tag{28}$$

where it is clear that the total flows from any origin node or location i vary as

$$p_i' \propto P_i^{-\lambda_i} , \tag{29}$$

and the flows into any destination zone vary as

$$p'_j \propto P_j^{-\lambda_j} , \tag{30}$$

with the parameters λ_i and λ_j relating to the mean of the observed logarithmic populations associated with the constraint Eq. (27). Note that the probabilities for each origin and destination node or zone are independent from one another as there is no constraint tying them together as in the classic spatial interaction model where distance or travel cost is intrinsic to the specification.

These power laws can be related to recent explorations in network science which suggest that the number of in-degrees – the volume of links entering a destination in our terms – and the number of out-degrees – the volume emanating from an origin, both follow power laws [2]. These results have been widely observed in topological rather than planar networks where the focus is on the numbers of physical links associated with nodes rather than the volume of traffic on each link. Clearly the number of physical links in planar graphs is limited and the general finding from network science that the number of links scales as a power law cannot apply to systems that exist in two-dimensional Euclidean space [24]. However a popular way of transforming a planar graph into one which is non-planar is to invoke a rule that privileges some edges over others merging these into long links and then generating a topology which is based on the merged edges as constituting nodes and the links between the new edges as arcs. This is the method that is called space syntax [47] and it is clear that by introducing order into the network in this way, the in-degrees and out-degrees of the resulting topological graph can be scaling. Jiang [48] illustrates this quite clearly although there is some reticence to make such transformations and where planar graphs have been examined using new developments in network science based on small worlds and scale-free graph theory, the focus has been much more on deriving new network properties than on appealing to any scale-free structure [33].

However to consider the scale-free network properties of spatial interaction systems, each trip might be considered a physical link in and of itself, albeit that it represents an interaction on a physical network as a person making such an interaction is distinct in space and time. Thus the connections to network science are close. In fact the study of networks and their scaling properties has not followed the static formulations which dominate our study of cities in equilibrium for the main way in which such power laws are derived for topological networks is through a process of preferential attachment which grows networks from a small number of seed nodes [8]. Nevertheless, such dynamics appear quite consistent with the evolution of spatial interaction systems.

These models will be introduced a little later when urban dynamics are being dealt with. For the moment, let us note that there are various simple dynamics which can account not only for the distribution of network links following power laws, but also for the distribution of city sizes, incomes, and a variety of other social (and physical) phenomena from models that grow the number of objects according to simple proportionate growth consistent with the generation of lognormal distributions. Suffice it to say that although we have focused on urban densities as following either power laws or negative exponential functions in this section, it is entirely possible to use the entropy-maximizing framework to generate distributions which are log-normal, another alternative with a strong spatial logic. Most distributions which characterize urban structure and activities however are not likely to be normal and to conclude this section, we will review albeit very briefly, some empirical results that indicate the form and pattern of urban activities in western cities.

Empirical Applications: Rank-Size Representations of Urban Distributions

The model in Eq. (26) gives the probability of location in a zone i as an inverse power function of the population or size of that place which is also proportional to the frequency

$$f(p_i) \propto p_i = KP_i^{-\varphi} . \tag{31}$$

It is possible to estimate the scaling parameter φ in many different ways but a first test of whether or not a power law is likely to exist can be made by plotting the logarithms of the frequencies and population sizes and noting whether or not they fall onto a straight line. In fact a much more preferable plot which enables each individual observation to be represented is the cumulative function which is formed from the integral of Eq. (31) up to a given size; that is $F_i \propto P_i^{-\varphi+1}$. The counter-cumulative $F - F_i$ where F is the sum of all frequencies in the system – that is the number of events or cities – also varies as $P_i^{-\varphi+1}$ and is in fact the rank of the city in question. Assuming each population size is different, then the order of $\{i\}$ is the reverse of the rank, and we can now write the rank r of i as $r = F - F_i$. The equation for this rank-size distribution (which is the one that is usually used to fit the data) is thus

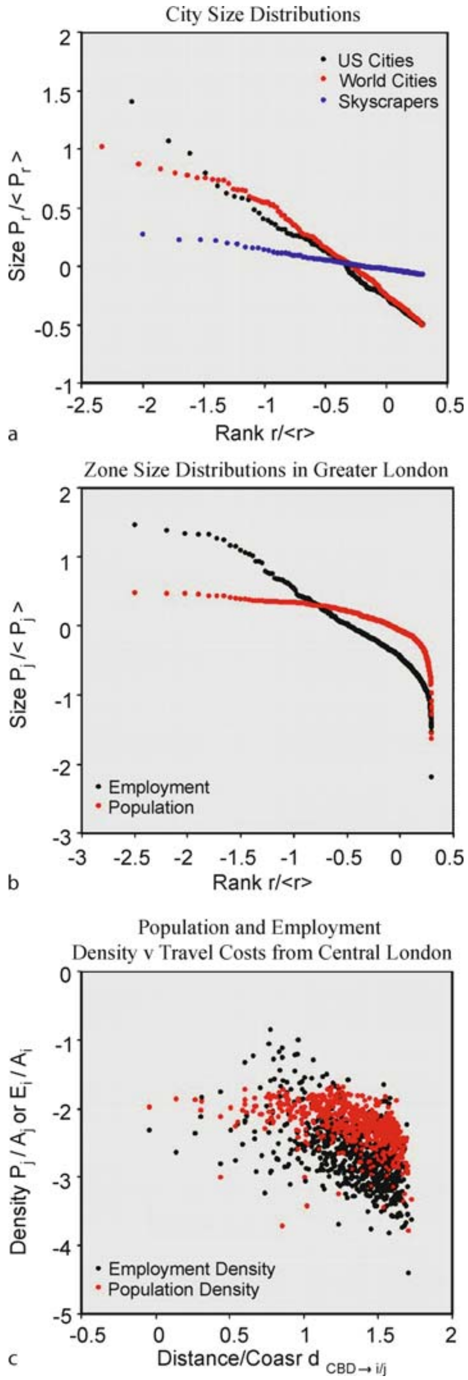
$$r = GP_r^{-\varphi+1} \tag{32}$$

where G is a scaling constant which in logarithmic form is $\ln r = G - (\varphi - 1) \ln P_r$. This is the equation that is implicit in the rank-size plots presented below which reveal evidence of scaling.

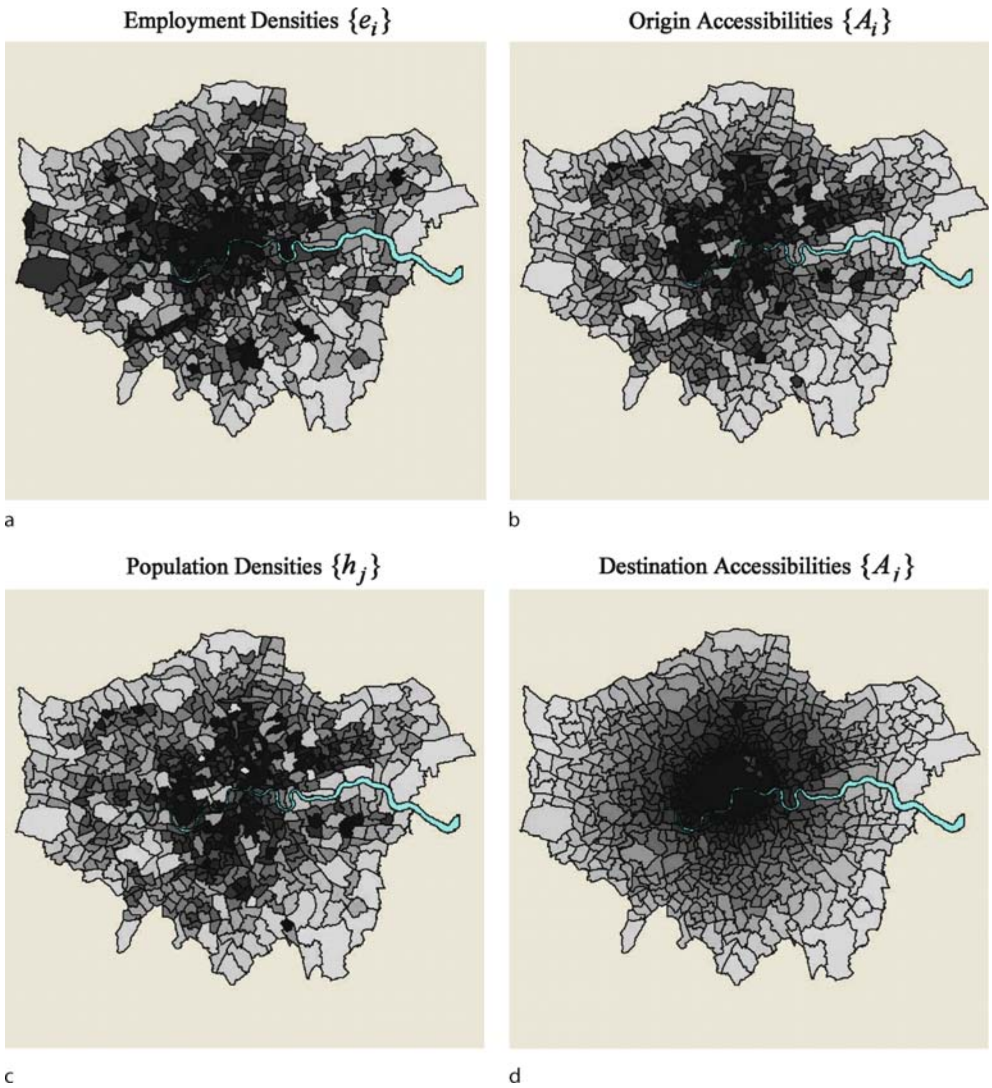
First let us examine the scaling which is implicit in urban size distributions for the largest world city populations over 1 million in 2005, for cities over 100,000 in the USA in 2000, and for the 200 tallest buildings in the world in 2007. We could repeat such examples ad nauseum but these provide a good selection which we graph in rank-size logarithmic form in Fig. 1a, noting that we have normalized all the data by their means, that is by $<P_r>$ and $<r>$, as $P_r/<P_r>$ and $r/<r>$. We are only examining a very small number at the very top of the distribution and this is clearly not definitive evidence of scaling in the rest of the distribution but these plots do show the typical distributions of city size activities that have been observed in this field for over 50 years. As we will imply later, these signatures are evidence of self-organization and fractal structure which emerge through competition from the bottom up [15].

To illustrate densities in cities, we take employment and working population in small zones in Greater London, a city which has some 4.4 million workers. We rank-order the distribution in the same way we have done for world cities, and plot these, suitably normalized by their means, logarithmically in Fig. 1b. These distributions are in fact plotted as densities so that we remove aerial size effects. Employment densities $e_i = E_i/A_i$ can be interpreted as the number of work trips originating in employment zones $e_i = \sum_j T_{ij}$ – the volume of the out-degrees of each employment zone considered as nodes in the graph of all linkages between all places in the system, and population densities $h_j = P_j/A_j$ as the destination distributions $h_j = \sum_i T_{ij}$ – the in-degrees which measure all the trips destined for each residential zone from all employment zones. In short if there is linearity in the plots, this is evidence that the underlying interactions on the physical networks that link these zones are scaling. Figure 1b provides some evidence of scaling but the distributions are more similar to lognormal distributions than to power laws. This probably implies that the mechanisms for generating these distributions are considerably more complex than growth through preferential attachment which we will examine in more detail below [15].

Lastly, we can demonstrate that scaling in city systems also exists with respect to how trips, employment and population activities vary with respect to distance. In Fig. 1c, we have again plotted the employment densities $e_i = E_i/A_i$ at origin locations and population densities $h_j = P_j/A_j$ at destination locations but this time against distances $d_{\text{CBD}\to i}$ and $d_{\text{CBD}\to j}$ from the center of London's CBD in logarithmic terms. It is clear that there is significant correlation but also a very wide spread of values around the log-linear regression lines due to the fact that the city is multi-centric. Nevertheless the relationships appears to be scaling with these estimated as $e_i = 0.042 d^{-0.98}_{\text{CBD}\to i}$, $(r^2 = -0.30)$, and $h_j = 0.029 d^{-0.53}_{\text{CBD}\to j}$,



Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 1
Scaling distributions in world cities and in Greater London


Employment Densities $\{e_i\}$
Origin Accessibilities $\{A_i\}$
a
b
Population Densities $\{h_j\}$
Destination Accessibilities $\{A_j\}$
c
d

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 2
Employment, population and accessibilities in Greater London (greatest extent is 55kms east to west; 45kms north to south)

($r^2 = -0.23$). However more structured spatial relationships can be measured by accessibilities which provide indices of overall proximity to origins or destinations, thus taking account of the fact that there are several competing centers. Accessibility can be measured in many different ways but here we use a traditional definition of potential based on employment accessibility A_i to populations at destinations, and population accessibility A_j to employment at origins defined as

$$\left.\begin{aligned} A_i &\propto \sum_j \frac{h_j}{c_{ij}} \\ A_j &\propto \sum_i \frac{e_i}{c_{ij}} \end{aligned}\right\} , \tag{33}$$

where c_{ij} is, as before, the generalized cost of travel from employment origin i to population destination j. In Fig. 2a, b, we compare the distribution of employment densities e_i with accessibility origins A_i and in Fig. 2c, d, population densities h_j with accessibility destinations A_j. Each set of maps is clearly correlated with higher associations than in Fig. 1c which take account of only the single CBD. Regressing $\{\ln e_i\}$ on $\{\ln A_i\}$ and $\{\ln p_j\}$ on $\{\ln A_j\}$ gives an approximate scaling with 31% of the variance accounted for in terms of origin accessibility and 41% for destination accessibility. These relations appear linear but there is still considerable noise in the data which undoubtedly reflects the relative simplicity of the models and the fact that accessibility is being measured using current transport costs without any reference to the historical evolution of the city's structure. It is, however, building blocks such as these that constitute the basis for operational land use transport models that have developed for comparative static and quasi-dynamic forecasting that we will discuss below.

Urban Dynamics

Aggregate Development

Models of city systems have largely been treated as static for at first sight, urban structure in terms of its form and to some extent its function appears stable and long-lasting. During the industrial era, cities appeared to have a well-defined structure where land uses were arranged in concentric rings according to their productivity and wealth around a central focus, usually the central business district (CBD), the point where most cities were originally located and exchange took place. Moreover data on how cities had evolved were largely absent and this reinforced the focus on statics and equilibria. Where the need to examine urban change was urgent, models were largely fashioned in terms of the simplest growth dynamics possible and we will begin with these here.

The growth of human populations in their aggregate appears to follow an exponential law where the rate of change σ is proportional to the size of the population itself $P(t)$, that is

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = \sigma P(t) . \tag{34}$$

It is easy to show that starting from an initial population $P(0)$, the growth is exponential, that is

$$P(t) = P(0)\exp(\sigma t) , \tag{35}$$

which is the continuous form of model. When formulated discretely, at time steps $t = 1, 2, \dots, T$, Eq. (34) can be written as $P(t) - P(t-1) = \beta P(t-1)$ which leads to

$$P(t) = (1+\beta)P(t-1) . \tag{36}$$

Through time from the initial condition $P(0)$, the trajectory is

$$P(t) = (1+\beta)^t P(0) . \tag{37}$$

$1+\beta$ is the growth rate. If $\beta > 0$, Eq. (37) shows exponential growth, if $\beta < 0$, exponential decline, and if $\beta = 0$, the population is in the steady state and simply reproduces itself.

This simple growth model leads to *smooth change*, and any discontinuities or breaks in the trajectories of growth or decline must come about through an external change in the rate from the outside environment. If we assume the growth rate fluctuates around a mean of one with β varying randomly, above -1, then it is not possible to predict the trajectory of the growth path. However if we have a large number of objects which we will assume to be cities whose growth rates are chosen randomly, then we can write the growth equation for each city as

$$P_i(t) = [1+\beta_i(t)]P_i(t-1) , \tag{38}$$

which from an initial condition $P_i(0)$ gives

$$P_i(t) = \prod_{\tau=1}^{t} [1+\beta_i(\tau)]P_i(0) . \tag{39}$$

This is growth by proportionate effect; that is, each city grows in proportion to its current size but the growth rate in each time period is random. In a large system of cities, the ultimate distribution of these population sizes will be lognormal. This is easy to demonstrate for the logarithm of Eq. (39) can be approximated by

$$\ln P_i(t) = \ln P_i(0) + \sum_{\tau=1}^{t} \beta_i(\tau) , \tag{40}$$

where the sum of the random components is an approximation to the log of the product term in Eq. (39) using Taylor's expansion. This converges to the lognormal from the law of large numbers. It was first demonstrated by Gibrat [43] for social systems but is of considerable interest here in that the fat tail of the lognormal can be approximated by an inverse power law. This has become the default dynamic model which underpins an explanation of the rank-size rule for city populations first popularized by Zipf [81] and more recently confirmed by Gabaix [41] and Blank and Solomon [23] among others. We demonstrated this in Fig. 1a for the world city populations greater than 1 million and for US city populations greater than 100,000. As such, it is the null hypothesis for the distribution of urban populations in individual cities as well as population locations within cities.

Although Gibrat's model does not take account of interactions between the cities, it does introduce diversity into the picture, simulating a system that in the aggregate is non-smooth but nevertheless displays regularity. These links to aggregate dynamics focus on introducing slightly more realistic constraints and one that is of wide relevance is the introduction of capacity constraints or limits on the level to which a population might grow. Such capacitated growth is usually referred to as logistic growth. Retaining the exponential growth model, we can limit this by moderating the growth rate σ according to an upper limit on population $P_{\max}$ which changes the model in Eq. (34) and the growth rate σ to

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = \left[\sigma\left(1 - \frac{P(t)}{P_{\max}}\right)\right] P(t)\,. \tag{41}$$

It is clear that when $P(t) = P_{\max}$, the overall rate of change is zero and no further change occurs. The continuous version of this logistic is

$$P(t) = \frac{P_{\max}}{1 + \left(\frac{P_{\max}}{P(0)} - 1\right)\exp(-\sigma t)}\,, \tag{42}$$

where it is easy to see that as $t \to \infty$, $P(t) \to P_{\max}$.

The discrete equivalent of this model in Eq. (41) follows directly from $P(t) - P(t-1) = \beta[1 - (P(t-1)/P_{\max})]P(t-1)$ as

$$P(t) = \left[1 + \beta\left(1 - \frac{P(t-1)}{P_{\max}}\right)\right] P(t-1)\,, \tag{43}$$

where the long term dynamics is too intricate to write out as a series. Equation (43) however shows that the growth component β is successively influenced by the growth of the population so far, thus preserving the capacity limit through the simple expedient of adjusting the growth rate downwards. As in all exponential models, it is based on proportionate growth. As we noted above, we can make each city subject to a random growth component $\beta_i(t)$ while still keeping the proportionate effect.

$$P(t) = \left[1 + \beta_i(t)\left(1 - \frac{P(t-1)}{P_{\max}}\right)\right] P(t-1)\,. \tag{44}$$

This model has not been tested in any detail but if $\beta_i(t)$ is selected randomly, the model is a likely to generate a lognormal-like distribution of cities but with upper limits being invoked for some of these. In fact, this stochastic equivalent also requires a lower integer bound on the size of cities so that cities do not become too small [14]. Within these limits as long as the upper limits are not too tight, the sorts of distributions of cities that we observe in the real world are predictable.

In the case of the logistic model, remarkable and unusual discontinuous nonlinear behavior can result from its simple dynamics. When the β component of the growth rate is $\beta < 2$, the predicted growth trajectory is the typical logistic which increases at an increasing rate until an inflection point after which the growth begins to slow, eventually converging to the upper capacity limit of $P_{\max}$. However when $\beta \cong 2$, the population oscillates around this limit, bifurcating between two values. As the value of the growth rate increases towards 2.57, these oscillations get greater, the bifurcations doubling in a regular but rapidly increasing manner. At the point where $\beta \cong 2.57$, the oscillations and bifurcations become infinite, apparently random, and this regime persists until $\beta \cong 3$ during which the predictions look entirely chaotic. In fact, this is the regime of 'chaos' but chaos in a controlled manner from a deterministic model which is not governed by externally induced or observed randomness or noise.

These findings were found independently by May [52], Feigenbaum [38], Mandelbot [51] among others. They relate strongly to bifurcation and chaos theory and to fractal geometry but they still tend to be of theoretical importance only. Growth rates of this magnitude are rare in human systems although there is some suggestion that they might occur in more complex coupled biological systems of predator-prey relations. In fact one of the key issues in simulating urban systems using this kind of dynamics is that although these models are important theoretical constructs in defining the scope of the dynamics that define city systems, much of these dynamic behaviors are simplistic. In so far as they do characterize urban systems, it is at the highly aggregate scale as we demonstrate a little later. The use of these ideas in fact is much more applicable to extending the static equilibrium models of the last section and to demonstrate these, we will now illustrate how these models might be enriched by putting together logistic behaviors with spatial movement and interaction.

One way of articulating urban dynamics at the intra-urban level is to identify different speeds of change. In particular we can define a fast dynamics that relates to how people might move around the city on daily basis, for example, in terms of the journey to work, and a slower dynamics that relates to more gradual change that relates to the size of different locations affected by residential migrations. We can model the fast dynamics using a singly-constrained spatial interaction which distributes workers to residential locations which we define using previous notation where all variables are now time scripted by (t): $T_{ij}(t)$ trips between zones i and j, employment $E_i(t)$ at origin zone i, population $P_j(t)$ at destination zone j, the friction of distance parameter $\gamma(t)$, and the travel cost $c_{ij}(t)$ between zones i and j. The model is defined as

$$T_{ij}(t) = E_i(t)\frac{P_j(t)\exp[-\gamma(t)c_{ij}(t)]}{\sum_j P_j(t)\exp[-\gamma(t)c_{ij}(t)]}\,, \tag{45}$$

from which we can predict residential population $P'_j(t)$ as

$$\begin{aligned} P'_j(t) &= \sum_i T_{ij}(t) \\ &= \sum_i E_i(t) \frac{P_j(t) \exp[-\gamma(t) c_{ij}(t)]}{\sum_j P_j(t) \exp[-\gamma(t) c_{ij}(t)]} \,. \end{aligned} \tag{46}$$

This is the fast dynamics but each zone is capacitated by an unchanging upper limit on population where the zonal population changes slowly in proportion to its existing size through internal migration and in response to the upper limit $P_{j\max}$. The change in terms of this slower dynamic from t to $t+1$ is modeled as

$$\Delta P_j(t+1) = \beta[P_{j\max} - P'_j(t)] P'_j(t) \tag{47}$$

with the long term trajectory thus given as

$$P_j(t+1) = \left(1 + \beta[P_{j\max} - P'_j(t)]\right) P'_j(t) \,. \tag{48}$$

Clearly $P_j(t)$ will converge to $P_{j\max}$ as long as $P'_j(t)$ is increasing while the fast dynamics is also updated in each successive time period from

$$\begin{aligned} P'_j(t+1) &= \sum_i T_{ij}(t+1) \\ &= \sum_i E_i(t+1) \frac{P_j(t+1) \exp[-\gamma(t+1) c_{ij}(t+1)]}{\sum_j P_j(t+1) \exp[-\gamma(t+1) c_{ij}(t+1)]} \,. \end{aligned} \tag{49}$$

We may have an even slower dynamics relating to technological or other social change which changes $P_{j\max}$ while various other models may be used to predict employment for example, which itself may be a function of another fast dynamics relating to industrial and commercial interactions. The time subscripted variables travel $c_{ij}(t+1)$ and the friction of distance parameter $\gamma(t+1)$ might be changes that reflect other time scales. We might even have lagged variables independently introduced reflecting stocks or flows at previous time periods $t-1$, $t-2$ etc. Wilson [75,76] has explored links between these spatial interaction entropy-maximizing models and logistic growth and has shown that in a system of cities or zones within a city, unusual bifurcating behavior in terms of the emergence of different zonal centers can occur when parameter values, particularly the travel cost parameter $\gamma(t+1)$, cross certain thresholds.

There have been many proposals involving dynamic models of city systems which build on the style of nonlinear dynamics introduced here and these all have the potential to generate discontinuous behavior. Although Wilson [75] pioneered embedding dynamic logistic change into spatial interaction models, there have been important extensions to urban predator-prey models by Dendrinos and Mullaly [36] and to bifurcating urban systems by Allen [3,4], all set within a wider dynamics linking macro to micro through master equation approaches [45]. A good summary is given by Nijkamp and Reggiani [57] but most of these have not really led to extensive empirical applications for it has been difficult to find the necessary rich dynamics in the sparse temporal data sets available for cities and city systems; at the macro-level, a lot of this dynamics tends to be smoothed away in any case. In fact, more practical approaches to urban dynamics have emerged at finer scale levels where the agents and activities are more disaggregated and where there is a stronger relationship to spatial behavior. We will turn to these now.

Dynamic Disaggregation: Agents and Cells

Static models of the spatial interaction variety have been assembled into linked sets of sub-models, disaggregated into detailed types of activity, and structured so that they simulate changes in activities through time. However, the dynamics that is implied in such models is simplistic in that the focus has still been very much on location in space with time added as an afterthought. Temporal processes are rarely to the forefront in such models and it is not surprising that a more flexible dynamics is emerging from entirely different considerations. In fact, the models of this section come from dealing with objects and individuals at much lower/finer spatial scales and simulating processes which engage them in decisions affecting their spatial behavior. The fact that such decisions take place through time (and space) makes them temporal and dynamic rather through the imposition of any predetermined dynamic structures such as those used in the aggregate dynamic models above. The models here deal with individuals as agents, rooted in cells which define the space they occupy and in this sense, are highly disaggregate as well as dynamic. These models generate development in cities from the bottom up and have the capability of producing patterns which are emergent. Unlike the dynamic models of the last section, their long term spatial behavior can be surprising and often hard to anticipate.

It is possible however to use the established notation for equilibrium models in developing this framework based on the generic dynamic $P_i(t) = P_i(t-1) + \Delta P_i(t)$ where the change in population $\Delta P_i(t)$ can be divided into two components. The first is the usual proportionate effect, the positive feedback induced by population on itself which is defined as the *reactive* element of change $\omega P_i(t-1)$. The second is the *interactive* element, change that is generated from some action-at-a-distance which is often regarded as a diffusion of population from other locations in the system. We can model this in the simplest way using the traditional gravity model in Eq. (24) but noting that we must sum the effects of the diffusion over the destinations from where it is generated as a kind of accessibility or potential. The second component of change is $\phi P_i(t-1)K\sum_j P_j(t-1)/c_{ij}^{\eta}$ from which the total change between t and $t-1$ is

$$\Delta P_i(t) = \omega P_i(t-1) + \phi P_i(t-1)K\sum_j \frac{P_j(t-1)}{c_{ij}^{\eta}} + \varepsilon_i(t-1). \tag{50}$$

We have also added a random component $\varepsilon_i(t-1)$ in the spirit of our previous discussion concerning growth rates. We can now write the basic reaction-diffusion equation, as it is sometimes called, as

$$\begin{aligned} &P_i(t) \\ &= P_i(t-1) + \Delta P_i(t) \\ &= P_i(t-1)\left(1+\omega+\phi K\sum_j \frac{P_j(t-1)}{c_{ij}^{\eta}} + \varepsilon_i(t-1)\right). \end{aligned} \tag{51}$$

This equation looks as though it applies to a zonal system but we can consider each index i or j simply a marker of location, and each population activity can take on any value; for single individuals it can be 0 or 1 while it might represent proportions of an aggregate population or total numbers for the framework is entirely generic. As such, it is more likely to mirror a slow dynamics of development rather than a fast dynamics of movement although movement is implicit through the diffusive accessibility term.

We will therefore assume that the cells are small enough, space-wise, to contain single activities – a single household or land use which is the cell state – with the cellular tessellation usually forming a grid associated with the pixel map used to visualize data input and model output. In terms of our notation, population in any cell i must be $P_i(t) = 1$ or 0, representing a cell which is occupied or empty with the change being $\Delta P_i(t) = -1$ or 0 if $P_i(t-1) = 1$ and $\Delta P_i(t) = 1$ or 0 if $P_i(t-1) = 0$. These switches of state are not computed by Eq. (51) for the way these cellular variants are operationalized is through a series of rules, constraints and thresholds. Although consistent with the generic model equations, these are applied in more ad hoc terms. Thus these models are often referred to as automata and in this case, as cellular automata (CA).

The next simplification which determines whether or not a CA follows a strict formalism, relates to the space over which the diffusion takes place. In the fast dynamic equilibrium models of the last section and the slower ones of this, interaction is usually possible across the entire space but in strict CA, diffusion is over a local neighborhood of cells around i, Ω_i, where the cells are adjacent. For symmetric neighborhoods, the simplest is composed of cells which are north, south, east and west of the cell in question, that is $\Omega_i = n, s, e, w$ – the so-called von Neumann neighborhood, while if the diagonal nearest neighbors are included, then the number of adjacent cells rises to 8 forming the so-called Moore neighborhood. These highly localized neighborhoods are essential to processes that grow from the bottom up but generate global patterns that show emergence. Rules for diffusion are based on switching a cell's state on or off, dependent upon what is happening in the neighborhood, with such rules being based on counts of cells, cell attributes, constraints on what can happen in a cell, and so on.

The simplest way of showing how diffusion in localized neighborhoods takes place can be demonstrated by simplifying the diffusion term in Eq. (50) as follows. Then $\phi K\sum_j P_j(t-1) = \phi K\sum_j P_j(t-1)c_{ij}^{-\eta}$ as $c_{ij} = 1$ when $\Omega_i = n, s, e, w$. The cost is set as a constant value as each cell is assumed to be small enough to incur the same (or no) cost of transport between adjacent cells. Thus the diffusion is a count of cells in the neighborhood i. The overall growth rate is scaled by the size of the activity in i but this activity is always either $P_i(t-1) = 1$ or 0, presence or absence. In fact this scaling is inappropriate in models that work by switching cells on and off for it is only relevant when one is dealing with aggregates. This arises from the way the generic equation in (51) has been derived and in CA models, it is assumed to be neutral. Thus the change Eq. (50) becomes

$$\Delta P_i(t) = \omega + \phi K\sum_j P_j(t-1) + \varepsilon_i(t)\,, \tag{52}$$

where this can now be used to determine a threshold $Z_{\max}$ over which the cell state is switched. A typical rule might be

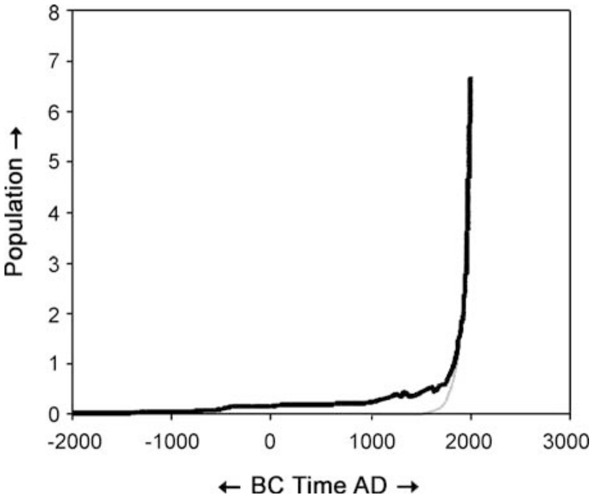
$$P_i(t) = \begin{cases} 1, & \text{if } [\omega + \phi K\sum_j P_j(t-1) + \varepsilon_i(t)] > Z_{\max} \\ 0, & \text{otherwise}\,. \end{cases} \tag{53}$$

It is entirely possible to separate the reaction from the diffusion and consider different combinations of these effects sparking off a state change. As we have implied, different combinations of attributes in cells and constraints within neighborhoods can be used to effect a switch, much depending on the precise specification of the model.

In many growth models based on CA, the strict limits posed by a local neighborhood are relaxed. In short, the diffusion field is no longer local but is an information or potential field consistent with its use in social physics where action-at-distance is assumed to be all important. In the case of strict CA, it is assumed that there is no action-at-a-distance in that diffusion only takes place to physically adjacent cells. Over time, activity can reach all parts of the system but it cannot hop over the basic cell unit. In cities, this is clearly quite unrealistic as the feasibility of deciding what and where to locate does not depend on physical adjacency. In terms of applications, there are few if any urban growth models based on strict CA although this does rather beg the question as to why CA is being used in the first place. In fact it is more appropriate to call such models cell-space or CS models as Couclelis [32] has suggested. In another sense, this framework can be considered as one for agent-based modeling where the cells are not agents and where there is no assumption of a regular underlying grid of cells [12,13]. There may be such a grid but the framework simply supposes that the indices *i* and *j* refer to locations that may form a regular tessellation but alternatively may be mobile and changing. In such cases, it is often necessary to extend the notation to deal with specific relations between the underlying space and the location of each agent.

Empirical Dynamics: Population Change and City Size

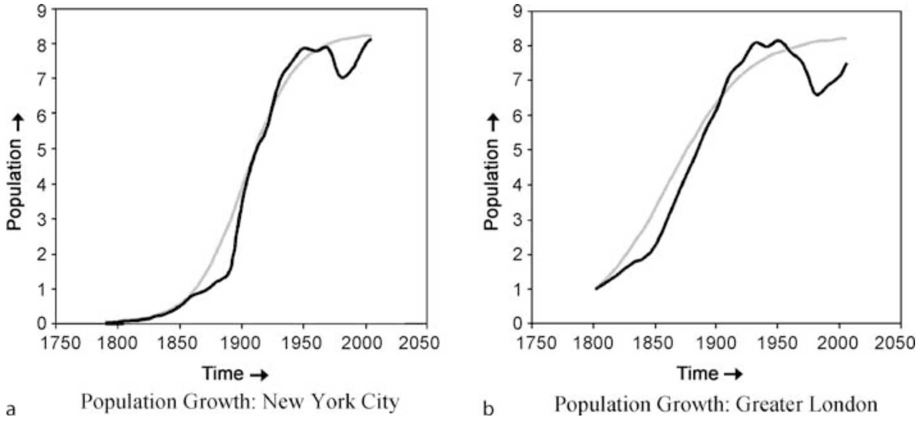
We will now briefly illustrate examples of the models introduced in this section before we then examine the construction of more comprehensive models of city systems. Simple exponential growth models apply to rapidly growing populations which are nowhere near capacity limits such as entire countries or the world. In Fig. 3, we show the growth of world population from 2000 BCE to date where it is clear that the rate of growth may be faster than the exponential model implies, although probably not as fast as double exponential. In fact world population is likely to slow rapidly over the next century probably mirroring global resource limits to an extent which are clearly illustrated in the growth of the largest western cities. In Figs. 4a, b, we show the growth in population of New York City (the five boroughs) and Greater London from 1750 to date and it is clear that in both cases, as the cities developed, population grew exponentially only to slow as the upper density limits of each city were reached.



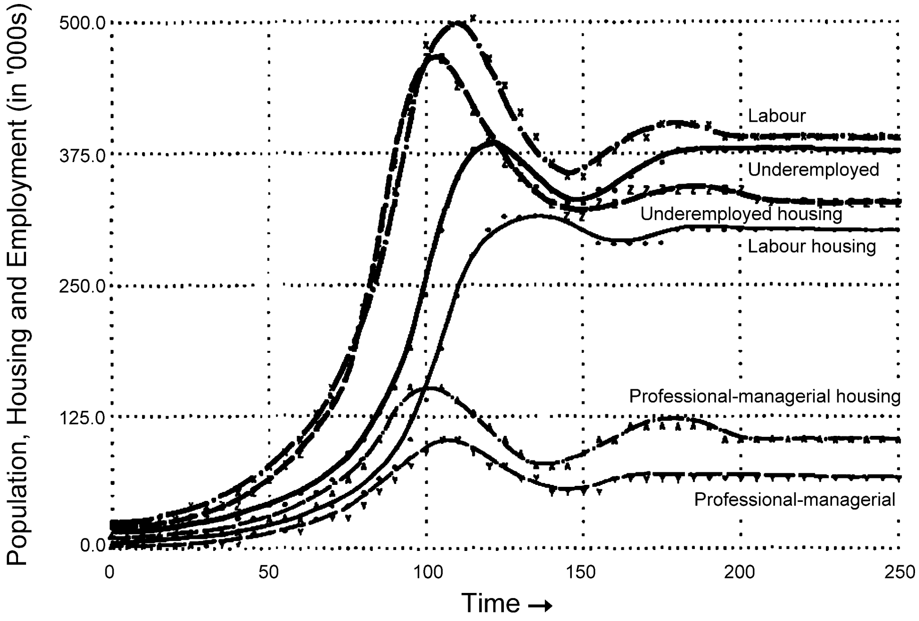
Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 3
Exponential world population growth. The *fitted exponential curve* is shown in *grey* where for the most part it is coincident with the observed growth, except for the very long period before the Industrial Revolution (before 1750)

Subsequent population loss and then a recent return of population to the inner and central city now dominate these two urban cores, which is reminiscent of the sorts of urban dynamic simulated by Forrester [39] where various leads and lags in the flow of populations mean that the capacity limit is often overshot, setting up a series of oscillations which damp in the limit. Forrester's model was the one of the first to grapple with the many interconnections between stocks and flows in the urban economy although these relationships were predicated hypothetically in simple proportionate feedback terms. Together they generated a rich dynamics but dominated by growth which was capacitated, thus producing logistic-like profiles with the leads and lags giving damped oscillations which we illustrate from his work in Fig. 5 [11]. We will see that the same phenomena can be generated from the bottom up as indeed Forrester's model implies, using cellular automata within a bounded spatial system.

Dynamics which arise from bottom-up urban processes can be illustrated for a typical CA/CS model, DUEM (Dynamic Urban Evolutionary Model) originally developed by Xie [78]. In the version of the model here, there are five distinct land uses – housing, manufacturing/primary industry, commerce and services, transport in the form of the street/road network, and vacant land. In principle, at each time period, each land use can generate quantities and locations of any other land use although in practice only industry, commerce and housing can gener-



Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 4
Logistic population growth



Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 5
Oscillating capacitated growth in a version of the Forrester *Urban Dynamics* model (from [11])

ate one another as well as generating streets. Streets do not generate land uses other than streets themselves. Vacant land is regarded as a residual available for development which can result from a state change (decline) in land use. The way the generation of land uses takes place is through a rule-based implementation of the generic Eq. (51) which enables a land use k, $P_i^k(t)$, to be generated from any other land use ℓ, $P_j^\ell(t-1)$. Land uses are also organized across a life cycle from initiating through mature to declining. Only initiating land uses which reflect their relative newness can spawn new land use. Mature remain passive in these terms but still influence new location while declining land uses disappear, thus reflecting completion of the life cycle of built form.

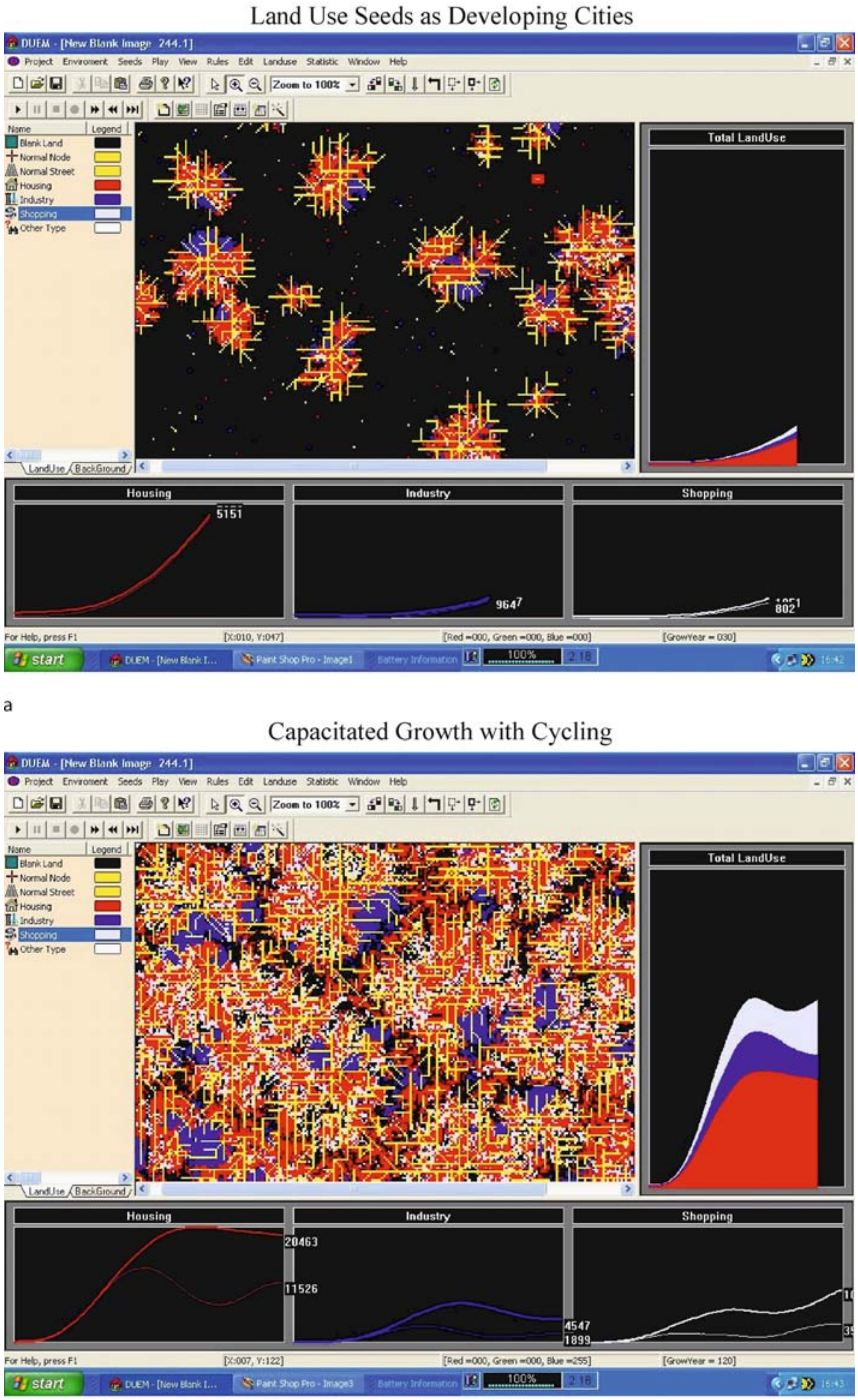
We are not able to present the fine details of the model here (see Batty, Xie and Sun [19], and Xie and Batty [79]) but we can provide a broad sketch. The way initiating land uses spawn new ones is structured according to rule-based equations akin to the thresholding implied in Eq. (53). In fact, there are three spatial scales at which these thresholds are applied ranging from the most local neighborhood through the district to the region itself. The neighborhood exercises a trigger for new growth or decline based on the existence or otherwise of the street network, the district uses the densities of related land uses and distance of the new land use from the initiating use to effect a change, while the region is used to implement hard and fast constraints on what cells are available or not for development. Typically an initiating land use will spawn a new land use in a district only if the cells in question are vacant and if they are not affected by some regional constraint on development with these rules being implemented first. The probability of this land use occurring in a cell in this district is then fixed according to its distance from the initiating location. This probability is then modified according to the density of different land uses that exist around each of these potential locations – using compatibility constraints – and then in the local neighborhood, the density of the street network is examined. If this density is not sufficient to support a new use, the probability is set equal to zero and the cell in question does not survive this process of allocation. At this point, the cell state is switched from 'empty' or 'vacant' to 'developed' if the random number drawn is consistent with the development probability determined through this process.

Declines in land use which are simply switches from developed to vacant in terms of cell state are produced through the life cycling of activities. When a mature land use in a cell reaches a certain age, it moves into a one period declining state and then disappears at the end of this time period, the cell becoming vacant. Cells remain vacant for one time period before entering the pool of eligible locations for new development. In the model as currently constituted, there is no internal migration of activities or indeed any mutation of uses but these processes are intrinsic to the model structure and have simply not been invoked. The software for this model has been written from scratch in *Visual C++* with the loosest coupling possible to GIS through the import of raster files in different proprietary formats. The interface we have developed, shown below, enables the user to plant various land use seeds into a virgin landscape or an already developed system which is arranged on a suitably registered pixel grid which can be up to 3K x 3K or 9 million pixels in size. A map of this region forms the main window but there are also three related windows which show the various trajectories of how different land uses change through time with the map and trajectories successively updated in each run.

A feature which is largely due to the fact that the model can be run quickly through many time periods, is that the system soon grows to its upper limits with exponential growth at first which then becomes logistic or capacitated. In Fig. 6, we show how this occurs from planting a random selection of land use seeds in the region and then letting these evolve until the system fills. Because there are lags in the redevelopment of land uses in the model due to the life cycle effects, as the system fills, land is vacated. This increases the space available for new development leading to oscillations of the kind reflected in Forrester's model shown in Fig. 5 and more controversially in the real systems shown for New York City and Greater London in Fig. 4. In this sense, a CA model has a dynamics which is equivalent to that of the more top-down dynamics where growth is modeled by exponential or logistic functions. CA models however generate this as an emergent phenomena from the bottom up.

Our last demonstration of CA really does generate emergent phenomena. This is a model of residential movement that leads to extreme segregation of a population classified into two distinct groups which we will call red R and green G. Let us array the population on a square grid of dimension 51 x 51 where we place an R person next to a G person in alternate fashion, arranging them in checker board style as in Fig. 7a. The rule for being satisfied with one's locational position viz a viz one's relationship to other individuals is as follows: persons of a different group will live quite happily, side by side with each other, as long as there are as many persons of the same persuasion in their local neighborhood. The neighborhood in this instance is the eight cells that surround a person on the checkerboard in the *n, s, e, w,* and *nw, se, sw,* and *ne* positions. If however a person finds that the persons of the opposing group outnumber those of their own group, and this would occur if there were more than 4 persons of the opposite persuasion, then the person in question would change their allegiance. In other words, they would switch their support to restore their own equilibrium which ensures that they are surrounded by at least the same number of their own group. There is a version of this model that is a little more realistic in which a person would seek another location – move – if this condition were not satisfied rather than change their support, but this is clearly not possible in the completely filled system that we have assumed; we will return to this slightly more realistic model below.

In Fig. 7a, the alternative positioning shown in the checker board pattern meets this rule and the locational pattern is in 'equilibrium': that is, no one wants to change their support to another group. However let us suppose that just six persons out of a total of 2601 (51 x 51 agents sitting on the checker board) who compose about 0.01 percent of the two populations, change their allegiance. These six changes are easy to see in Fig. 7a where we assume that four R persons of the red group, change in their allegiance to support the green group, and two Gs change the op-



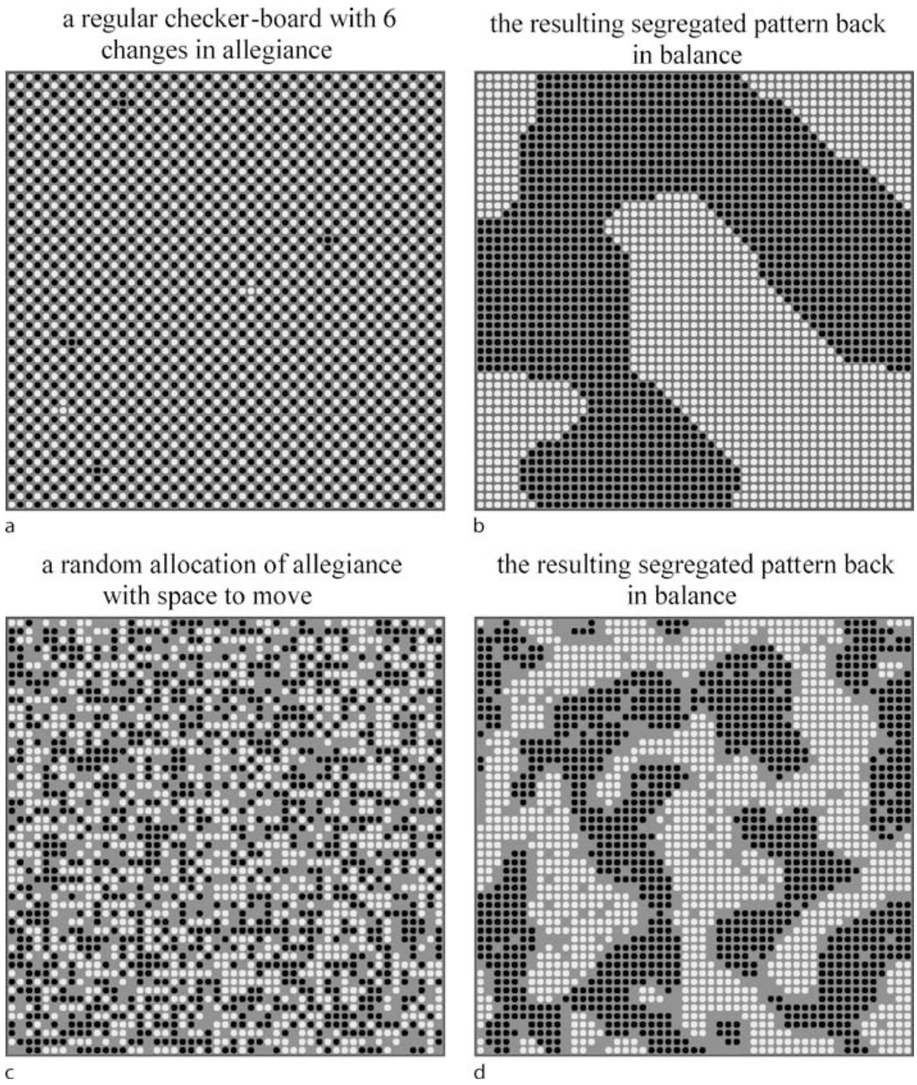
Land Use Seeds as Developing Cities

a

Capacitated Growth with Cycling

b

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 6
Cellular growth using the *DUEM* model


a regular checker-board with 6 changes in allegiance
the resulting segregated pattern back in balance
a
b
a random allocation of allegiance with space to move
the resulting segregated pattern back in balance
c
d

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 7
Emergent segregation: A fragile equality (a) gives way to segregation (b); A random mix with available space (c) gives way to segregation (d)

posite way. What then happens is the equilibrium is upset in these locations but instead of being quickly restored by local changes, this sets off a mighty unraveling which quickly changes the locational complexion of the system to one where the Rs are completely and utterly segregated from the Gs. We show this in Fig. 7b. From a situation where everyone was satisfied and mixed completely, we get dramatic segregation which is a most unusual consequence. At first sight, one would never imagine that with so mild a balance of preferences, such segregation would take place. The ultimate pattern implies that Rs will live nowhere near Gs unless they really have to and there is nowhere else to live and vice versa. If an R or a G could not tolerate more than one person of a different kind living near them, then such segregation would be understandable but this is not the case: Rs are quite content to live in harmony with Gs as long as the harmony is equality.

This model was first proposed more than 30 years by Schelling [61,62]. In fact we can make this a little more realistic if we provide some free space within the system. In this case, we assume that 1/3 of the lattice is empty of persons of any kind, 1/3 composed of Rs, and 1/3 of Gs, and we mix these randomly as we show in Fig. 7c. Now the rule is slightly different in that if there are more opposition persons around a person of one persuasion, then that person will try to move his or her location to a more preferential position. This sets up a process of shuffling around the checker board but as we show in Fig. 7d, quite dramatic shifts take place in location which leads to the segregation shown. This is the kind of effect that takes place in residential areas in large cities where people wish to surround themselves with neighbors of their own kind. What is surprising about the phenomena which makes it 'emergent' is that for very mild preferential bias, dramatic segregation can take place. Of course if the preferences for like neighbors are very strong anyway, then segregation will take place. But in reality, such preferences are usually mild rather than strong, yet extreme segregation takes place anyway. The conclusion is that cities often look more segregated around racial and social lines than the attitudes of their residents might suggest.

Comprehensive System Models of Urban Structure

Integrated Land Use Transport Models

The various components used to model cities in equilibrium were quickly assembled into structures that attempted to simulate urban structure and growth from the 1960s onwards. These models were referred to as land use transport models in that their aim was to simulate the locations of different land uses and their consequent patterns of traffic generation, usually according to spatial interaction principles based on gravitational assumptions. But they usually represented cities as demographic and economic activities – population, households, employment and so on – rather than as residential, commercial or industrial land use. In short the city system was seen to operate at the level of the location of activities which then consumed space through land use from which traffic was generated, and once urban activities and their interactions were predicted, appropriate translations were made into land use. As we shall see, this is not as unproblematic as was originally thought.

The integration of urban activities and their interactions – land use and transport – can be accomplished using a variety of economic frameworks built around economic relationships between activities. Traditionally these have been represented as input-output models where one activity is linked to another and it is possible to predict the chain of linkages between all the activities using multipliers. We will illustrate this for two activities: we assume that employment E is divided into an unpredictable component, sometimes considered as employment that is *basic* B and export orientated in the economy, and employment that is *non-basic* S where $E = B + S$. Non-basic employment services the population P from which it is derived as $S = bP$. If we then consider that population can be generated by applying an activity rate a to employment as $P = aE$, we have the rudiments of a generative sequence that forms a structure for predicting activities and their locations which are highly interdependent. Simple manipulation of these relationships shows that $E = B(1 - ba)^{-1}$ where $(1 - ba)^{-1}$ is the multiplier central to traditional macro-economic theory.

If we now consider that employment and population are related spatially through their interactions, we model the relationship between employment as population using a singly-constrained sub-model

$$P_j = a \sum_i T_{ij} = a \sum_i E_i \frac{F_j c_{ij}^{-\psi}}{\sum_k F_k c_{ik}^{-\psi}} , \qquad (54)$$

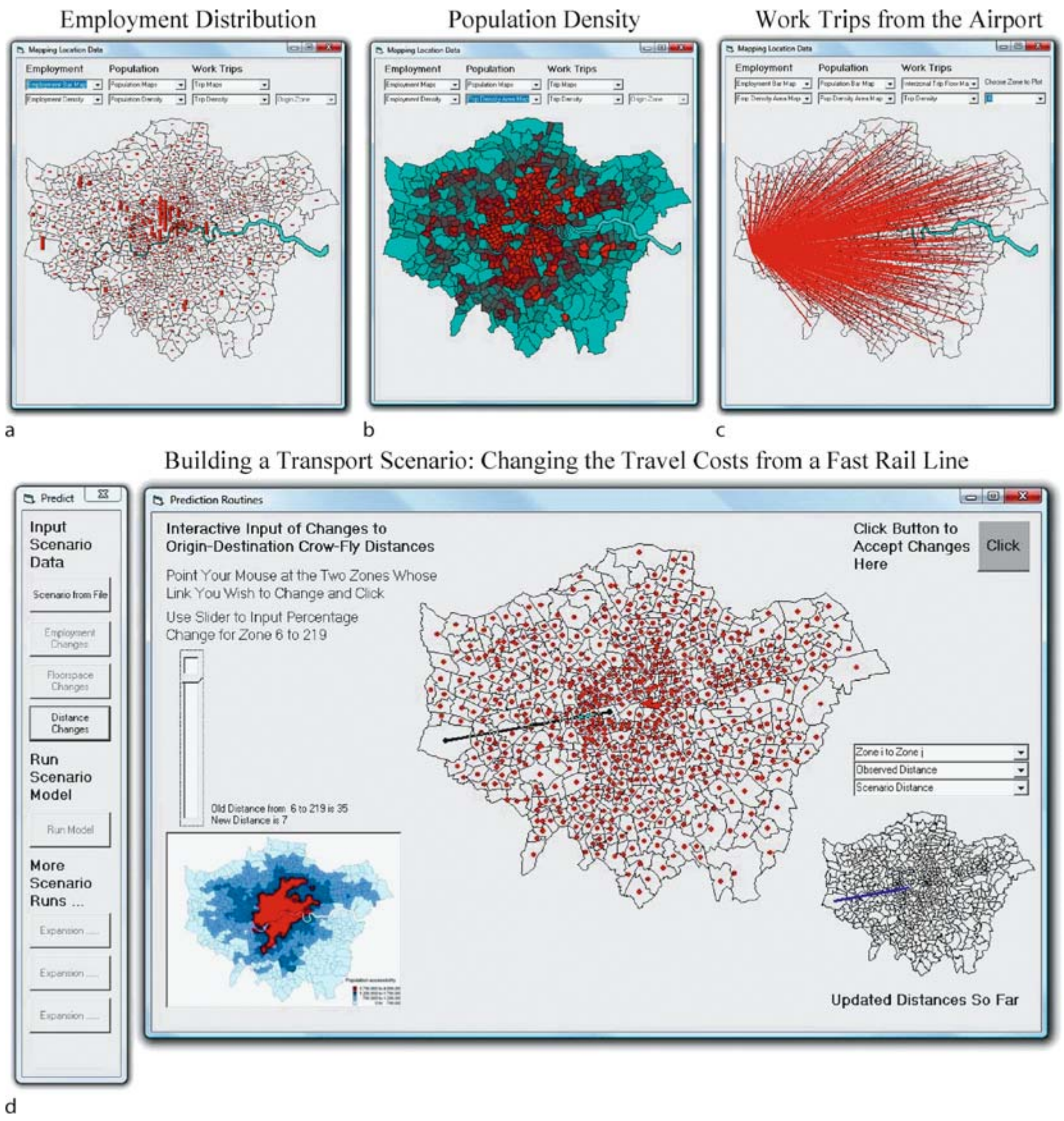
where T_{ij} are work trips between i and j, F_j is some measure of attraction at residential location j, and ψ is the friction of distance/travel cost parameter. Employment is modeled in reverse direction as

$$E_i = b \sum_j S_{ji} = b \sum_j P_j \frac{F_i c_{ji}^{-\varpi}}{\sum_k F_k c_{ki}^{-\varpi}} , \qquad (55)$$

where S_{ji} are employment demands in j from i, F_i is some measure of attraction at residential location i, and ϖ is the friction of distance/travel cost parameter. These two equations for the two sectors are not usually solved simultaneously but the chain is broken in that we start with basic employment B in Eq. (54), predicting basic population, then using this basic population in Eq. (55) to produce an increment of non-basic employment which in turn is used to predict the next increment of non-basic population in Eq. (54). This iteration converges to the multiplier relationships $E = B(1 - ba)^{-1}$ and $P = bB(1 - ba)^{-1}$.

This kind of sequence can be disaggregated indefinitely with respect to population and employment types and linked demands to other sectors. Education, leisure and so on can be added to the framework making the model ever more comprehensive. This was the model first developed by Lowry [50]. It is still the most widely applied of all operational urban models and has been elaborated in various ways, some of them dealing with partial dynamics [11]. Their theoretical pedigree is rooted largely in regional economics, location theory and the new urban economics which represent the spatial equivalents of classical macro and micro economics. The most coherent recent statement in this vein is based on applications of trade theory to the urban economy as reflected in the work of Fujita, Krugman and Venables [40] but there is a long heritage of empirical models in the Lowry [50] tradition which continue to be built [71].

These models now incorporate the four-stage transportation modeling process of trip generation, distribution, modal split and assignment explicitly and they are consistent with discrete choice methods based on utility maximizing in their simulation of trip-making [20]. They have been slowly adapted to simulate dynamic change although they still tend to generate the entire activity pattern of the city in one go, and they remain parsimonious in that the assumption is that all the outcomes from the model can be tested in terms of their goodness of fit. They have also become more disaggregate and there are now links to physical land use although they still remain at the level of activity allocation despite their nomenclature as land use transport models. In short, this class of models is the most operational in that newer styles tend to be less comprehensive in their treatment of urban activities and transportation. Probably the most highly developed of these models currently is the UrbanSim model [69] although the MEPLAN, TRANUS and IRPUD models, whose most recent versions were developed in the EU Propolis [60] project, also represent the state-of-the-art.



Interactive Input of Changes to Origin-Destination Crow-Fly Distances

Point Your Mouse at the Two Zones Whose Link You Wish to Change and Click

Use Slider to Input Percentage Change for Zone 6 to 219

Old Distance from 6 to 219 is 35
New Distance is 7

Click Button to Accept Changes Here

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 8
Visualization of outputs from a Greater London land use transport model

To conclude this section, it is worth showing a visualization from one of these land use transport models which we have recently built for the London region as part of an integrated assessment of climate change in the metropolis. The component we show is a residential location model which predicts the flow of workers from employment locations to residential areas using four different modes of transport and disaggregated into five employment and five household types. In Fig. 8a, we show some outputs from the model – the observed employment distribution, the pattern of population density, and total work trips from the airport (Heathrow) zone in the base year simulation 2005. This kind of model assumes that employment and the travel cost network are exogenously determined and thus 'what-if' style questions can be thrown at the model to be evaluated in terms of the impact of changes in the transport network and employment volumes on the location of population. We illustrate such a scenario builder for changes in the transport routes and costs in Fig. 8b which provides some sense of how such complexity can be visualized. These are key issues in planning policy for the future growth of London, particularly with respect to flooding in the Thames Estuary which is likely to be affected by climate change. These kinds of models are hardly routine but they are being developed now in many places.

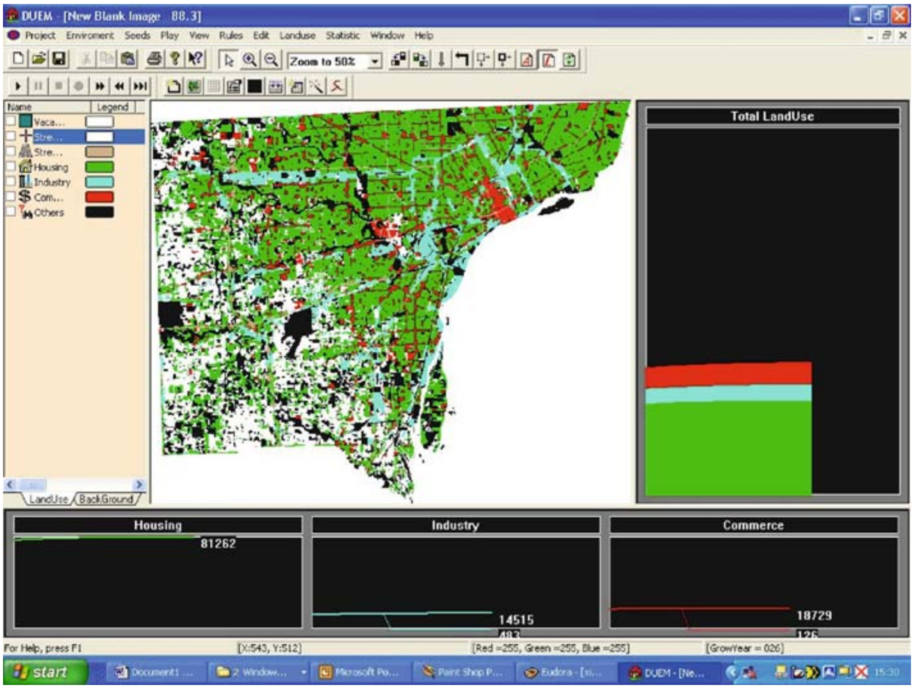
Agent-Based and Cellular Automata Models of Land Development

The first bottom-up CA models applicable to urban structure and growth can be traced back to the 1960s. Chapin and Weiss [27] used cell-space (CS) simulation whose locational attractions were based on linear regression, in their models of urban growth in Greensboro, North Carolina. Lathrop and Hamburg [49] used gravitational models to effect the same in simulating growth in the Buffalo-Niagara region while from a rather different perspective, Tobler [65] used CA-like simulation to generate a movie of growth in the Detroit region. All these applications were on the edge of the mainstream which 30 years ago was based not on formal dynamics but on cross-sectional equilibrium models of the variety presented above. In the intervening years, CA insofar as it was considered a simulation tool, was regarded as important mainly for its pedagogic and analytical value [32]. It was not until the early 1990s that models began to emerge which were considered to be close enough to actual urban growth patterns to form the basis for simulation and prediction. In fact, there still exists a recurrent debate about whether or not CA models are more important for their pedagogic value rather than for their abilities to simulate real systems. These require gross simplifications of model processes and spatial units, sometimes rendering them further from reality than the static cross-sectional models that came before.

The three earliest attempts at such modeling were geared to simulating rapid urban growth for metropolitan regions, medium-sized towns, and suburban areas. Batty and Xie [18] developed simulations of suburban residential sprawl in Amherst, New York, where a detailed space-time series of development was used to tune the model. Clarke and Gaydos [31] embarked on a series of simulations of large-scale metropolitan urban growth in the Bay Area and went on to model a series of cities in the US in the Gigalopolis project. White and Engelen [72] developed a CA model for Cincinnati from rather crude temporal land use data and in all these cases, the focus was on land development, suburbanization, and sprawl. Since then, several other groups have developed similar models focusing on suburbanization in Australian cities [70], 'desakota' – rapid urban growth in rural areas in China – specifically in the Pearl River Delta [80], diffused urban growth in Northern Italy [22], and rapid urbanization in Latin American cities [35]. Other attempts at modeling and predicting sprawl have been made by Papini et al. [58] for Rome and Cheng [28] for Wuhan, while Engelen's group at RIKS in the Netherlands has been responsible for many applications of their model system to various European cities [9].

There are at least four applications which do not focus on urban growth per se. Wu and Webster [77] have been intent on adding spatial economic processes and market clearing to such models, while Portugali and Benenson [59] in Tel-Aviv have focused their efforts on intra-urban change, particularly segregation and ghettoization. Semboloni [64] has worked on adding more classical mechanisms to his CA models reflecting scale and hierarchy as well as extending his simulations to the third dimension, while there have been several attempts by physicists to evolve a more general CA framework for urban development which links to new ideas in complexity such as self-organized criticality and power law scaling [7,63].

It is worth showing some graphics from such CA models as they are being applied to real cities. In Fig. 9, we show how the DUEM model can be used to simulate the pattern of development change in the Detroit region of South East Michigan. In a sense because we live in world dominated by a somewhat unhealthy interest in growth, it might be assumed that all the models we have presented here are only geared to simulating new development. In fact, each of these models can simulate decline or reproduce the steady state because CA models can solely deal with transitions and change in the existing fabric as we il-



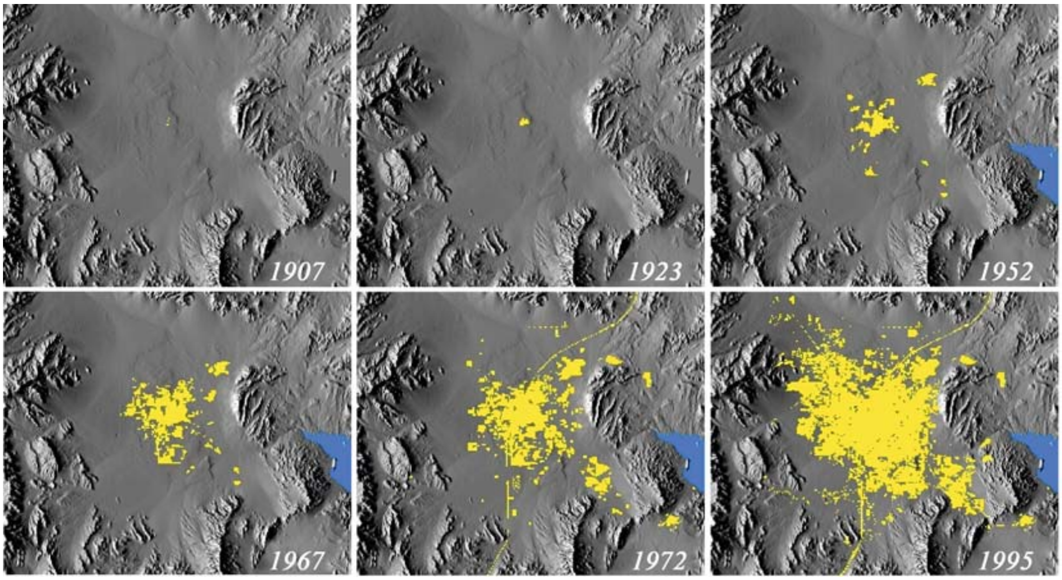
Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 9
Simulating very slow growth and rapid decline in the Detroit region using the CA *DUEM* model

lustrated earlier in the Schelling segregation model. This is the case in Detroit where the population has rapidly adjusted and segregated its locations in the last 50 years but in a context where the overall growth has been extremely modest with many areas growing very fast in the suburbs but the central areas declining at similar rates. The profile in Fig. 9 is akin to a steady state rather than the overall exponential growth or decline shown in previous examples.

There are some agent-based models at the land use or activities level which enable predictions of future urban patterns but the main focus is at the very micro-level where local movements in terms of traffic are being simulated [25]. Several models that approach the agent-based ideal originate from other areas. TRANSIMS is a hybrid in that its roots are in agent-based simulation of vehicles but it has been scaled to embrace urban activities [55] and even UrbanSim can be interpreted through the agent paradigm. A parallel but significant approach to individualistic modeling is based on micro-simulation which essentially samples individual behavior from more aggregate distributions and constructs synthetic agent-based models linked to spatial location [30]. This is a rapidly changing field at the present time with no agreement about terminology. The term agent is being used to describe many different types of models with some focusing on unique objects ranging from cells or points in space where activities or individuals exist to models of institutions and groups with only implicit spatial positioning [44].

Models of Urban Morphology

The models introduced above do not capture many of the physical features of cities and regions in terms of their morphology. Cities are highly organized with respect to their form, displaying as we have already seen in terms of city size, clusters of activity on all scales, in short, fractals [16]. Insofar as static equilibrium models are able to reproduce this form and to an extent they are able to do so, this is largely because some of the structure of the city is input into these models through existing employment and population distributions which have already captured elements of the morphology. There are competitive effects in these models too that are intrinsic to these simulations with the dynamic models based on cellular automata closest to reflecting these processes in urban form. This is because the process of development is generated from the bottom up and agglomeration is a key feature of the processes of development that are simulated as in some of the



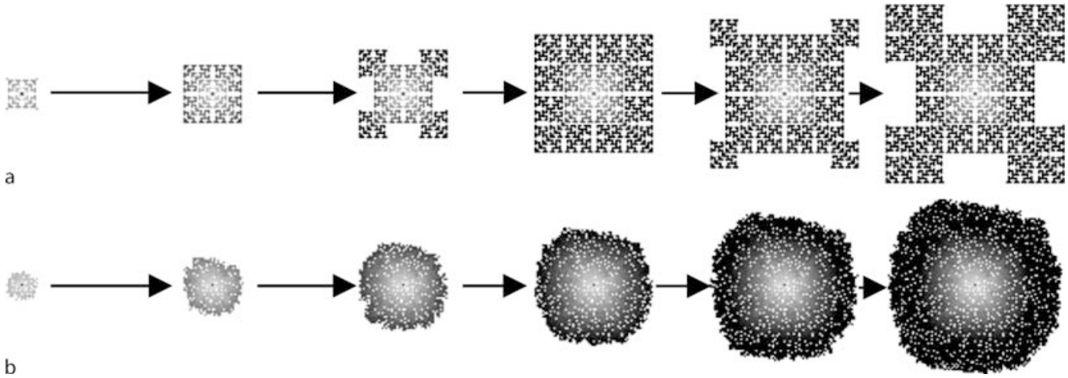
Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 10
The growth of Las Vegas from 1907 to 1995 (from [1])

models discussed in the last section. Here we will simply illustrate some of the evolving forms that various combinations of the models already discussed are able to simulate. This shows how various processes of land development and travel behavior can come together to generate structures that are close to what we observe in the real world.

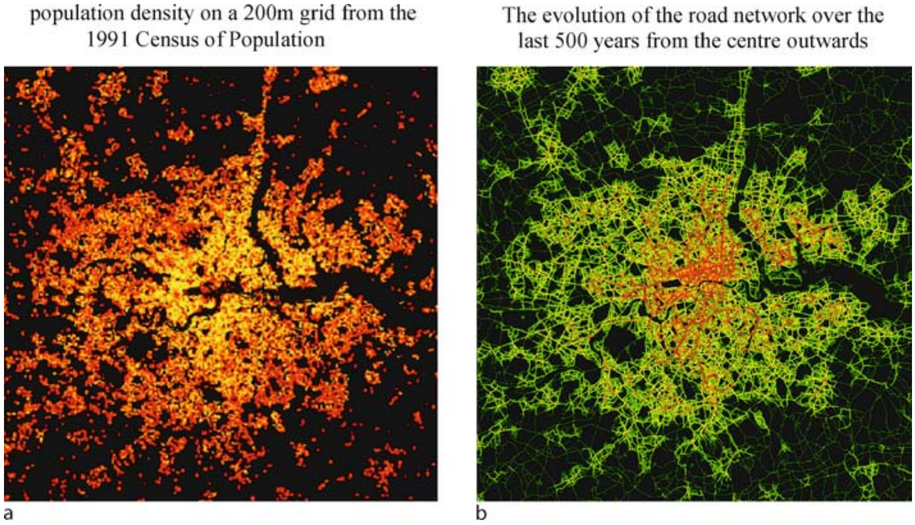
A good example of the urban growth which has been rapid over the last 50 years is Las Vegas, the fastest growing metropolitan area in the United States which is illustrated in Fig. 10 [1]. The sprawl does not look very different from time period to time period although it is clear that growth is clustered and these clusters tend to merge as the city grows. In this sense, the pattern always looks like more of the same from time period to time period but inside the city, things have changed rather more dramatically as the place has moved from desert oasis and staging post prior to 1950 to the entertainment and gambling capital of the US. Exponential growth of population, employment and tourism is implied by this volume of urban development mirroring the simplest 'un-capacitated' growth model in Eqs. (36) and (37). The fact that the city has grown in some directions rather than others is largely due to a combination of physical and accidental historical factors and does not imply any differences in the way growth has occurred from one time period to the next.

Cellular automata models can generate such growth where entirely local development rules are operated uniformly across the space to grow a city from a single seed. This can lead to fractal patterns, patterns that are self-similar in form with respect to scale, of the kind observed in real cities. In Fig. 11a, we show how the operation of deterministic rules where a cell is developed if there is one and only one cell already developed in its immediate neighborhood, leads to a growing structure. This is a typical example of a modular principle that preserves a certain level of density and space when development occurs but when operated routinely and exhaustively leads to cellular growth that is regular and self-similar across scales, hence fractal. In Fig. 11b, the shape of the structure generated is now circular in that development eventually occurs everywhere. The city fills up completely but the order in which this takes place is a result of development taking place at each time period with random probability. This is the effect of introducing 'noise' or 'diversity' into the model used to generate the sequence in Fig. 11a.

If urban growth is modular and scales in the simplistic way that is portrayed in these models of fractal growth, then it is not surprising that there is a tendency to explain such patterns generically, without regard to growth per se; to study these as if they represent systems with an equilibrium pattern that simply scales through time. But this is a trap that must be avoided. Dig below the surface, and examine the processes of growth and the activities that occupy these forms, disaggregate the scale and change the time interval, and this image of an implied stability changes quite radically. During the era pictured in



Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 11
Growth from the bottom up. a deterministic growth based on developing cells if one and only one cell is already developed in their 8 cell adjacent neighborhood, and b stochastic growth based on developing cell if any cell is developed in the adjacent neighborhood according to a random probability


population density on a 200m grid from the 1991 Census of Population
The evolution of the road network over the last 500 years from the centre outwards
a
b

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 12
Greater London: self-similar clusters and the connectivity network within the sprawl

Fig. 10, technology has changed dramatically. Las Vegas did not acquire its gambling functions until the 1950s but by then it was already growing fast and the subsequent injection of cash into its local economy, the largest per capita in the western world for those who reside there, did little to change the pattern of explosive growth that followed. The manner in which people moved in the early Las Vegas was by horse and wagon but the city could only grow with the car, the plane and air-conditioning, not to say the incredible information technologies that now dictate how one gambles, wins, and loses.

Our six frame 'movie' of the growth of Las Vegas does reveal that the established pattern of adding to the periphery is not entirely the complete story for small blobs of development seem to attach themselves and then are absorbed back into the growing mass as growth catches them up. In this case, this is simply housing being constructed a little beyond the edge due to the mechanics of the development process. In older, more established settlement patterns such as those in Western Europe for example, this might be the absorption of older villages and freestanding towns into the growing sprawl. Consider the picture of population density in London recorded in 1991 and illustrated in Fig. 12a. Here there are many towns and villages that existed long before London grew to embrace them. If we define the metropolis as the connected network of settlement that fills an entire space where everyone can connect to everybody else either directly or indirectly, the picture is similar as we show in Fig. 12b.

One could envisage London being connected in this way with a much sparser network of links while at the other extreme the entire space could be filled. In fact, it would seem that the level of connectivity which has evolved with respect to the density of the space filled is just enough for the city to function as a whole. It is this morphology and degree of connectivity that marks the fact that the city has reached a level of self-organization which is regarded as critical. If connectivity were greater, more space would be filled and many more connections put in place but the structure would contain a certain degree of redundancy making it inefficient. Below this, the system would not be connected at all and it would not function as a metropolis. In fact there are strong relationships to this characterization of urban settlement as a porous media in which a phase transition might take place as the system fills up which in network terms, is like a percolation threshold [13]. The models that we have sketched above all provide ways of generating these kinds of morphology, albeit through somewhat different mechanisms than the obvious way in which growth in physical systems takes place. The forms generated constitute an essential check on the adequacy or otherwise of these system models.

Future Directions

The biggest problems facing the development of complex systems models in general and those applied to cities in particular involve validation. The move from articulating systems as organized entities structured from the top down based on some sort of centralized control mechanisms to systems that grow in an uncoordinated way from the bottom up have also shifted our perspective from developing systems model in a parsimonious way to developing much richer models requiring more detailed data. In short, complexity theory has changed the basis for theory and model selection from an insistence that all models must be testable against data to an acceptance that if there is a strong reason why some non-testable propositions should be included in a model (as models with very rich behaviors and processes imply), then these should be included even if they cannot be tested. This is consistent with the shift from aggregate to disaggregate modeling, from the focus on equilibrium to dynamics, and on processes and behaviors rather than simply outcomes.

This changes the entire basis of validation and combined with the difficulties of articulating processes which are clearly relevant but often unobservable, the way in which models might be useful in policy making in complex systems is changing too. Modeling is now much more contingent on context and circumstance than at any time in the past. The use of multiple models, counter modeling and the synthesis of different and often contradictory model structures is now taken for granted in systems where we consider there may be no optimal solutions and where there will always be dissent from what is regarded as acceptable. Many newer models such as those based on cellular and agent-based structures and those which postulate a dynamics that involves bifurcations that are often of only theoretical interest until one such dynamic is observed, are unlikely to meet the canons of parsimony in which unambiguous tests can be made against data. These limits to validation begin to suggest that complex system models need to be classified on a continuum of ways in which they can be tested and used in practice which will depend on the type of model, the context, and the users involved [17].

In terms of more substantive developments, the question of dynamics is still of burning importance in developing better models of cities. There is an intrinsic problem of articulating urban processes of change from sparsely populated data bases which often contain only the aggregate outcomes of multiple processes. The way in which our commonsense observations of decision making in cities can be linked to more considered outcomes represented in data has barely been broached in developing good models of urban spatial behavior. In agent-based modeling, the role of cognition is important while the question of defining agents at appropriate levels is a major research focus, particularly when it comes to aggregates which are of a more abstract nature, such as groups and institutions. However what is of clear importance is the fact that as our focus becomes finer and as we disaggregate to ever more detailed levels, we then begin to represent policy processes into which these models might be nested in more detailed ways, implying that policy making and planning itself might be simply one other feature of these system models.

In short in our quest for more detail and for embracing a wider environment, city models have come to encapsulate the control mechanisms themselves as intrinsic to their functioning. It is at this point that we need much better ways of showing how such models can be used in practice. To an extent, this implies that we need to link these system models to their wider context of use and application, showing how other conceptions, other systems models, might be related to them in less formal ways than in terms of the science we have presented here. This has always been a challenge for the application of complexity theory to human and social systems, and it will remain the cutting edge of this field whose rationale is the prediction and design of more efficient, equitable, and sustainable cities.

Bibliography

Primary Literature

1. Acevedo W, Gaydos L, Tilley J, Mladinich C, Buchanan J, Blauer S, Kruger K, Schubert J (1997) Urban land use change in the Las Vegas valley. US Geological Survey, Washington
2. Albert R, Jeong H, Barabási A-L (1999) Diameter of the world wide web. Nature 401:130–131
3. Allen PM (1982) Evolution, modelling, and design in a complex world. Environ Plan B 9:95–111
4. Allen PM (1998) Cities and regions as self-organizing systems: models of complexity. Taylor and Francis, London
5. Alonso W (1964) Location and land use. Harvard University Press, Cambridge
6. Anas A (1983) Discrete choice theory, information theory and the multinomial logit and gravity models. Transp Res B 17B:13–23
7. Andersson C, Rasmussen S, White R (2002) Urban settlements transition. Environ Plan B 29:841–865
8. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
9. Barredo JI, Kasanko M, McCormick N, Lavalle C (2003) Modeling dynamic spatial processes: simulation of urban future scenarios through cellular automata. Landscape Urban Plan 64:145–160
10. Batty M (1974) Spatial entropy. Geograph Anal 6:1–31
11. Batty M (1976) Urban modelling: algorithms, calibration, predictions. Cambridge University Press, Cambridge
12. Batty M (2005) Agents, cells, and cities: new representational models for simulating multiscale urban dynamics. Environ Plan A 37(8):1373–1394
13. Batty M (2005) Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. MIT Press, Cambridge
14. Batty M (2006) Rank clocks. Nature 444:592–596
15. Batty M (2008) The size, scale, and shape of cities. Science 319(5864):769–771
16. Batty M, Longley PA (1994) Fractal cities: a geometry of form and function. Academic Press, San Diego
17. Batty M, Torrens PM (2005) Modelling and prediction in a complex world. Futures 37(7):745–766
18. Batty M, Xie Y (1994) From cells to cities. Environ Plan B 21:s31–s48
19. Batty M, Xie Y, Sun Z (1999) Modeling urban dynamics through GIS-based cellular automata. Comput Environ Urban Syst 23:205–233
20. Ben Akiva M, Lerman S (1985) Discrete choice analysis. MIT Press, Cambridge
21. Berry BJL (1964) Cities as systems within systems of cities. Papers Proc Region Sci Assoc 13:147–164
22. Besussi E, Cecchini A, Rinaldi E (1998) The diffused city of the italian north-east: identification of urban dynamics using cellular automata urban models. Comp Environ Urban Syst 22:497–523
23. Blank A, Solomon S (2000) Power laws in cities population, financial markets and internet sites: scaling and systems with a variable number of components. Physica A 287:279–288
24. Cardillo A, Scellato S, Latora V, Porta S (2006) Structural properties of planar graphs of urban street patterns. Phys Rev E 73:066107-1–8
25. Castle CJE, Crooks AT (2006) Principles and concepts of agent-based modelling for developing geospatial simulations. Working Paper 110. Centre for Advanced Spatial Analysis, University College London, London
26. Chadwick GF (1971) A systems view of planning. Pergamon Press, Oxford
27. Chapin FS, Weiss SF (1968) A probabilistic model for residential growth. Transp Res 2:375–390
28. Cheng J (2003) Modelling spatial and temporal urban growth. Ph D Thesis, ITC Dissertation 99, ITC, Enschede, Netherlands
29. Clark C (1951) Urban population densities. J Royal Stat Soc Ser A 114:490–496
30. Clarke G (ed) (1996) Microsimulation for urban and regional policy analysis. Pion Press, London
31. Clarke KC, Gaydos LJ (1998) Loose coupling a cellular automaton model and GIS: long-term growth prediction for San Francisco and Washington/Baltimore. Int J Geograph Inform Sci 12:699–714
32. Couclelis H (1985) Cellular worlds: a framework for modeling micro-macro dynamics. Environ Plan A 17:585–596
33. Crucitti P, Latora V, Porta S (2006) Centrality measures in spatial networks of urban streets. Phys Rev E 73:036125-1-5
34. Curry L (1964) The random spatial economy: an exploration in settlement theory. Ann Assoc Amer Geograph 54:138–146
35. de Almeida CM, Batty M, Câmara G, Cerqueira GC, Monteiro AMV, Pennachin CP, Soares-Filho BS (2003) Stochastic cellular automata modeling of urban land use dynamics: empirical development and estimation. Comput Environ Urban Syst 27:481–509
36. Dendrinos DS, Mullally H (1985) Urban evolution: studies in the mathematical ecology of cities. Oxford University Press, Oxford
37. Epstein JM, Axtell RL (1996) Growing artificial societies: social science from the bottom up. MIT Press, Cambridge
38. Feigenbaum MJ (1980) The metric universal properties of period doubling bifurcations and the spectrum for a route to turbulence. Ann New York Acad Sci 357:330–336
39. Forrester JW (1969) Urban dynamics. MIT Press, Cambridge
40. Fujita M, Krugman P, Venables AJ (1999) The spatial economy: cities, regions, and international trade. MIT Press, Cambridge
41. Gabaix X (1999) Zipf's law for cities: an explanation. Quart J Econom 114:739–767
42. Gell-Man M (1994) The quark and the jaguar: adventures in the simple and the complex. Freeman and Company, New York
43. Gibrat R (1931) Les inégalités économiques. Librarie du Recueil Sirey, Paris
44. Gilbert N (2007) Agent-based models. Sage Inc., Thousand Oaks
45. Haag G (1989) Dynamic decision theory: applications to urban and regional topics. Kluwer, Dordrecht
46. Helbing D, Nagel K (2004) The physics of traffic and regional development. Contemp Phys 45:405–426
47. Hillier B (1996) Space is the machine. Cambridge University Press, Cambridge
48. Jiang B (2007) A topological pattern of urban street networks: universality and peculiarity. Physica A 384:647–655

49. Lathrop GT, Hamburg JR (1965) An opportunity-accessibility model for allocating regional growth. J Amer Inst Plan 31:95–103
50. Lowry IS (1964) Model of metropolis. Memorandum RM-4035-RC. Rand Corporation, Santa Monica
51. Mandelbot BB (1983) The fractal geometry of nature. Freeman, New York
52. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
53. McLoughlin JB (1969) Urban and regional planning: a systems approach. Faber and Faber, London
54. Miller JH, Page SE (2007) Complex adaptive systems: an introduction to computational models of social life. Princeton University Press, Princeton
55. Nagel K, Beckman RJ, Barrett CL (1999) TRANSIMS for urban planning. LA-UR 984389. Los Alamos National Laboratory, Los Alamos
56. Newman M, Barabási A-L, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton
57. Nijkamp P, Reggiani A (1992) Interaction, evolution and chaos in space. Springer, Berlin
58. Papini L, Rabino GA, Colonna A, Di Stefano V, Lombardo S (1998) Learning cellular automata in a real world: the case study of the rome metropolitan area. In: Bandini S, Serra R, Suggi Liverani F (eds) Cellular automata: research towards industry: ACRI'96. Proc of the 3rd Conference on cellular automata for research and industry. Springer, London, pp 165–183
59. Portugali J, Benenson I (1996) Human agents between local and global forces in a self-organizing city. In: Schweitzer F (ed) Self-organization of complex structures: from individual to collective dynamics. Gordon and Breach, London, pp 537–545
60. Propolis (2004) PROPOLIS (policies and research of policies for land use and transport for increasing urban sustainability). Final report for the Commission of the European Communities. LT Consultants Ltd, Helsinki
61. Schelling TC (1969) Models of segregation. Amer Econom Rev Papers Proc 58(2):488–493
62. Schelling TC (1978) Micromotives and macrobehavior. Norton and Company, New York
63. Schweitzer F, Steinbrink J (1997) urban cluster growth: analysis and computer simulation of urban aggregations. In: Schweitzer F (ed) Self-organization of complex structures: from individual to collective dynamics. Gordon and Breach, London, pp 501–518
64. Semboloni F (2000) The growth of an urban cluster into a dynamic self-modifying spatial pattern. Environ Plan B 27:549–564
65. Tobler WR (1970) A computer movie simulating population growth in the detroit region. Econom Geograph 42:234–240
66. Tribus M (1969) Rational, descriptions, decisions and designs. Pergamon Press, New York
67. von Bertalanffy L (1969) General system theory: foundations, development, applications. George Braziller, New York. Revised Edition (1976)
68. von Thünen JH (1826) Von Thunen's isolated state. Pergamon, Oxford. (1966 translation from the 1826 German Edition Der isolierte Staat in Beziehung auf Landwirtschaft und Nationaloekonomie by P. G. Hall)
69. Waddell P (2002) UrbanSim: modeling urban development for land use, transportation and environmental planning. J Amer Plan Assoc 68:297–314
70. Ward DP, Murray AT, Phinn SR (2000) A stochastically constrained cellular model of urban growth. Comput Environ Urban Syst 24:539–558
71. Wegener M (2005) Urban land-use transportation models. In: Maguire DJ, Batty M, Goodchild MF (eds) GIS, spatial analysis, and modeling. ESRI Press, Redlands, pp 203–220
72. White RW, Engelen G (1993) Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land use patterns. Environ Plan A 25:1175–1193
73. Wiener N (1965) Cybernetics: or the control and communication in the animal and the machine, 2nd edn. MIT Press, Cambridge
74. Wilson AG (1970) Entropy in urban and regional modelling. Pion Press, London
75. Wilson AG (1981) Catastrophe theory and bifurcation; applications to urban and regional systems. University of California Press, Berkeley
76. Wilson AG (2007) Boltzmann, Lotka, and Volterra and spatial structural evolution: a integrated methodology for some dynamical systems. J Royal Soc Interface 1–7. doi:10.1098/rsif.2007.1288
77. Wu F, Webster CJ (1998) Simulation of land development through the integration of cellular automata and multicriteria evaluation. Environ Plan B 25:103–126
78. Xie Y (1994) Analytical models and algorithms for cellular urban dynamics. Unpublished Ph D dissertation, State University of New York at Buffalo, Buffalo
79. Xie Y, Batty M (2005) Integrated urban evolutionary modeling. In: Atkinson PM, Foody GM, Darby SE, Wu F (eds) Geodynamics. CRC Press, Boca Raton, pp 273–293
80. Yeh A G-O, Li X (2000) A 'Grey-Cell' constrained ca model for the simulation of urban forms and developments in the planning of sustainable cities using GIS. Centre of Urban Planning and Environmental Management. University of Hong Kong, Pokfulam
81. Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Cambridge

Books and Reviews

Barabási A (2002) Linked: the new science of networks. Perseus Publishing, New York

Batty M, Couclelis H, Eichen M (1997) Urban systems as cellular automata. Environ Plan B 24:159–164

Benenson I, Torrens PM (2004) Geosimulation: automata-based modeling of urban phenomena. Wiley, London

Clarke M, Wilson AG (1993) Dynamics of urban spatial structure: progress and problems. J Region Sci 21:1–18

Couclelis H (1997) From cellular automata models to urban models: new principles for model development and implementation. Environ Plan B 24:165–174

Dendrinos DS, Sonis M (1990) Chaos and socio-spatial dynamics. Springer, New York

Haggett P, Chorley R (1969) Network analysis in geography. Edward Arnold, London

Helbing D, Molnar P Farkas IJ, Bolay K (2001) Self-organizing pedestrian movement. Environ Plan B 28:361–383

Holland JH (1995) Hidden order: how adaptation builds complexity. Addison-Wesley, Reading

Isard W (1956) Location and space-economy: a general theory relating to industrial location, market areas, land use, trade and urban structure. MIT Press, Cambridge

Jacobs J (1961) The death and life of great american cities. Vintage Books, Random House

Krugman PR (1996) The self-organizing economy. Blackwell, Cambridge

Portugali J (2000) Self-organization and the city. Springer, Berlin

Resnick M (1994) Termites, turtles and traffic jams: explorations in massively parallel micro-worlds. MIT Press, Cambridge

Sanders L, Pumain D, Mathian H, Guerin-Pace F, Bura S (1997) SIMPOP: a multiagent system for the study of urbanism. Environ Plan B 24:287–305

Simon HA (1969, 1996) Sciences of the artificial. MIT Press, Cambridge

Stewart JQ, Warntz W (1958) Physics of population distribution. J Region Sci 1:99–123

White RW (1998) Cities and cellular automata. Discrete Dyn Nature Soc 2:111–125

Willumsen LG, de Ortuzar JD (1990) Modelling transport. Wiley, Chichester

Wilson AG (2000) Complex spatial systems: the modelling foundations of urban and regional analysis. Pearson Education, Harlow

Wilson AG (1974) Urban and regional models in geography and planning. Wiley, Chichester

Climate Change and Agriculture

Cynthia Rosenzweig
NASA/Goddard Institute for Space Studies,
Columbia University, New York, USA

Article Outline

Glossary

Anthropogenic emissions Greenhouse gas emissions that are produced as a result of humans through such developments as industry or agriculture.

Greenhouse gases The gases of the atmosphere that create the greenhouse effect, which keeps much of the heat from the sun from radiating back into outer space. Greenhouse gases include, in order of relative abundance: Water vapor, carbon dioxide, methane, nitrous oxide, ozone, and CFCs. Greenhouse gases come from natural sources and human activity; present CO_2 levels are $\sim$380 ppmv, approximately 100 ppmv higher than they were in pre-industrial times.

Soybean cyst nematode *Heterodera glycines*, a plant-parasite that infects the roots of soybean, with the female becoming a cyst. Infection causes various symptoms, including a serious loss of yield.

El Niño-southern oscillation (ENSO) A phenomenon in the equatorial Pacific Ocean characterized by a positive sea-surface temperature departure from normal (for the 1971–2000 base period) in the Niño 3.4 region greater than or equal in magnitude to 0.5°C, averaged over three consecutive months.

North atlantic Oscillation (NAO) A hemispheric, meridional oscillation in atmospheric mass with centers of action near Iceland and over the subtropical Atlantic.

Vegetative index A simple numerical indicator used to analyze remote sensing measurements, often from space satellites, to determine how much photosynthesis is occurring in an area.

Soil organic carbon All the organic compounds within the soil without living roots and animals.

Introduction

The term climate change refers to an overall shift of mean climate conditions in a given region. The warming trend associated with anthropogenic emissions of greenhouse gases and the enhanced greenhouse effect of the atmosphere can and should be regarded as a "climate change" when viewed on the time scale of decades or a few centuries.

Climate change exacerbates concerns about agricultural production and food security worldwide. At global and regional scales, food security is prominent among the human concerns and ecosystem services under threat from dangerous anthropogenic interference in the earth's climate [17,29,50]. At the national scale, decision-makers are concerned about potential damages that may arise in coming decades from climate change impacts, since these are likely to affect domestic and international policies, trading patterns, resource use, regional planning, and human welfare.

While agro-climatic conditions, land resources and their management are key components of food production, both supply and demand are also critically affected by distinct socio-economic pressures, including current and projected trends in population and income growth and distribution, as well as availability and access to technology and development. In the last three decades, for instance, average daily per capita intake has risen globally from 2,400 to 2,800 calories, spurred by economic growth, improved production systems, international trade, and globalization of food markets [12]. Feedbacks of such growth patterns on cultures, personal tastes, and lifestyles have in turn led to major dietary changes – mainly in developing countries – where shares of meat, fat, and sugar in total food intake have increased significantly [13]. Thus, the consequences of climate change on world food demand and supply will depend on many interactive dynamic processes.

Agriculture plays two fundamental roles in human-driven climate change. It is one of the key human sectors that will be affected by climate change over the coming decades, thus requiring adaptation measures. Agriculture is also a major source of greenhouse gases to the atmosphere, including carbon dioxide (CO_2) due to land-use change and farm operations; methane (CH_4) from rice production and livestock husbandry, and nitrous oxide (N_2O) from application of nitrogen fertilizer. As climate changes as well as socio-economic pressures shape future demands for food, fiber and energy, synergies can be identified between adaptation and mitigation strategies, so that robust options that meet both climate and societal challenges can be developed. Ultimately, farmers and others in the agricultural sector will be faced with the dual task of contributing to global reductions of carbon dioxide and other greenhouse gas emissions, while coping with an already-changing climate.

A changing climate due to increasing anthropogenic emissions of greenhouse gases will affect both the productivity and geographic distribution of crop and pasture species. The major climate factors contributing to these responses include increasing atmospheric carbon dioxide, rising temperature, and increasing extreme events, especially droughts and floods. These factors in turn will affect water resources for agriculture, grazing lands, livestock, and associated agricultural pests. Effects will vary, depending on the degree of change in temperature and precipitation and on the particular management system and its location. Several studies have suggested that recent warming trends in some regions may have already had discernible effects on some agricultural systems [17].

Climate change projections are uncertain in regard to both the rate and magnitude of temperature and precipitation variation in the coming decades. This uncertainty arises from a lack of precise knowledge of how climate system processes will change and of how population growth, economic and technological developments, and land-use patterns will evolve in the coming century [16,17]. Despite these uncertainties, the ultimate significance of the climate change issue is related to its global reach, affecting agricultural regions throughout the world in complex ways. After approximately two decades of research, ten major conclusions may be drawn in regard to climate change and agriculture.

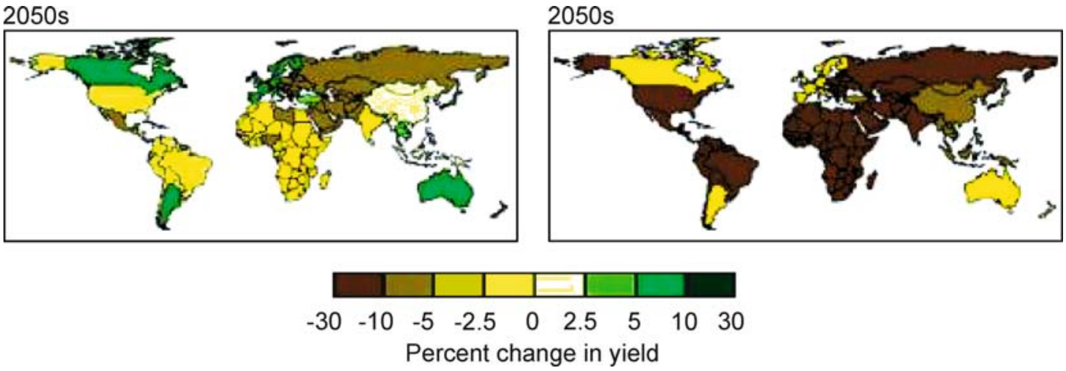
Effects on Agricultural Systems Will Be Heterogeneous

Global studies on projected climate change effects on agriculture show that negative and positive effects will occur both within countries and across the world. In large countries such as the United States, Russia, Brazil, and Australia, agricultural regions will likely be affected quite differently. Some regions will experience increases in production and some declines (see, e. g., [34]). At the international level, this implies possible shifts in comparative advantage for production of export crops. This also implies that adaptive responses to climate change will necessarily be complex and varied. Due to differences in global climate model projections and decadal variability, it is impossible to project exact effects in any one location for any given decade.

Developing Countries Are More Vulnerable

Despite general uncertainties about the rate and magnitude of climate change and especially about consequent hydrological changes, regional and global studies have consistently shown that agricultural production systems in the mid and high latitudes are more likely to benefit in the near term (to mid-century), while production systems in the low-latitudes are more likely to decline (Fig. 1) [17]. In biophysical terms, rising temperatures will likely push many crops beyond their limits of optimal growth and yield. Higher temperatures will intensify the evaporative demand of the atmosphere, leading to greater water stress, especially in semi-arid regions. Since most developing countries are located in lower-latitude regions (some which are indeed semi-arid) while most developed countries are located in the more humid mid- to high latitudes, this finding suggests a divergence in vulnerability between these groups of nations, with far-reaching implications for future world food security [31,37].

Furthermore, developing countries often have fewer resources with which to devise appropriate adaptation measures to meet changing agricultural conditions. The combination of potentially greater climate stresses and lower adaptive capacity in developing countries creates different degrees of vulnerability between rich and poor nations as they confront global warming. This difference is due in part to the potentially greater detrimental impacts of a changing climate in areas that are already warm (particularly if such areas are also dry), and in part to the generally lower levels of adaptive capacity in developing countries.



Climate Change and Agriculture, Figure 1

Potential changes (%) in national cereal yields for the 2050s (compared with 1990) under the HadCM3 SRES A1FI with (*left*) and without (*right*) CO_2 effects [31]

Development Path Matters

Since climate is not the only driving force on agriculture, researchers now conduct scenario analysis that include linked sets of population projections, economic growth rates, energy technology improvements, land-use changes, and associated emissions of greenhouse gases [31]. Regional patterns related to economic development and adaptive capacity contribute to differing levels of climate change impacts [17]. Scenarios with higher economic growth rates and less attention to environmental issues lead to high temperatures and reduced adaptive capacity, which in turn lead to pronounced decreases in yields both regionally and globally. Scenarios with lower greenhouse gas emissions and greater attention to environmental issues lead to lower amounts of temperature rise and crop production declines.

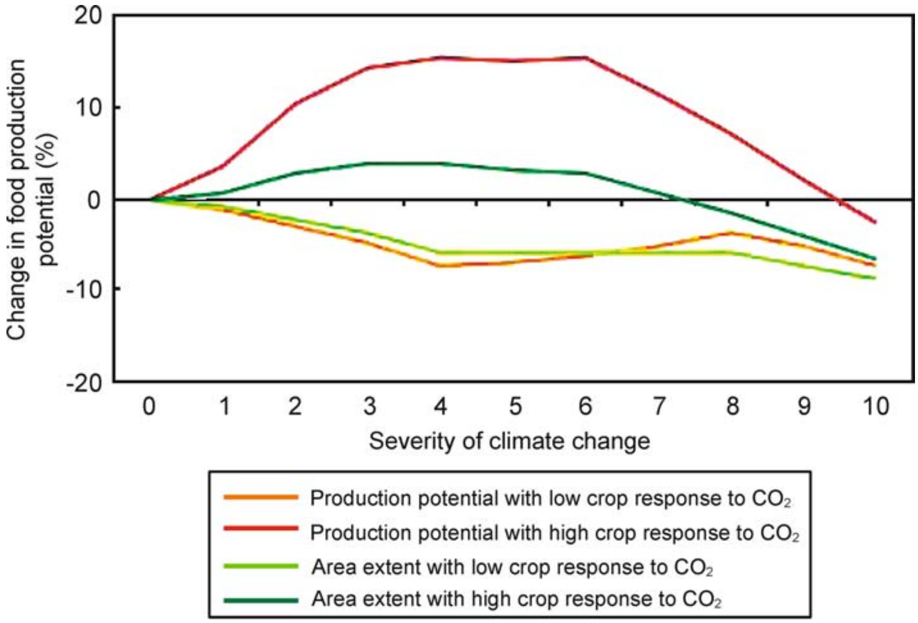
Long-Term Effects Are Negative for Both Developed and Developing Countries

If the effects of climate change are not abated, production in the mid- and high-latitudes is likely to decline in the longer term (i. e., above ~ 3°C warming) (Fig. 2) [16]. These results are consistent over a range of temperature, precipitation, and direct CO_2 effects tested, and are due primarily to the detrimental effects of heat and water stress as temperatures rise. While the beneficial effects of CO_2 may eventually level out, the detrimental effects of warmer temperatures and greater water stress are more likely to be progressive in all regions. Although the precise levels of CO_2 effects on crops and their contribution to global crop production are still active areas of research [47], global impacts are likely to turn negative in all regions sometime during the second half of the century, and perhaps before then for some major crops. For instance, by 2050 climate change is projected to have a downward pressure on yields of almonds, walnuts, avocados, and table grapes in California. Opportunities for expansion into cooler regions have been identified, but this adaptation would require substantial investments and may be limited by non-climatic constraints [24].

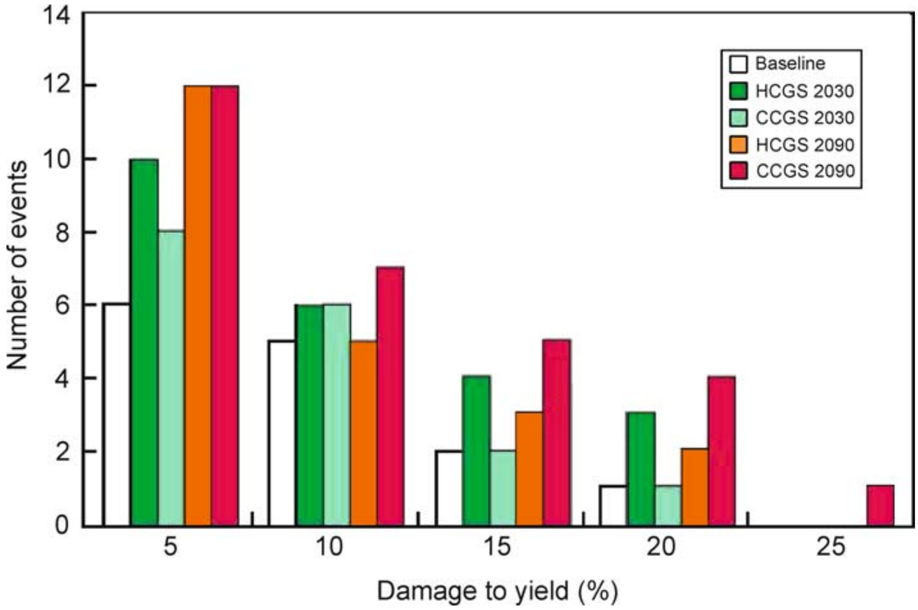
Water Resources Are Key

Recent flooding and heavy precipitation events in the US and worldwide have caused great damage to crop production. If the frequency of these weather extremes were to increase in the near future, as recent trends for the US indicate and as projected by global climate models [17,48], the cost of crop losses in the coming decades could rise dramatically. US corn production losses due to excess soil moisture, already significant under current climate, may double during the next thirty years, causing additional damages totaling an estimated $3 billion per year (Fig. 3) [41]. These costs may either be borne directly by those farmers affected or may need to be transferred to private or governmental insurance and disaster relief programs. There is also concern for tractability in the spring and water-logging in the summer in mid and high latitudes where precipitation is projected to increase.

Changes in crop water demand and water availability will affect the reliability of irrigation, which competes for growing municipal and industrial demands [42]. Studies link climate change scenarios with hydrologic, agricultural, and planning models to estimate water availability for agriculture under changing climate conditions, to explore changes in ecosystem services, and to evaluate adap-



Climate Change and Agriculture, Figure 2
Change in food production potential in relation to severity of climate change [13]



Climate Change and Agriculture, Figure 3
Number of events causing damage to maize yields due to excess soil moisture conditions, averaged over all study sites, under current baseline (1951–1998) and climate change conditions. The Hadley Center (HC) and Canadian Center (CC) scenarios with greenhouse gas and sulfate aerosols (GS) were used. Events causing a 20% simulated yield damage are comparable to the 1993 US Midwest floods [41]

tation strategies for the water resources and agriculture sectors. Major irrigated agricultural regions are very likely to be affected by changing supplies of and demands for water due a changing climate especially under conditions of expansion of irrigated lands [42].

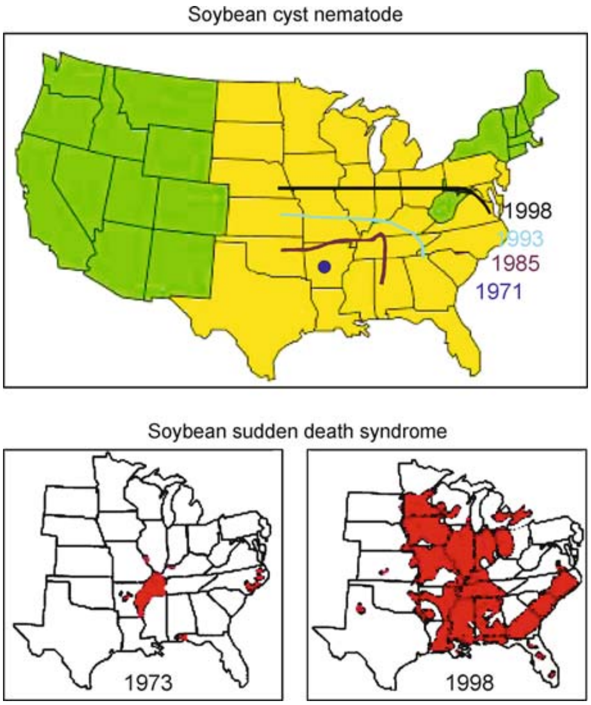
Cultivars are available for agricultural adaptation to the projected changes, but their demand for water may be higher than currently adapted varieties. Thus, even in relatively water-rich areas, changes in water demand due to climate change effects on agriculture and increased demand from urban growth will require timely improvements in crop cultivars, irrigation and drainage technology, and water management. In tropical regions, the use of agroforestry may be an economically feasible way to protect crop plants from extremes in microclimates and soil moisture [22].

Agricultural Pests and Diseases May Spread

Increased pest damage arises from changes in production systems, enhanced resistance of some pests to pesticides, and the production of crops in warmer and more humid climatic regions where crops are more susceptible to pests. Changes in crop management techniques, particularly the intensification of cropping, reduction in crop rotations, and increase in monocultures, have increased the activity of pests. The expansion of worldwide trade in food and plant products has also increased the impact of weeds, insects, and diseases on crops. The geographical ranges of several important insects, weeds, and pathogens in the US have recently expanded, including soybean cyst nematode (*Heterodera glycines*) and sudden death syndrome (*Fusarium solani* f. sp. *glycines*) (Fig. 4) [15,39,43].

Current climate trends and extreme weather events may be directly and indirectly contributing to the increased pest damage [17,39,52]. Downy mildew (*Plasmopara viticola*) epidemics on grape, the most serious grapevine disease in northern Italy, may increase under climate change, even though reduced precipitation may have a counterbalancing effect on disease pressure [44].

Such changes need to be put in the context of the global increases in pest-induced losses of crops in all regions since the 1940s [30,33] and the more than 33-fold increase in both the amount and toxicity of pesticide used over the same period [33]. Climate change thus may exacerbate environmental and public health issues related to agricultural chemicals [39], since increased applications of agricultural chemicals are likely to be needed in response to increasing disease pressure. Improved knowledge of the effects of climate on host–pathogen interactions will contribute to the adaptive capacity of agro-ecosystems.

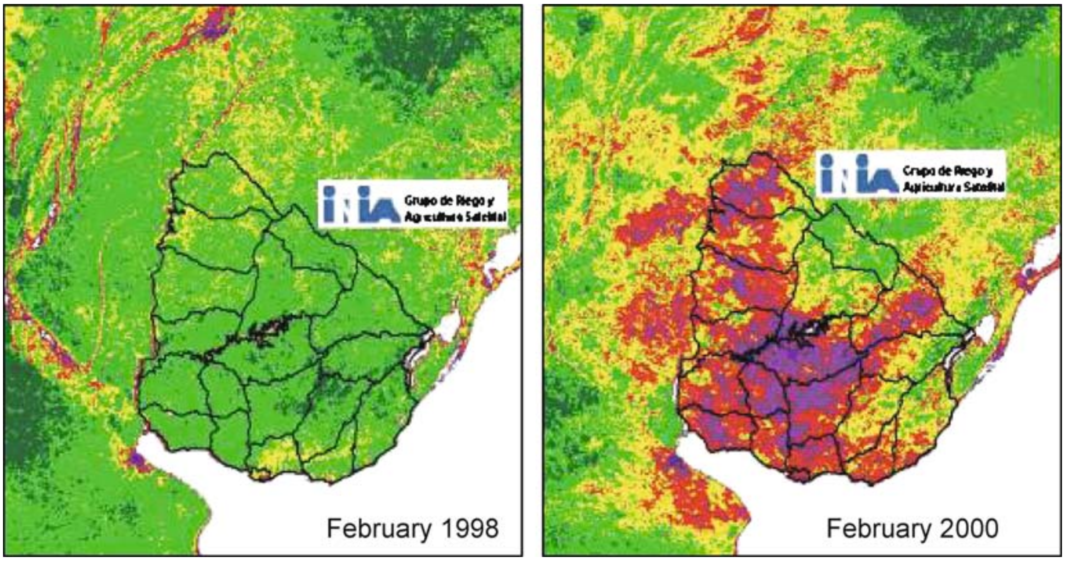


Climate Change and Agriculture, Figure 4
Range expansion of soybean cyst nematode (*Heterodera glycines*) from 1971 to 1989 (*top*) and soybean sudden death syndrome (*Fusarium solani f. sp. Glycines*) from 1973 to 1998 (*bottom*) in North America [40]

Current Climate Stress Is a Key Entry Point for Climate Change

There is an important interplay between current and future climate stresses. Since farmers have dealt with climatic fluctuations since the advent of agriculture, improving strategies for dealing with present climate extremes – such as droughts, floods, and heatwaves – is an important way to prepare for climate change. Many agricultural regions are affected by the major climate variability systems, including the processes known as the El Niño-Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO) [36]. The El Niño phase of the ENSO cycle tends to bring rainfall to Uruguay, while La Niña brings drought, as shown in Fig. 5 for 1998, an El Niño year, and 2000, the following La Niña.

In terms of prediction tools, ENSO models provide the opportunity for testing and validation of climate prediction and assessment on shorter seasonal-to-interannual time-scales. Skill in predicting climate changes on shorter time-scales, particularly the ENSO periods of the last twenty years when good observations exist, may lend credence to projections of global warming over the longer-term. As global climate models are further developed with



Climate Change and Agriculture, Figure 5
Vegetative Index (NDVI) for El Niño (1998) and La Niña (2000) years in Uruguay. *Green* = adequate water conditions; *red*/*purple* = drought conditions [2]

improved parametrizations and higher spatial resolution, they are likely to improve simulations of ENSO and other large-scale variability processes. The interaction of these systems with underlying anthropogenic trends caused by increasing greenhouse gas concentrations in the atmosphere is an active area of contemporary climate science. For regions directly affected by ENSO and other systems, such changes, if they do indeed occur, may become important manifestations of global warming.

Adaptation Is Necessary

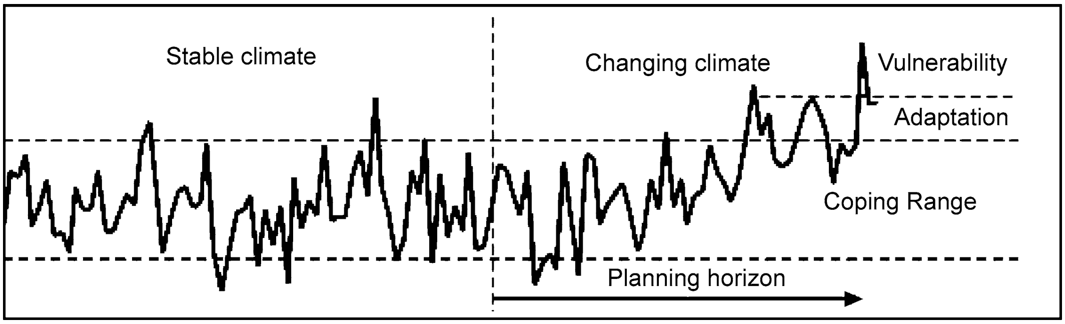
'Coping range' is a useful paradigm for improving responses to climate stresses of today and preparing for the climate changes of tomorrow [20]. An agricultural system may currently exist within a 'coping range' of climate variability that may be exceeded as incidence of extreme events increases under changing climate conditions (Fig. 6). The goal is to increase the coping range over which an agricultural system may thrive under such changes through the process of adaptation.

Adaptation can help farmers to minimize negative impacts of climate on human activities and ecosystems and to take advantage of potential beneficial changes. *Adaptation* to climate change can be defined as the range of actions taken in response to changes in local and regional climatic conditions [45]. Adaptation responses include both *autonomous adaptation* actions (i. e., those taken spontaneously and independently by individual farmers), and *planned adaptation* actions (i. e., those facilitated by climate-specific regulations and incentives put in place by regional, national and international policies) (see Table 1) [17]. In terms of the multiple factors impinging on agriculture, however, system responses to socio-economic, institutional, political or cultural pressures may outweigh response to climate change alone in driving the evolution of agricultural systems. The *adaptive capacity* of a system, in the context of climate change, can be viewed as the full set of system skills – i. e., technical solutions available to farmers in order to respond to climate stresses – as determined by the socio-economic and cultural settings, plus institutional and policy contexts, prevalent in the region of interest [17].

While current agronomic research confirms that at the field level crops would respond positively to elevated CO_2 in the absence of climate change (e. g., [1,17,19,21], the associated impacts of high temperatures, altered patterns of precipitation, and possibly increased frequency of extreme events (such as drought and floods) are likely to require a range of adaptation responses, some of which are listed in Table 2.

Mitigation Reduces Long-Term Risk

Agriculture has an important role to play in mitigation of climate change. *Mitigation* is defined as intervention aimed at reducing the severity of climate change by reducing the atmospheric concentration of greenhouse gases, ei-



Climate Change and Agriculture, Figure 6
Coping range of climate variability (adapted from [20])

Climate Change and Agriculture, Table 1
Adaptation approaches to climate impacts on agriculture

Approach	Definition	Operation
Autonomous	Adaptation that does not constitute a conscious response to climatic stimuli but is triggered by ecological changes in natural systems and by market or welfare changes in human systems. Also referred to as spontaneous adaptation.	Crop calendar shifts (planting, input schedules, harvesting) Cultivar changes Crop-mix changes
Planned	Adaptation that is the result of a deliberate policy decision, based on an awareness that conditions have changed or are about to change and that action is required to return to, maintain, or achieve a desired state.	Land-use incentives Irrigation infrastructure Water pricing Germplasm development programs

Climate Change and Agriculture, Table 2
Key agronomic impacts and adaptation responses

Agricultural impacts	Adaptation response
Biomass increase under elevated CO_2	Cultivar selection and breeding to maximize yield
Acceleration of maturity due to higher temperature	Cultivar selection and breeding of slower maturing types
Heat stress during flowering and reproduction	Early planting of spring crops
Crop losses due to increased droughts and floods	Changes in crop mixtures and rotations; warning systems; insurance
Increased pest damage	Improved management; increased pesticide use; biotechnology

ther by reducing emissions or by enhancing sinks. There are several major ways that the agricultural sector can contribute to climate change mitigation.

Soil Carbon Sequestration Of the approximately 150 GT of carbon that were lost in the last century due to land conversion to agriculture and subsequent production, about two thirds were lost due to deforestation and one-third, roughly 50 GT, due to cultivation of current agricultural soils and exports as food products [50]. The latter figure represents the maximum theoretical amount of carbon that could be restored in agricultural soils. In practice, as long as 40–50% of total above-ground grain or fruit production is exported as food to non-agricultural areas, the actual carbon amount that can be restored in agricultural soils is much lower.

Efforts to improve soil quality and raise soil organic carbon (SOC) levels include crop management and conservation tillage techniques. These practices have evolved as means to enhance sustainability and resilience of agricultural systems, rather than with soil carbon sequestration as primary motivation. They include so-called "best practice" agricultural techniques, such as use of cover crops and/or nitrogen fixers in rotation cycles, judicious use of fertilizers and organic amendments, soil water management improvements to irrigation and drainage, and improved varieties with high biomass production.

Conventional tillage is defined [27] as the mechanical manipulation of the topsoil that leaves no more than 15% of the ground surface covered with crop residues. In contrast, no-till management is defined as the avoidance of mechanical manipulation of the topsoil so as to leave it undisturbed and covered with surface residues from harvesting the prior crop to planting the new crop.

Best agricultural practices can result in a net augmentation of soil carbon and in enhanced productivity due to better soil structure and soil moisture management. The relevant practices include precise and timely applications and spatial allocation of fertilizers, use of slow-release fertilizers, prevention of erosion, shortening or elimination of fallow periods, use of high-residue cover crops and green-manure crops, and minimized mechanical disturbance of soil (e. g., zero tillage). Altogether, such practices may lead to partial or even complete restoration of the soil's organic carbon content where it had been depleted. In some cases, it might even be possible to store more carbon than had originally been present in the "virgin" soil. Where the soils had been severely degraded and their agricultural productivity greatly impaired, they may be converted to grassland or afforested so as to serve as carbon sinks.

The overall potential for carbon storage depends in each case on such factors as climate, type of vegetation, topography, depth and texture of the soil, past use (or abuse), and current management.

Along with sequestering carbon, these practices have the potential to improve soils in developing countries. In areas such as West Africa, soil fertility depletion has been described as the single most important constraint to food security [4]. Studies in smallholder agricultural farms in Africa have already illustrated significant increases in system carbon and productivity through organic-inorganic resources management (Roose and Barthes [35])

Biofuels Agriculture may help to mitigate anthropogenic greenhouse emissions through the production of biofuels. As has been demonstrated by ethanol based on corn production, issues involved with biofuel production include potential competition with food production, increased pollution from fertilizers and pesticides, and loss of biodiversity. Biofuels derived from low-input high-diversity mixtures of native grassland perennials can provide more usable energy, greater greenhouse gas reduction, and less agrichemical pollution per hectare than corn grain ethanol or soybean biodiesel [46]. The higher net energy results arise because perennial grasses require lower energy inputs and produce higher bioenergy yield. Furthermore, all aboveground biomass of the grasses can be converted to energy, rather than just the seed of either corn or soybean. These perennial grasses also sequester carbon at significant rates [46].

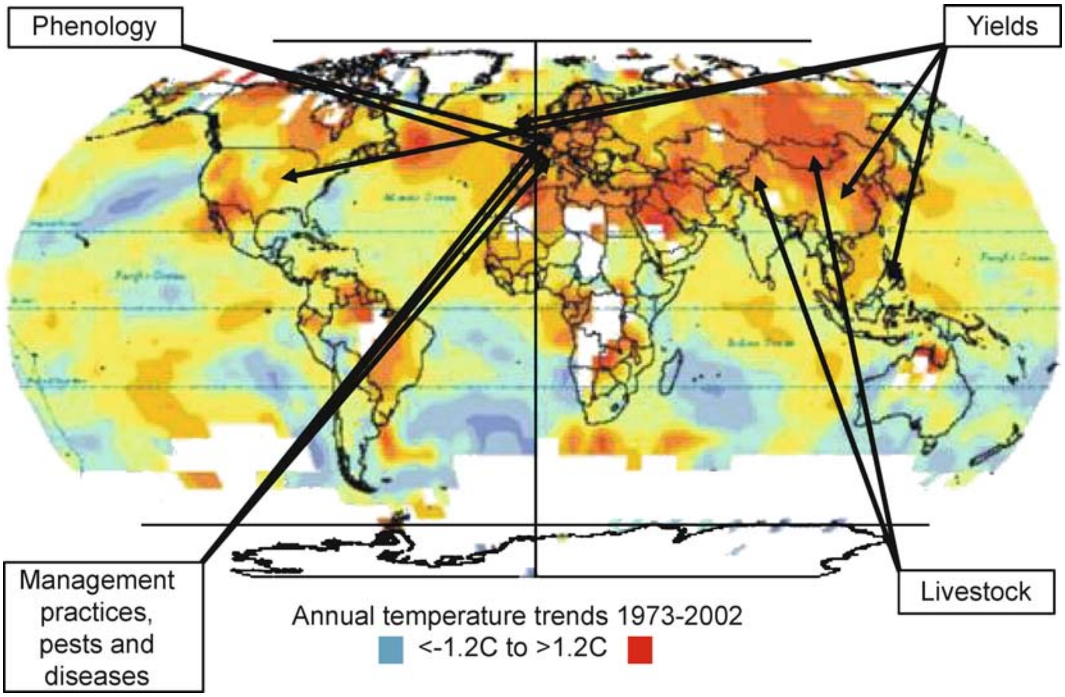
Other Greenhouse Gases Because of the greater global warming potential (GWP) of methane (21) and nitrous oxide (310) compared to carbon dioxide (1), reductions of non-CO_2 greenhouse gas emissions from agriculture can be quite significant and achieved via the development of more efficient rice (for methane) and livestock production systems (for both methane and nitrous oxide). In intensive agricultural systems with crops and livestock production, direct CO_2 emissions are predominantly connected to field crop production and are typically in the range of 150–200 kg C ha^{-1} yr^{-1} [14,51]. Recent full greenhouse gas analysis of different farm systems in Europe showed that such CO_2 emissions represent only 10–15% of the farm total, with emissions of CH_4 contributing 25–30% and emissions of N_2O accounting for as much as 60% of total CO_2-equivalent greenhouse gas emissions from farm activities. The N_2O contribution arises from substantial nitrogen volatilization from fertilized fields and animal waste, but it is also a consequence of its very high GWP.

In Europe, methane emissions are mostly linked to cattle digestive pathways; its contribution also dominates that of CO_2, due in part to methane's high GWP. Mitigation measures for methane production in livestock include improved feed and nutrition regimes, as well as recovery of bio-gas for on-farm energy production. Effective reduction of N_2O emissions is more difficult, given the largely heterogeneous nature of emissions in space and time and thus the difficulty of timing fertilizer applications and/or manure management. Large uncertainties in emission factors also complicate the assessment of efficient N_2O-reduction strategies. Current techniques focus on reduction of absolute amounts of fertilizer nitrogen applied to fields, as well as on livestock feeding regimes that reduce animal excreta.

Climate Change Effects on Agriculture Are Occurring Already

Agricultural effects of observed climate changes are being documented in many regions of the world (Fig. 7). Changes in crop phenology provide important evidence of responses to recent regional climate change. Such changes are apparent in perennial crops, such as fruit trees and wine-making varieties of grapes, which are less dependent on yearly management decisions by farmers than annual crops and are also often easier to observe.

Phenological changes are often observed in tandem with changes in management practices by farmers. Between 1951 and 2004 in Europe, agricultural crops have advanced 2.1 days/decade in response to recent warming in spring [28]. In Sahelian countries, increasing temperature in combination with rainfall reduction has led to


Phenology
Yields
Management practices, pests and diseases
Livestock
Annual temperature trends 1973-2002
🟦 <-1.2C to >1.2C 🟥

Climate Change and Agriculture, Figure 7
Locations of observed changes in agriculture in response to climate changes

a reduced length of vegetative period, no longer allowing present varieties to complete their cycle [5].

A negative effect of warming for local rice production has been observed by the International Rice Research Institute (IRRI) in the Philippines (yield loss of 15% for 1°C increase of growing-season minimum temperature in the dry season) [32]; a similar effect has been noted on hay yield in the UK (1°C increase in July–August led to a 0.33 t/ha loss) [6]. At the county level, US maize and soybean yields are demonstrating a positive effect of cooler and wetter years in the Midwest and hotter and drier years in the North-west plains [23]. In the case of the Sahel region of Africa, warmer and drier conditions have served as a catalyst for a number of other factors that have accelerated a decline in groundnut production [49]. For livestock, one study in Tibet reports a significant relationship of improved performance with warming in high mountainous conditions [9], while pasture biomass in Mongolia has been negatively affected by the warmer and drier climate, as observed at a local station [3] or at the regional scale by remote sensing [10].

Conclusions

Climate change brings both challenges and opportunities to agriculture. Farmers and researchers are being called on to simultaneously adapt to and mitigate climate change through a myriad of activities involving management practices, crop breeding, and new production systems. Some of these can be mutually re-enforcing, especially in view of the projected increased climate variability under climate change. This is because, most mitigation techniques currently considered in agriculture, including reduced tillage, were originally designed as “best practice” management strategies, aimed at enhancing the long-term stability and resilience of cropping systems in the face of climate variability or of increased cultivation intensity. By increasing the ability of soils to hold soil moisture and to better withstand erosion, and by enriching ecosystem biodiversity through the establishment of more diversified cropping systems, mitigation techniques implemented locally for soil carbon sequestration may also help cropping systems to better withstand droughts and/or floods, both of which are projected to increase in frequency and severity in future warmer climates. As climate change progresses, agriculture will continue to play a leading role in responding to a dynamic environment.

Future Directions

The time has come to incorporate climate as an essential factor in development planning and implementation. In the past, responses to climate variability were often too narrowly focused and lacking in institutional fit. Development programs are now beginning to include recommendations to mainstream responses to climate variability and change.

Magalhães [25] gave as an example of the need for a broad focus when considering climate in planning the early drought policies in Northern Brazil. Responses had focused on improving the water-supply infrastructure (e. g. building dams and digging wells) rather than on redressing the social and economic vulnerabilities and the need to build human capital by means of education, institutions, and market incentives for sustainability. Adaptive policies should be broadly conceived so as to increase and secure household entitlements, to change land-use patterns that lead to degradation, and to develop means of support for inhabitants that are less sensitive to the vagaries of climate.

At what levels or scales of organization (national, regional, household, and individual) can the variability of climate and the sustainability of agricultural production be addressed effectively? Because many nations encompass several and often most numerous climatic zones, the challenge faced by national agricultural managers and policy make is to foster sustainability at the regional level while building a foundation from the bottom-up at the individual as well as at the household and community levels. Dilley [8] believes that greater benefits of food security could be realized if knowledge were made more readily available at the household level, thereby improving the ability of more people to make even small adjustments based on anticipated climatic conditions. Multiple-scale efforts are clearly needed, with pathways of communication among the various levels and sectors of society.

Regions do not exist in isolation, as evidenced by the effects that extreme climate events occurring along the Atlantic coast of South America have on the sustainability of the inland Amazon rainforest. At least part of the pressure to deforest the Amazon region arises from the westward migration of farmers from Northern Brazil who suffer from ENSO-related droughts there (Magalhães [25]). Thus, policies related to sustainability and climate issues need to take regional interactions and their direct and indirect linkages into account.

Beyond interconnections among regions within a nation, there are the larger national concerns of economics, finance, and international relations that may affect the range of climate-adaptation policy choices. Engaging with the international community on issues of climate variability and change can lead to capacity building in both developed and developing countries, as climate and societal processes are studded and as improved understanding is incorporated into policies. This can be accomplished through interactions with international bodies dealing with climate variability and change, such as the World Meterological Organization and the United Nations Framework Convention on Climate Change. There is a growing realization that climate plays an important role in sustainable development: It is a component of natural capital, an occasional trigger to socio-economic crises caused by extreme events, and a long-term component of global environmental change.

Bibliography

1. Ainsworth EA, Long SP (2005) What have we learned from 15 years of free-air CO_2 enrichment (FACE)? A meta-analysis of the response of photosynthesis, canopy properties and plant production to rising CO_2. New Phytologist 165:351–275
2. Baethgen WE, Giménez A (2002) Seasonal climate forecasts and the agricultural sector of Uruguay. In: Examples of ENSO-Society Interactions. The International Research Institute for Climate and Socity, New York. http://iri.columbia.edu/climate/ENSO/societal/resource/example/Baethgen.html
3. Batimaa P (2005) The potential impact of climate change and vulnerability and adaptation assessment for the livestock sector of Mongolia. Assessments of Impacts and Adaptations to Climate Change. AIACC, Washington DC, 20 pp
4. Bationo A, Kihara J, Vanlauwe B, Waswa B, Kimetu J (2007) Soil organic carbon dynamics, functions and management in West African agro-ecosystems. Agric Syst 94:13–25
5. Ben Mohamed A, Duivenbooden NV, Abdoussallam S (2002) Impact of climate change on agricultural production in the Sahel, Part 1. Methodological approach and case study for millet in Niger. Clim Chang 54:327–348
6. Cannell MGR, Palutikof JP, Sparks TH (1999) Indicators of climate change in the UK. DETR, London, 87 pp
7. Chmielewski FM, Muller A, Bruns E (2004) Climate changes and trends in phenology of fruit trees and field crops in Germany, 1961–2000. Agric Forest Meteorol 121(1-2):69–78
8. Dilley M (2003) Regional responses to climate variability in Southern Africa. In: O'Brian K, Vogel C (eds) Coping with climate variability: The use of seasonal climate forecasts in Southern Africa. Ashgate, Hampshire, pp 35–47
9. Du MY, Kawashima S, Yonemura S, Zhang XY, Chen SB (2004) Mutual influence between human activities and climate change in the Tibetan Plateau during recent years. Global Planet. Change 41:241–249
10. Erdenetuya M (2004) Application of remote sensing in climate change study: Vulnerability and adaptation assessment for grassland ecosystem and livestock sector in Mongolia project. AIACC Annual Report, Washington DC
11. Fischer G, Shah M, Velthuizen H, Nachtergael FO (2001) Global agro-ecological assessment for agriculture in the 21st century. International Institute for Applied Systems Analysis. IIASA, Laxenburg
12. Fischer G, Shah M, van Velthuizen H (2002) Climate change and agricultural vulnerability, special report to the UN World Summit on Sustainable Development, Johannesburg 2002. IIASA, Laxenburg

13. Fischer, Shah GM, Tubiello FN, van Velhuizen H (2005) Socioeconomic and climate change impacts on agriculture: An integrated assessment, 1990–2080. Philos Trans R Soc B-Biol Sci 360(1463):2067–2083
14. Flessa H, Ruser R, Dörsch P, Kamp T, Jimenez MA, Munch JC, Beese F (2002) Integrated evaluation of greenhouse gas emissions (CO_2, CH_4, N_2O) from two farming systems in southern Germany. Agric Ecosyst Environ 91:175–189
15. Hartman GL, Noel GR, Gray LE (1995) Occurrence of soybean sudden death syndrome in east-central Illinois and associated yield losses. Plant Disease 79:314–318
16. IPCC (2007) Climate change 2001: The scientific basis. Contributions of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
17. IPCC (2007) Climate change 2007: Impacts, adaptation, and vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
18. IPCC (2007) Climate Change 2001: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, 1076 pp
19. Jablonski LM, Wang X, Curtis PS (2002) Plant reproduction under elevated CO_2 conditions: A meta-analysis of reports on 79 crop and wild species. New Phytologist 156(1):9–26
20. Jones PD, Mann ME (2004) Climate over past millennium. Rev Geophys 42:RG2002
21. Kimball BA, Kobayashi K, Bindi M (2002) Response of agriculture crops to free-air CO_2 enrichment. Adv Agron 77:293–368
22. Lin BB (2007) Agroforestry management as an adaptive strategy against potential microclimate extremes in coffee agriculture. Agric Forest Meteorol 144(1-2):85–94
23. Lobell DB, Asner GP (2003) Climate and management contributions to recent trends in US agricultural yields. Science 299:1032
24. Lobell DB, Field CB, Cahill KN, Bonfils C (2006) Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties. Agricult Forest Meteorol 141:208–218
25. Magalhães AR (2000) Sustainable development: Climate and policy linkages. In: Proceedings of the International Forum on Climate Prediction, Agriculture and Development, 26-–28 April. International Research Institute for Climate Prediction, Palisades, New York, pp 3–10
26. Magalhães AR, Glantz MH (eds) (1992) Socioeconomic impacts of climate variations and policy response in Brazil. Esquel Brazil Foundation, Brasilia
27. Marland G, West TO, Schlamadinger B, Canella L (2003) Managing soil organic carbon in agriculture: The net effect on greenhouse gas emissions. Tellus 55B:613–621
28. Menzel A, von Vopelius J, Estrella N, Schleip C, Dose V (2006) Farmers' annual activities are not tracking speed of climate change. Climate Res 32:201–207
29. Millennium Ecosystem Assessment (2005) Ecosystem and human well-being: Synthesis. Island, Washington DC
30. Oerke EC, Dehne HW, Schohnbeck F, Weber A (1995) Crop production and crop protection: Estimated losses in major food and cash crops. Elsevier, Amsterdam, 830 pp
31. Parry ML, Rosenzweig C, Iglesias A, Livermore M, Fischer G (2004) Effects of climate change on global food production under SRES emissions and socio-economic scenarios. Glob Environ Chang 14:53–67
32. Peng SB, Huang JL, Sheehy JE, Laza RC, Visperas RM, Zhong XH, Centeno GS, Khush GS, Cassman KG (2004) Rice yields decline with higher night temperature from global warming. Proc Natl Acad Sci USA 101(27):9971–9975
33. Pimentel D (1997) Pest management in agriculture, In: Pimentel D (ed) Techniques for reducing pesticide use: Environmental and economic benefits. Wiley, Chichester, pp 1–12
34. Reilly J, Tubiello F, McCarl B, Abler D, Darwin R, Fuglie K, Hollinger S, Izaurralde C, Jagtap S, Jones J, Mearns L, Ojima D, Paul E, Paustian K, Riha S, Rosenberg N, Rosenzweig C (2003) US agriculture and climate change: New results. Clim Chang 57:43–69
35. Roose E, Barthes B (2001) Organic matter management for soil conservation and productivity restoration in Africa: a contribution from Francophone research. Nutrient Cycling in Agroecosystems 61(1–2):159–170
36. Rosenzweig C, Hillel D (2008) Climate variability and the global harvest. Oxford University Press, Oxford
37. Rosenzweig C, Parry ML (1994) Potential impacts of climate change on world food supply. Nature 367:133–138
38. Rosenzweig C, Tubiello F (2007) The interactions of adaptions and mitigation strategies in agriculture. Mit Adapt Strategies Glob Change 12(5):855–873
39. Rosenzweig C, Iglesias A, Yang XB, Epstein PR, Chivian E (2000) Implications of climate change for US agriculture: Extreme weather events, plant diseases, and pests. Center for Health and the Global Environment, Harvard Medical School. Cambridge, 56 pp
40. Rosenzweig C, Iglesias A, Yang XB, Epstein PR, Chivian E (2000) Climate change and extreme weather events: Implications for food production, plant diseases, and pests. Global Change and Human Health 2(2)
41. Rosenzweig C, Tubiello FN, Goldberg R, Mills E, Bloomfield J (2002) Increased crop damage in the US from excess precipitation under climate change. Glob Environ Chang 12:197–202
42. Rosenzweig C, Strzepek KM, Major DC, Iglesias A, Yates DN, McClusky A, Hillel D (2004) Water resources for agriculture in a changing climate: International case studies. Glob Environ Chang 14:345–360
43. Roy KW, Rupe JC, Hershman DE, Abney TS (1997) Sudden death syndrome. Plant Dis 81:1100–1111
44. Salinari F, Giosue S, Tubiello FN, Rettori A, Rossi V, Spanna F, Rosenzweig C, Gullino ML (2006) Downy mildew (Plasmopara viticola) epidemics on grapevine under climate change. Glob Chang Biol 12(7):1299–1307
45. Smit B, Burton I, Klein RJT et al (2000) An anatomy of adaptation to climate change and variability. Climate Chang 45(1):223–251
46. Tillman D, Hill J, Lehman C (2006) Carbon-negative biofuels from low-input high-diversity grassland biomass. Science 314:1598–1600
47. Tubiello FN, Amthor JS, Boote KJ, Donatelli M, Easterling W, Fischer G, Gifford RM, Howden M, Reilly J, Rosenzweig C (2006) Crop response to elevated CO2 and world food supply – A comment on "Food for Thought …" by Long et al. Science 312:1918–1921:2006. Eur J Agron 26(3):215–223 APR 2007
48. National Assessment Synthesis Team (2001) Climatic Change Impacts on the US: The potential consequences of climate

variability and change. US Global Change Research Program, Washington DC

49. Van Duivenbooden N, Abdoussalam S, Mohamed AB (2002) Impact of climate change on agricultural production in the Sahel. Part 2. Case study for groundnut and cowpea in Niger. Clim Chang 54:349–368
50. Watson RT, Noble IR, Bolin B, Ravindranath NH, Verando DJ, Dokken DJ (2000) IPCC special reports. Land use, land-use change, and forestry. Cambridge Univ Press, Cambridge, 324
51. West TO, Marland G (2002) A synthesis of carbon sequestration, carbon emissions, and net carbon flux in agriculture: Comparing tillage practices in the United States. Agr Ecosyst Environ 91:217–232
52. Yang XB, Scherm H (1997) El Niño and infectious disease. Science 275:739

Climate Change, Economic Costs of

Richard S. J. Tol[1,2,3,4]

[1] Economic and Social Research Institute, Dublin, Ireland

[2] Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands

[3] Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands

[4] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, USA

Article Outline

Glossary

Direct costs The direct cost equals quantity times price.

Discount factor The discount factor t years into the future equals one over one plus the discount rate, raised to the power t.

Discount rate The discount rate is the annual rate of decline of the value of consumption. It is roughly equal to the rate of interest, or the opportunity cost of capital. The discount rate consists of two components: the rate at which consumers get better off, and the rate of pure time preference (or impatience). The discount rate is not related to the rate of inflation, which is the annual rate of decline of the value of money.

Economics Economics is the social science which studies human behavior with regard to the relationship between ends and scarce means which have alternative uses.

Equity weights Equity weights are applied to aggregate national impacts to a global total. Equity weights are often set to unity, but sometimes equity weights equal the ratio of nationally average per capita income over global average per capita income.

Indirect costs The indirect costs equal all costs that are not direct costs. This includes the price change induced by the change in quantity (partial equilibrium effects), the changes in other markets (general equilibrium effects), and the changes at later times (dynamic effects).

Marginal costs The marginal cost of greenhouse gas emissions equals the first partial derivative of the net present value of the total costs of climate change to emissions.

Monetary valuation Monetary valuation is a set of techniques and their application that attempts to express in monetary terms the value to humans of changes in environmental goods and services. Negative impacts are typically expressed as an income loss that would give an equivalent loss in welfare.

Neo-classical economics Although historians would refer to neo-classical economics as the dominant form of economic research between 1860 and 1910, common usage has neo-classical economics as a synonym for mainstream or orthodox economics. In that sense, neo-classical economics is a style of research, characterized by empirical rigour, mathematical rigour, and micro-founded macro-relationships.

Net present value The net present value is the sum of all future costs and benefits, weighted by the discount factor.

Total costs The total cost of climate change equals the direct and the indirect costs of climate change, that is, the difference in welfare between a scenario with climate and a scenario without.

Definition of the Subject

The economic costs of climate change include all positive and negative impacts of the enhanced greenhouse effect and the resulting changes in the atmosphere and ocean on all human consumers and producers. Total costs refer to the difference in human welfare between a scenario with climate change and a scenario without climate change.

Marginal costs refer to the difference in human welfare between two scenarios with a slightly different climate, normalized by the amount of greenhouse gas emissions that would induce that difference. Estimates of the economic costs of climate change are important to assess the size of the climate problem relative to other problems, and to compare the costs of climate change to the costs of greenhouse gas emission reduction.

Introduction

Calls for greenhouse gas emission reduction are often phrased as a moral imperative. While tempting, this is wrong. Firstly, there is no moral agreement. Emission reduction could save polar bears but it would cost coal miner's jobs and raise the price of food for the malnourished. Moral imperatives are easy if a policy has only benefits. As soon as a policy has both costs and benefits, one has to make trade-offs and choose the lesser evil. Secondly, there is no avoiding dangerous interference with the climate system. Emission reduction would slow down the melting of the Greenland ice cap, and reduce the probability of a collapse of the West-Antarctic Ice Sheet – but it would not stop the melting or bring the change to zero. Thirdly, we have no obligations to future generations or poor people. Such duties are self-imposed. And even if we choose to help others, there are many ways to do this. Would our grandchildren prefer a richer but warmer world, or a poorer but colder one? Would the grandchildren of the Bangladeshis like us to reduce greenhouse gas emissions, help them to adapt to climate change, or help their grandparents grow rich?

This chapter looks into such questions. It is written from the thoroughly relativistic perspective of a neo-classical economist. The basic principles of the economic theory of climate change are quite simple. Greenhouse gas emissions are an externality. Externalities are unintended – we burn coal to make electricity, not to emit carbon dioxide – and uncompensated – carbon dioxide is freely dumped in the atmosphere – consequences of economic activity. Externalities should be internalized, that is, emitters should pay for their emissions. The price of emissions should equal the damage done by the emissions. That is all.

There is now a vast literature on the economics of climate change – started by Nordhaus [73,74]. A large part of that literature is about the deviation between the simple policy prescription of economic theory and the complexities of actual policy. Another large part of the literature is about the costs of greenhouse gas emission reduction. The literature on the economic costs of climate change is only a small one. It is reviewed here.

Section "Issues" discusses the methodological, conceptual, and moral issues one has to confront when estimating the economic impacts of climate change. Section "Total Costs" reviews estimates of the total economic impact. Section "Marginal Costs" surveys estimates of the marginal impacts. Section "Policy Implications" concludes by assessing the policy implications.

Issues

Scenarios

A scenario is a set of assumptions on future conditions that is coherent, internally consistent, and not implausible [70]. Climate scenarios are usually derived from modeling experiments with General Circulation Models (GCM). Climate scenarios include simple statistics such as the global mean surface air temperature, and complex results such as spatial patterns of rainfall extremes. Climate scenarios may also include low-probability events, such as a disruption of the thermohaline circulation in the Atlantic Ocean, or the collapse of the West Antarctic Ice Sheet.

Scenarios also include population, economic activity, greenhouse gas emissions, and land use. Besides driving the climate models, these components are also important as they determine the vulnerability of social and economic systems to climate change over time. Although poorer societies are generally believed to be more vulnerable to climate change, this is by no means a simple relationship. Some impacts tend to fall with economic growth. The impacts of climate change on infectious diseases are a prime example. Malaria does not kill middle income people, because they can afford prevention and (if necessary) cure. Some impacts tend to rise with economic growth. The impact of climate change on biodiversity and species loss is one example. People tend to care more about these matters as their income grows, and further developed economies put more pressure on nature. Other impacts may rise first and then fall with economic growth. Urban air quality is one example. The very poor have nothing to foul the air with, and the very rich do not like foul air and have the wherewithal to prevent it. Climate change is likely to increase malaria, reduce biodiversity and worsen air quality. In some case, climate change and economic growth work together to increase the impacts, while in other cases they pull in opposite directions.

Valuation Approach

There are various techniques for the monetary valuation of climate change impacts. Some values of impacts are directly based on observed prices. Agriculture and dike building are examples. Other values can be indirectly measured on the basis of observed market prices for surrogate products or services. One example is human health, which is not traded directly but indirectly through safety measures, labor markets, and health care. The challenge in these instances is to model future market prices that are consistent with the underlying socioeconomic scenario. For yet other impacts, no market values exist, and hypothetical prices are needs. Notable impacts are on non-commercial ecosystems and biodiversity.

Because it is practically impossible to estimate each exposure-response relationship or value at the respective geographical location of a climate change impact, data from previous studies focusing on different locations and different policy contexts are inevitable. Furthermore, most climate change impacts will take place in the future, for which by definition no data are available. Therefore it is important to know when data from other studies can be used and under what conditions, and how to extrapolate values from today to tomorrow.

The majority of recent studies still adopt benefit transfer methods for the evaluation of climate impacts. However, benefit transfer is not very reliable [91]. For this reason, more attention should be given to original valuation research in the context of climate change.

An example of such a study is Li et al. [54] who analyze the willingness-to-pay (WTP) of American citizens for climate policy by means of the contingent valuation method. They find that the median American citizen is willing to pay about $ 15/tC. Berrens et al. [7] find a willingness-to-pay between $ 200 and $ 1760 per US household per year (0.2–2.3% of income) for US ratification of the Kyoto Protocol. (Manne and Richels [60] estimate that the costs of US ratification would be 0.75% of GDP in 2010.) Hersch and Viscusi [46] find that Europeans are willing to pay up to 3.7% more for petrol if that helps combat climate change. Viscusi and Zeckhauser [117] find that Harvard students are willing to pay $ 0.50/gallon (a 25% price increase) or 3% of their expected annual income for greenhouse gas emission reduction. (This study also showed that these students underestimate projected warming in Boston by about 50%, while the authors made them believe that carbon dioxide emission reduction would be effective for slowing climate change in the next 30 years.) There is scope for similar applications of WTP techniques, mainly to account for spatial and socio-economic differences in individuals' preferences.

Direct and Higher Order Impacts

Most studies to date have estimated the direct costs of climate change. Direct costs equal the physical change (e. g., the dikes to be reinforced) times their price (in this example, the costs per dike length and dike height). Direct costs are easy to compute (but see Subsect. "Valuation Approach"), but probably underestimate the real economic costs.

The higher order impacts come in three kinds. Firstly, climate change may impact the market under consideration. For example, if dikes are being reinforced everywhere, then the costs of dike building is likely to go up as materials, machinery, and skilled labour is difficult to get. Secondly, the impact of climate change on one market may spill over into other markets. For example, dike building may increase the costs of construction, as the same materials and skills are used. Dike building is capital intensive and may drive up the interest rate. Thirdly, the impact of climate change may affect economic growth. For example, money invested in dike building is not invested elsewhere.

A number of recent studies have examined the economy-wide implications of sea level rise [10], tourism [8], and health [9]. While it is perhaps too early to draw firm conclusions from this body of research, the studies suggest that the indirect effects of climate change impacts can both enlarge and diminish the direct economic impacts of climate change. The distribution of gains and losses is another difference between direct costs and general equilibrium effects. Whereas direct costs are limited to those directly affected, markets would spread the impact to their suppliers, clients, and competitors as well as to financial markets.

Fankhauser and Tol [35] show that the economic growth impact of climate change is as large as the direct impact of climate change. Acemoglu et al. [1] and Masters and McMillan [62] show that differences in climate explain part of observed differences in economic development. Easterly and Levine [28] show that the link is at most weak and indirect, and it is not clear whether the mechanisms that may have been active in the past, still hold for present and future.

Adaptation

One cannot study the costs of climate change impact without also studying, or at least making assumptions about the costs of adaptation [109]. Studies focusing on costs of the impacts make widely differing assumptions about the amount of adaptation that will take place. While some studies completely ignore adaptation, other studies consider arbitrary levels of adaptation, or assume optimal adaptation. No studies use realistic models of adaptation [109]. There is little research that shows how adaptation costs compare to the potential damages of not adapting. The impacts of climate change and the capacity to adapt would be affected with the level of development and flexibility of the economy [123]. Hence, the future success and nature of adaptation depends on the assumed socio-economic scenario.

Aggregation: Temporal

Climate change is a slow process. Today's emissions will affect the climate for decades to centuries, and sea level for centuries to millennia. As cause and effect are separated in time, so are costs (of emission reduction) and benefits (of avoided climate change). The procedure to make commensurate costs and benefits at different points in time is called discounting. Discounting is as common as it is controversial. See [3] for an excellent discussion.

Individuals discount future gains or losses because of two reasons. (People may also discount the future because it is more uncertain than the present, but in this case discounting is used as a shortcut for an uncertainty analysis.) First, money earns interest. Second, people are impatient. The first reason is widely accepted. Davidson [26] is one of the few exceptions. On the second reason, there is virtual consensus too. All ethical arguments show that people should not discount (e. g., [12]). All empirical evidence shows that people do nonetheless (e. g., [79,80]).

Climate change is a large-scale problem. Therefore, the discount rate of society is more relevant than the individual discount rate. The appropriate measure of the growth rate of money is the average growth rate of per capita consumption. Again, there is little dispute on this. But should the social rate of discount also include a measure of impatience? Again, philosophers agree: Impatience is immoral. However, this implies that a government would deviate from the will of the people. This may be defended with the argument that the government is the guardian of future, yet unborn people. However, the empirical evidence is clear in this case too: Governments are impatient [31].

Discounting is more profound over long periods than over short ones. Discounting implies that climate change damages that occur in a century or so are largely irrelevant. This realization has led people to rethink the fundamental principles of discounting, particularly

(a) the notion that the procedure of discounting results from the intertemporal allocation of resources of an individual agent; and
(b) the assumption that discounting is exponential.

To start with the individual perspective, Lind [56] and Lind and Schuler [55] argue that *earmarked* investment is a crucial assumption in discounting. The discount factor measures the trade-off between consumption now and consumption later, where consumption later is contingent on a specific investment plan. As the current generation cannot commit near-future generations to maintain their investments for the benefit of far-future generations, discounting breaks down between generations. Schelling [95] agrees. The alternative is to decide explicitly on the resource allocation between generations. Chichilnisky [19] shows that discounting coincides with a dictatorship of the present generation over future generations. Gerlagh and Keyzer [38] show that discounting is equivalent to the present generation owning all future resources. This is objectionable from a moral standpoint, but it is reality. This line of research has not led to practical alternatives to discounting.

Conventional discounting is exponential: The discount factor is $(1+r)^{-t}$, where r is the discount rate and t is time. Some people argue that the functional specification of conventional discounting is wrong. The first component is empirical. Conventional exponential discounting has that the relative difference between two years is always equal, regardless of their distance from the present. That is, the difference between year 10 and 11 is the same as the distance between year 100 and 101. However, many people would in fact argue that the difference between year 10 and 11 is equal to the difference between year 100 and 110. Such hyperbolic discounting [22] is very similar to exponential discounting for short periods, but the difference is substantial for long periods. The similarity between exponential and hyperbolic discounting in the short run is important, because a switch to hyperbolic discounting would imply a drastic overhaul of long-term decisions only.

There are two further arguments for hyperbolic discounting cf. Dasgupta and Maskin [25]. The first is due to Weitzman [118]. He shows that, if one is uncertain what discount rate to use, then the lowest discount rate becomes increasingly dominant over time. The certainty-equivalent discount rate falls with time, and the difference between years shrinks in the more distant future. Consider the following example. After one year, the average of a 1% and a 10% discount rate is

$$1-\left(\frac{\left(1.01^{-1}+1.10^{-1}\right)}{2}\right)^{1/1}=5.0\,\%$$

(and not 5.5%). After 100 years,

$$1-\left(\frac{\left(1.01^{-100}+1.10^{-100}\right)}{2}\right)^{1/100}=1.7\,\%\,.$$

That is, the average approaches the minimum as time progresses. One may criticize this as a short cut for a full uncertainty analysis. However, Gollier [39,40] shows that the same is true if a government somehow aggregates the individual discount rates of its citizens. In the long run, the preferences of the person with the lowest discount rate become increasingly important, and the discount rate declines over time.

Guo et al. [41] and Newell and Pizer [71] show that hyperbolic discounting leads to higher estimates of the social cost of carbon. However, the quantitative effect is limited by the fact that hyperbolic discount rates are high for the first decades.

Aggregation: Spatial

Climate change is a global problem. Carbon dioxide and other greenhouse gases mix uniformly in the atmosphere. This implies most of the impacts of one country's emissions fall on other countries. The same is true for the benefits of emission reduction. The impacts on different countries need to be aggregated somehow.

Two methods dominate the literature. In the first and oldest method, regional impacts are quantified in local currencies, converted to dollars, say, and added up [33,103]. This is simple, but the disadvantage is that similar impacts are treated differently. Most disturbingly, climate-change-induced deaths in rich countries receive a greater weight than climate-change-induced deaths in poor countries. The second method, known as equity weighing, corrects for this [6,36]. Rather than simply adding regional estimates, the regional utility-equivalents are added and then converted back to money according to an assumed global welfare function. A big disadvantage of this method is that climate-change-induced deaths are treated differently than deaths by other, national causes. The reason for this discrepancy is that equity weighing, as practiced in the literature, explicitly assumes a global decision maker.

In the meta-analysis of Tol [107], the median estimate of the marginal damage costs of carbon dioxide is $ 10/tC without equity weights, and $ 54/tC with equity weights. So, equity weighing is obviously important. The reason is simple. Poor countries are more vulnerable to climate change. Poor countries have little economic weight. Equity weights correct for this.

Morally, this may be the right thing to do. However, national governments also have a certain obligation to defend the interests of their citizens. A narrow interpretation of self-interest would suggest that impacts abroad be ignored (unless they spill over, e. g., through international migration). Then, climate change policy would be very limited, as most impacts will be abroad. However, the principle of good neighborhood is well established, both morally and legally. This entails that one should avoid doing harm to others; and should pay compensation if harm is done nonetheless (e. g., [113]).

A rational actor would avoid doing harm if that is cheaper than the compensation paid. From a national perspective, the relevant damages are then the impacts on the own country plus the compensation paid to other countries. Schelling [93] forcefully argues that compensation should equal the welfare loss of the victim rather than the welfare loss that the culprit would have experienced had she been the victim. This argues for aggregation of monetized impact estimates without equity weighing.

However, compensation would need to be paid only once. Furthermore, a country would also reasonably expect to be compensated itself. This implies that the damage to a country equals the global damage times its share in causing the problem. Defining the latter is a thorny issue, as the cause-effect chain is long, complex, and uncertain. One would need to make arbitrary decisions on cause, effect and their connection.

Uncertainty

Climate change is plagued by uncertainty [16]. Partly, this is because our understanding of climate change and its impacts is incomplete. For the larger part, however, this is because climate change will take place in the future, partly driven by future emissions, and impacting a future world. Future research and observations may reduce the uncertainty, although surprises may increase the uncertainty just as well, but uncertainty will never disappear. Learning and irreversibility play a crucial role in how to deal with uncertainty. Events that may or may not occur in some distant future, but whose consequences can be alleviated once it becomes clear if they would occur, should not worry us too much. On the other hand, if an effect is irreversible (e. g., species extinction), we may want to prevent it regardless of how uncertain it is and regardless of what future research will show (according to the "precautionary principle"). Another crucial part of dealing with uncertainty is risk aversion. Essentially, this determines how much weight we place on negative surprises. A risk neutral decision maker would cancel negative surprises against positive ones, but a risk adverse decision maker would not. Recent work has shown that the marginal damage costs of carbon dioxide are indeed very sensitive to the assumed degree of risk aversion. Although uncertainty and risk are often emphasized – often in a casual way – only few studies seek to quantify its implications (e. g., [51]).

In a recent paper, Weitzman [120] shows that, under a wide range of standard assumptions, the uncertainty about climate change is so large that the expected value of the social costs of climate change is infinite. Earlier, Tol [106] showed this for a specific model. This implies that uncertainty should take central stage in the analysis of climate policy. The Weitzman result throws up a number of methodological issues that will need to be resolved before the policy implications of this work become clear.

Completeness

The impacts of climate change that have been quantified and monetized include the impacts on agriculture and forestry, water resources, coastal zones, energy consumption, air quality, and human health. Obviously, this list is incomplete. Also within each impact category, the assessment is incomplete. Studies of the impacts of sea level rise on coastal zones, for instance, typically omit saltwater intrusion in groundwater [72]. Furthermore, studies typically compare the situations before and after climate change, but ignore that there will be a substantial period during which adaptation is suboptimal – the costs of this are not known.

Some of the missing impacts are most likely negative. Diarrhoea impacts have been quantified recently [57]. Like malaria, diarrhoea is a disease that is driven by poverty but sensitive to climate. Including diarrhoea tightens the link between development and climate policy. Increasing water temperatures would increase the costs of cooling power plants [101]. Redesigning urban water management systems, be it for more or less water, would be costly [4], as would implementing the safeguards against the increased uncertainty about future circumstances. Roads and bridges would suffer from weather conditions for which they were not designed; this would imply either disruption of traffic or expensive retrofits. Extratropical storms may well increase, leading to greater damage and higher building standards [27]. Expenditures on these things are relatively small. Even if climate change would double or triple the cost, the impact would be small. Ocean acidification would reduce marine biodiversity, and may well harm fisheries [52]. Ocean fisheries are only a small, and declining fraction of GDP, while there are ready substitutes for wild fish protein (notably fish farming). The value of biodiversity is unclear (see below).

Other missing impacts are probably positive. Higher wind speeds in the mid-latitudes would decrease the costs of wind and wave energy [11,44]. Less sea ice would improve the accessibility of arctic harbours, would reduce the costs of exploitation of oil and minerals in the Arctic, and may even open up new transport routes between Europe and East Asia [121]. Warmer weather would reduce expenditures on clothing and food, and traffic disruptions due to snow and ice [15]. Also in these cases, the impact of climate change is likely to be small relative to the economy.

Some missing impacts are positive in some places, and negative in others. Tourism is an example. Climate change may well drive summer tourists towards the poles and up the mountains [42,43]. People, however, are unlikely to change the time and money spent on holiday making. The effect is a redistribution of tourist revenue [8]. The global impact is close to zero, but regional impacts are measured in tens of billions of dollars – positive in temperate, rich countries, and negative in tropical, poor countries. This exacerbates the already skewed distribution of climate impacts. Some ski resorts may go out of business, and others would need expensive snowmaking equipment [29,97]. Other ski resorts would profit from the reduced competition. Although regional impacts may be substantial, at the global scale positives and negatives cancel.

Other impacts are simply not known. Some rivers may see an increase in flooding, and others a decrease [53]. At the moment, only a limited number of rivers have been studied in detail, and it is unclear how to extrapolate to other rivers. It is clear though, that land use and water management may greatly increase or reduce impacts. Although river floods wreak substantial havoc and damages of a single event can reach substantial numbers, average flood damage is in fact small relative to the economy [112]. Tropical storms do more damage, although a substantial share of the impact is due to bad planning rather than bad weather [14]. Nonetheless, tropical storms may prevent capital accumulation and the plantation of lucrative crops such as banana [30,69]. Unfortunately, it is not known how climate change would alter the frequency, intensity, and spread of tropical storms [63,89].

The missing impacts discussed above are probably small. There are also bigger gaps in the coverage of climate change impact studies. Climate change is likely to have a profound impact on biodiversity, but quantitative predictions are rare [13]. Although the economic impact of a small change in biodiversity is known to be small [88], the value of large biodiversity changes is unknown but could well be substantial [18]. There is a small but unknown chance that climate change will be more dramatic than is typically assumed in the impacts literature. This may be because of shutdown of the thermohaline circulation [61], a collapse of the Greenland or West-Antarctic Ice Sheet [84], or a release of large amounts of methane [45]. The economic analysis of such scenarios has only just begun [57]. It may be that climate change would lead to large-scale migration [64] and violent conflict, although there is only weak empirical support for this [49,124]. Finally, climate change impact studies stop at the end of the 21st century. In 2100, impacts are negative, and getting more negative at an accelerating pace. It is not known how rapidly things would get worse in the 22nd century without emission abatement.

Although the sign of the aggregate unknown impacts is not known, risk aversion would lead one to conclude that greenhouse gas emission reduction should be more stringent than suggested by a cost-benefit analysis based on the quantified impacts only. However, the size of the bias is unknown too – so the main policy implication is that more research is needed.

Total Costs

The first studies of the welfare impacts of climate change were done for the USA [21,74,98,102]. Although Nordhaus [74] (see also Ayres and Walter [5]) extrapolated his US estimate to the world, the credit for the first serious study of the global welfare impacts goes to Fankhauser [32,33], although Hohmeyer and Gaertner [48] earlier published some low quality estimates. Other global estimates include those by Nordhaus [76,77], Tol [103], Nordhaus and Yang [82], Plambeck and Hope [90], Nordhaus and Boyer [81], Mendelsohn et al. [66,68], Tol [105], Maddison [59], Hope [50], Rehdanz and Maddison [92] and Nordhaus [78]. Note that Stern et al. [100] is based on Hope [50].

This is a rather short list of studies, and an even shorter list of authors. This problem is worse if one considers that Nordhaus and Mendelsohn are colleagues; that Fankhauser, Maddison and Tol are students of Pearce; and that Rehdanz is a student of Maddison and Tol; while Hope's (and Stern's) estimates are averages of Fankhauser's and Tol's. Although most fields are dominated by a few people, dominance is here for want of challengers. The effect of this is hard to gauge. The reasons are lack of funding (this work is too applied for academic sources, while applied agencies do not like the typical results and pre-empt this by not funding it), lack of daring (this research requires making many assumptions, and taking on well-entrenched incumbents), and lack of reward (the economics profession frowns on the required interdisciplinarity). In addition, many people, including many economists, would argue that climate change is beyond cost-benefit analysis and that monetary valuation is unethical.

Table 1 shows some characteristics of these studies. A few insights emerge. First, the welfare impact of a doubling of the atmospheric concentration of carbon dioxide on the current economy is relatively small. Although the estimates differ, impacts are not more than a few percent of GDP. The estimates of Hope [50], Mendelsohn et al. [66,68] and Tol [105] even point to initial benefits of climate change. (Studies published after 1995 all have regions with net gains and net losses due to global warming, whereas earlier studies only find net losses.) With such estimates, it is no surprise that cost-benefit analyses of climate change recommend only limited greenhouse gas emission reduction – for instance, Nordhaus [75] argues that the optimal rate of emission reduction is 10–15%, one of the more contentious findings of the climate economics literature.

Second, although the impact is relatively small, it is not negligible. A few per cent of GDP in annual damage is a real concern.

Third, climate change may initially have positive impacts. This is partly because the higher ambient concentration of carbon dioxide would reduce water stress in plants and may make them grow faster – although this effect is now believed to be weaker [58]. Another reason is that the global economy is concentrated in the temperate zone, where a bit of warming may well be welcomed because of reductions in heating costs and cold-related health problems. At the same time, the world population is concentrated in the tropics, where the impacts of initial climate change are probably negative. Even though initial *economic* impacts are positive, it does not necessarily follow that greenhouse gas emissions should be subsidized. The climate responds rather slowly to changes in emissions, so the initial impacts cannot be avoided. Impacts start falling – that is, additional climate change reduces global welfare – roughly at the same time as climate change can be influenced by present and future emission reduction [47].

The fourth insight is that relative impacts are higher in poorer countries (see also Yohe and Schlesinger [122]). This is because poorer countries have a lower adaptive capacity [2], particularly in health [108], and have a greater exposure to climate change, particularly in agriculture and water resources. Furthermore, poorer countries tend to be hotter and therefore closer to temperature limits and short on spatial analogues should it get warmer still. At the same time, there are fewer studies on the impacts of climate change on developing countries than on developed countries. Although research is scarce [83], there is little reason to assume that climate change impacts would be homogeneous within countries; certainly, certain economic sectors (e. g., agriculture), regions (e. g., the coastal zone) and age groups (e. g., the elderly) are more heavily affected than

Climate Change, Economic Costs of, Table 1

Econonic impact estimates of climate change; numbers in brackets are either standard deviations or confidence intervals

Study	Warming	Impact	Minimum	Region	Maximum	Region
Nordhaus [76]	3.0	−1.3				
Nordhaus [77]	3.0	−4.8 (−30.0 to 0.0)				
Fankhauser [33]	2.5	−1.4	−4.7	China	−0.7	Eastern Europe and the former Soviet Union
Tol [103]	2.5	−1.9	−8.7	Africa	−0.3	Eastern Europe and the former Soviet Union
Nordhaus and Yang [82][a]	2.5	−1.7	−2.1	Developing countries	0.9	Former Soviet Union
Plambeck and Hope [60][a]	2.5	−2.5 (−0.5 to −11.4)	−8.6 (−0.6 to −39.5)	Asia (w/o China)	0.0 (−0.2 to 1.5)	Eastern Europe and the former Soviet Union
Mendelsohn et al. [66][a,b,c]	2.5	0.0 0.1	−3.6 −0.5	Africa	4.0 1.7	Eastern Europe and the former Soviet Union
Nordhaus and Boyer [81]	2.5	−1.5	−3.9	Africa	0.7	Russia
Tol [105]	1.0	2.3 (1.0)	−4.1 (2.2)	Africa	3.7 (2.2)	Western Europe
Maddison [59][a,d,e]	2.5	−0.1	−14.6	South America	2.5	Western Europe
Rehdanz and Maddison [92][a,c]	1.0	−0.4	−23.5	Sub-Saharan Africa	12.9	South Asia
Hope [50] [a]	2.5	0.9 (−0.2 to 2.7)	−2.6 (−0.4 to 10.0)	Asia (w/o China)	0.3 (−2.5 to 0.5)	Eastern Europe and the former Soviet Union
Nordhaus [78]	2.5	−0.9 (0.1)				

[a] Note that the global results were aggregated by the current author.
[b] The top estimate is for the "experimental" model, the bottom estimate for the "cross-sectional" model.
[c] Note that Mendelsohn et al. only include market impacts.
[d] Note that the national results were aggregated to regions by the current author for reasons of comparability.
[e] Note that Maddison only considers market impacts on households.

others. This has two policy implications. Firstly, recall that greenhouse gas mix uniformly in the atmosphere. It does not matter where they are emitted or by whom, the effect on climate change is the same. Therefore, any justification of stringent emission abatement is an appeal to consider the plight of the poor and the impacts imposed on them by the rich [94,95]. While this makes for wonderful rhetoric and fascinating research (e. g., [104]), reality shows little compassion for the poor by the rich. Secondly, if poverty is the root cause for vulnerability to climate change, one may wonder whether stimulating economic growth or emission abatement is the better way to reduce impacts. Indeed, Tol and Yohe [115] argue that the economic growth foregone by stringent abatement more than offsets the avoided impacts of climate change, at least for malaria, while Tol [108] shows that development is a cheaper way of reducing climate-change-induced malaria than is emission reduction. Moreover, richer countries may find it easier and cheaper to compensate poorer countries for the climate change damages caused, than to reduce greenhouse gas emissions. Such compensation may be explicit and financial, but would more likely take the shape of technical and financial assistance with adaptation (cf. [85]).

The agreement between the studies is remarkable if one considers the diversity in methods. The studies of Fankhauser, Hope, Nordhaus, and Tol all use the enumerative method: 'physical' impact estimates are obtained one by one, from 'natural science' papers based on 'process-based' models or 'laboratory experiments'. These physical impacts are multiplied with their respective prices, and added up. The 'prices' are obtained by benefit transfer. In contrast, Mendelsohn's work is based on direct, empirical estimates of the welfare impacts, using observed variations in prices and expenditures to discern the effect of climate (e. g., [67]). Mendelsohn estimates are done per sector and then added up, but physical modelling and benefit transfer are avoided. Nordhaus [78] uses empirical estimates of the *aggregate* climate impact on income, while Maddison [59] looks at patterns of *aggregate* household consumption. Like Mendelsohn, Nordhaus and Maddison rely exclusively on observations, but they assume that all climate effects are aggregated by the economy into incomes and expenditures. Rehdanz and Maddison [92] also empirically estimate the aggregate impact, but use self-reported happiness as an indicator; their approach is similar to that of Nordhaus and Maddison, but the indicator is subjective rather than objective. The enumerative studies of Fankhauser etc rely on controlled experiments (albeit with detailed, process-based models in most cases). This has the advantages of ease of interpretation and physical realism, but the main disadvantage is that certain things are kept constant that would change in reality; adaptation is probably the key element. The statistical studies of Mendelsohn etc rely on uncontrolled experiments. This has the advantage that everything varies as in reality, but the disadvantages are that the assessment is limited to observed variations (which may be small compared to projected changes, particularly in the case of carbon dioxide concentration) and that effects may be spuriously attributed to climate. Therefore, the variety of methods enhances confidence, not in the individual estimates, but in the average.

The shortcomings of the estimates are at least as interesting. Welfare losses are approximated with direct costs, ignoring general equilibrium and even partial equilibrium effects (see below). In the enumerative studies, impacts are assessed independently of one another, even if there is an obvious overlap as between water resources and agriculture. Estimates are often based on extrapolation from a few detailed case studies, and extrapolation is to climate and levels of development that are very different from the original case study. Valuation is based on benefit transfer, driven only by difference in per capita income. Realistic modelling of adaptation is problematic, and studies either assume no adaptation or perfect adaptation. Many impacts are unquantified, and some of these may be large (see below). The uncertainties are unknown – only 4 of the 14 estimates in Table 1 have some estimate of uncertainty. These problems are gradually solved, but progress is slow. Indeed, the above list of caveats is similar to that in Fankhauser and Tol [34].

Marginal Costs

Although the number of studies of the *total* costs of climate change is small, a larger number of studies estimate the *marginal* costs. The marginal damage cost of carbon dioxide is defined as the net present value of the incremental damage due to an infinitesimally small increase in carbon dioxide emissions. If this is computed along the optimal trajectory of emissions, the marginal damage cost equals the Pigou tax. Marginal damage cost estimates derive from total cost estimates – the fact that there are more estimates available, does not imply that we know more about the marginal costs than we do about the total costs. In fact, some of the total cost estimates [59,66,68,78,92] have yet to be used for marginal cost estimation, so that the empirical basis is actually smaller.

Tol [110] gathers 211 estimates of the SCC from 47 studies. The studies were grouped in those that were peer-reviewed and those that were not. Some studies are based on original estimates of the total costs of climate change, while other studies borrow total costs estimates from other studies. Most studies use incremental or marginal calculus to estimate the SCC, as they should, while a few others use average impacts or an unspecified method. Some studies assume that climate changes but society does not, while other studies include a dynamic model of vulnerability. A few studies use entirely arbitrary assumptions about future climate change, while most studies are based on internally consistent scenarios. These classifications are used as quality indicators. More recent studies were given a higher weight. Many studies report multiple estimates. Most of the estimates are sensitivity analyses around a central estimate, and some estimates are only included to (approximately) reproduce an earlier study. Tol [110] introduces additional weights to account for this.

Tol [110] adjusts a Fisher–Tippett kernel density estimator to 211 data points, weighted as describe above. The 211 estimates provide the modes. Only a few of the studies provide an estimate of the uncertainty. Therefore, the standard deviation is set equal to the sample standard deviation.

Table 2 shows selected characteristics of the kernel distribution for the whole sample and selected sub-samples.

Splitting the sample by discount rate used has the expected effect: A higher discount rate implies a lower estimate of the SCC and a thinner tail. Table 2 also shows that estimates in the peer reviewed literature are lower and less uncertain than estimates in the gray literature.

Splitting the sample by publication date, shows that the estimates of the SCC published before AR2 [87] were larger than the estimates published between AR2 and AR3 [99], which in turn were larger than the estimates published since. Note that these differences are not statistically significant if one considers the means and standard deviation. However, the kernel distribution clearly shifts to the left. Therefore, AR4 [96] were incorrect to conclude that the economic estimates of the impact of climate change have *increased* since 2001. In their words (pp. 781): "There is some evidence that initial new market benefits from climate change will peak at a lower magnitude and sooner than was assumed for the TAR, and it is likely that there will be higher damages for larger magni-

Climate Change, Economic Costs of, Table 2

Selected characteristics (mode, mean, standard deviation, median, 90-percentile, 95-percentile, 99-percentile, percentile of the Stern estimate) of the joint probability density of the social cost of carbon for the whole sample (all) and selected subsamples (pure rate of time preference, review process, and publication date)

	All	PRTP			Review		Publication date		
		0%	1%	3%	peer	gray	<1996	96–01	>2001
Mode	35	129	56	14	20	53	36	37	27
Mean	127	317	80	24	71	196	190	120	88
St.Dev.	243	301	70	21	98	345	392	179	121
Median	74	265	72	21	48	106	88	75	62
90%	267	722	171	51	170	470	397	274	196
95%	453	856	204	61	231	820	1555	482	263
99%	1655	1152	276	82	524	1771	1826	867	627
Stern	0.92	0.56	1.00	1.00	0.97	0.84	0.86	0.92	0.96

tudes of global mean temperature increases than was estimated in the TAR.” It is unclear how Schneider et al. [96] reached this conclusion, but it is not supported by the data presented here.

The SCC estimate by Stern et al. [100] is almost an outlier in the entire sample (excluding, of course, the Stern estimate itself). Depending on the kernel density, the Stern estimate lies between the 90th and the 94th percentile. It fits in better with estimates that use a low discount rate and were not peer-reviewed – characteristics of the Stern Review – but even in comparison to those studies, Stern et al. [100] are on the high side. The Stern estimate also fits in better with the older studies. This is no surprise, as the PAGE model (e. g., [50]) is calibrated to [87] and [99]. Other criticism of the Stern Review can be found in [24,65,79,80,115,119,120].

Policy Implications

The policy implications of the above findings are several, and not necessarily in line with the conceived wisdom of climate policy. First and foremost, the economic impacts literature points out that climate change is a problem. Initial climate change may be beneficial, but it cannot be avoided. This is a sunk benefit. The avoidable part of climate change is in all likelihood negative. This justifies greenhouse gas emission reduction.

Second, the estimates of the marginal damage costs justify some emission abatement, but not too much. For instance, the future price of carbon dioxide emission permits in the European Trading System is around $ 100/tC. Using the market rate of discount, the expected social cost of carbon is only $ 24/tC. This climate policy has a benefit-cost ratio of 0.24. EU climate policy is therefore too stringent. Of course, European climate policy does pass the cost-benefit test according to CEC [17], but this study does not meet conventional standards of academic quality [111]. Earlier, Pearce [89] similarly concluded that the UK cost-benefit analysis [20] is deficient, while also the *Stern Review* [100] has been criticized in the academic literature (e. g., [80,119]).

Third, climate policy is about ethics rather than about economics [37,116]. The judgment what to do about greenhouse gas emissions rests on the values one attaches to far-flung countries and distant futures. The ethics are not straightforward, however. If one places a lot of weight on the future, one should make a trade-off between increasing investment in capital goods, education, emission reduction, or technology. If one places a lot of weight on people in poor countries, one should make a trade-off between adaptation, development, emission reduction, and trade reform.

Fourth, the uncertainties about the economic impact of climate change are profound. Partly, this is because the subject is complex. A large share of the uncertainty can be explained, however, by the dearth of research funding. Although climate change is often said to be the largest (environmental) challenge of our times, very few researchers are funded to substantiate or refute that claim.

Future Directions

Further research is therefore needed. Several problems with past and present research are identified above. Firstly, research into the economic impact of climate change is rightly classified as “applied research”. This implies, however, that research funding comes from bodies with a stake in the result, and that quality and independence are not necessarily overriding concerns. The *Stern Review* is the most prominent example in the recent past of a study that started with the conclusions and worked back to identify the required assumptions. The new *Centre for Climate*

Change Economics and Policy at the *London School of Economics* may fall into the same trap. Because the stakes in climate policy are large, academic quality of research must be guaranteed.

Secondly, there are only two groups of independent academics who study the economic impact of climate change. These groups are not sufficiently funded. More importantly, these groups are rarely challenged. Combined with the first problem, it is therefore important to establish a third group of independent, academic economists to study the impact of climate change.

Thirdly, research on the impact of climate change, economic and otherwise, has been lamp-posting. After the groundbreaking work in the early 1990s, researchers have refined previous estimates. Little attention has been paid to those impacts for which no previous estimates exist. While this is the normal procedure of gradual progress in scientific research, the study of the impact of climate change is still in its formative stages. Not just *more*, but particularly *different* research is needed – into the economic effects of climate change on biodiversity, on violent conflict, on ice shelves and ocean current, and on economic development in the long term.

Acknowledgments

My thoughts on this subject were shaped by discussions with Hadi Dowlatabadi, Tom Downing, Sam Fankhauser, David Maddison, Rob Mendelsohn, Bill Nordhaus, David Pearce, Steve Schneider, Joel Smith, and Gary Yohe.

Bibliography

1. Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. Am Econ Rev 91:1369–1401
2. Adger WN (2006) Vulnerability. Glob Environ Chang 16:268–281
3. Arrow KJ, Cline WR, Maeler KG, Munasinghe M, Squitieri R, Stiglitz JE (1996) Intertemporal equity, discounting, and economic efficiency. In: Bruce JP, Lee H, Haites EF (eds) Climate change 1995: economic and social dimensions – contribution of working group iii to the second assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 125–144
4. Ashley RM, Balmfort DJ, Saul AJ, Blanskby JD (2005) Flooding in the future – predicting climate change, risks and responses in urban areas. Water Sci Technol 52(5):265–273
5. Ayres RU, Walter J (1991) The greenhouse effect: damages, costs and abatement. Environ Resour Econ 1:237–270
6. Azar C, Sterner T (1996) Discounting and distributional considerations in the context of global warming. Ecol Econ 19:169–184
7. Berrens RP, Bohara AK, Jenkins-Smith HC, Silva CL, Weimer DL (2004) Information and effort in contingent valuation surveys: application to global climate change using national internet samples. J Environ Econ Manag 47:331–363
8. Berrittella M, Bigano A, Roson R, Tol RSJ (2006) A general equilibrium analysis of climate change impacts on tourism. Tour Manag 27(5):913–924
9. Bosello F, Roson R, Tol RSJ (2006) Economy-wide estimates of the implications of climate change: human health. Ecol Econ 58:579–591
10. Bosello F, Roson R, Tol RSJ (2007) Economy-wide estimates of the implications of climate change: sea level rise. Environ Resour Econ 37:549–571
11. Breslow PB, Sailor DJ (2002) Vulnerability of wind power resources to climate change in the continental united states. Renew Energy 27(4):585–598
12. Broome J (1992) Counting the cost of global warming. White Horse Press, Cambridge
13. Burkett VR, Wilcox DA, Stottlemyer R, Barrow W, Fagre D, Baron J, Price J, Nielson JL, Allen CD, Peterson DL, Ruggerone G, Doyle T (2005) Nonlinear dynamics in ecosystem response to climate change: case studies and policy implications. Ecol Complex 2(4):357–394
14. Burton I, Kates RW, White GF (1993) The environment as hazard, 2nd edn. The Guilford Press, New York
15. Carmichael CG, Gallus Jr WA, Temeyer BR, Bryden MK (2004) A winter weather index for estimating winter road maintenance costs in the midwest. J Appl Meteorol 43(11):1783–90
16. CBO (2005) Uncertainty in analyzing climate change: policy implications. congress of the united states. Congressional Budget Office, Washington
17. CEC (2005) Winning the battle agains global climate change – background paper. Commission of the European Communities, Brussels
18. Champ PA, Boyle KJ, Brown TC (eds) (2003) A primer on nonmarket valuation. Kluwer, Dordrecht
19. Chichilnisky G (1996) An axiomatic approach to sustainable development. Soc Choice Welf 13(2):219–248
20. Clarkson R, Deyes K (2002) Estimating the social cost of carbon emissions. Working Paper 140. The Public Enquiry Unit – HM Treasury, London
21. Cline WR (1992) The economics of global warming. Institute for International Economics, Washington
22. Cropper ML, Aydede SK, Portney PR (1992) Rates of time preference for saving lives. Am Econ Rev 82(2):469–472
23. Darwin RF (1999) A FARMer's view of the ricardian approach to measuring agricultural effects of climatic change. Clim Chang 41(3–4):371–411
24. Dasgupta P (2007) Commentary: The stern review's economics of climate change. Natl Inst Econ Rev 199:4–7
25. Dasgupta P, Maskin E (2005) Uncertainty and hyperbolic discounting. Am Econ Rev 95(4):1290–1299
26. Davidson MD (2006) A social discount rate for climate damage to future generations based on regulatory law. Clim Chang 76:55–72
27. Dorland C, Tol RSJ, Palutikof JP (1999) Vulnerability of the netherlands and northwest europe to storm damage under climate change. Clim Chang 43:513–535
28. Easterly W, Levine R (2003) Tropics, germs, and crops: how endowments influence economic development. J Monet Econ 50:3–39
29. Elsasser H, Buerki R (2002) Climate change as a threat to tourism in the alps. Clim Res 20(3):253–257

30. Ennos AR (1997) Wind as an ecological factor. Trends Ecol Evol 12(3):108–111
31. Evans DJ, Sezer H (2004) Social discount rates for six major countries. Appl Econ Lett 11:557–560
32. Fankhauser S (1994) The economic costs of global warming damage: a survey. Glob Environ Chang 4(4):301–309
33. Fankhauser S (1995) Valuing climate change – the economics of the greenhouse, 1st edn. EarthScan, London
34. Fankhauser S, Tol RSJ (1996) Recent advancements in the economic assessment of climate change costs. Energy Policy 24(7):665–673
35. Fankhauser S, Tol RSJ (2005) On climate change and economic growth. Resour Energy Econ 27:1–17
36. Fankhauser S, Tol RSJ, Pearce DW (1997) The aggregation of climate change damages: a welfare theoretic approach. Environ Resour Econ 10:249–266
37. Gardiner SM (2006) A perfect moral storm: climate change, intergenerational ethics and the problem of moral corruption. Environ Values 15:397–413
38. Gerlagh R, Keyzer MA (2001) Sustainability and the intergenerational distribution of natural resource entitlements. J Public Econ 79:315–341
39. Gollier C (2002) Discounting an uncertain future. J Public Econ 85:149–166
40. Gollier C (2002) Time horizon and the discount rate. J Econ Theor 107:463–473
41. Guo JK, Hepburn C, Tol RSJ, Anthoff D (2006) Discounting and the social cost of carbon: a closer look at uncertainty. Environ Sci Policy 9(5):203–216
42. Hamilton JM, Maddison DJ, Tol RSJ (2005) Climate change and international tourism: a simulation study. Glob Environ Chang 15(3):253–266
43. Hamilton JM, Maddison DJ, Tol RSJ (2005) The effects of climate change on international tourism. Clim Res 29:255–268
44. Harrison GP, Wallace AR (2005) Sensitivity of wave energy to climate change. IEEE Trans Energy Convers 20(4):870–877
45. Harvey D, Huang Z (1995) Evaluation of the potential impact of methane clathrate destabilization on future global warming. J Geophys Res 100:2905–2926
46. Hersch J, Viscusi WK (2006) The generational divide in support for environmental policies: european evidence. Clim Chang 77:121–136
47. Hitz S, Smith JB (2004) Estimating global impacts from climate change. Glob Environ Chang 14:201–218
48. Hohmeyer O, Gaertner M (1992) The costs of climate change – a rough estimate of orders of magnitude. Fraunhofer-Institut fur Systemtechnik und Innovationsforschung, Karlsruhe
49. Homer-Dixon TF (1994) Environmental scarcities and violent conflict: evidence from cases. Int Secur 19(1):5–40
50. Hope CW (2006) The marginal impact of co2 from PAGE2002: an integrated assessment model incorporating the IPCC's five reasons for concern. Integr Assess J 6(1):19–56
51. Hope CW, Maul P (1996) Valuing the impact of CO_2 emissions. Energy Policy 24(3):211–219
52. Kikkawa T, Kita J, Ishumatsu A (2004) Comparison of the lethal effect of CO_2 and acidification on red sea bream (pagrus major) during the early developmental stages. Marine Pollut Bull 48(1–2):108–110
53. Kundzewicz ZW, Gracyk D, Maurer T, Pinskwar I, Radziejewski M, Svensson C, Szwed M (2005) Trend detection in river flow series: 1, annual maximum flow. Hydrol Sci J 50(5):797–810
54. Li H Berrens RP, Bohara AK, Jenkins-Smith HC, Silva CL, Weimer DL (2004) Exploring the beta model using proportional budget information in a contingent valuation study. Ecol Econ 28:329–343
55. Lind RC, Schuler RE (1998) Equity and discounting in climate change decisions. In: Nordhaus WD (ed) Economics and policy issues in climate change. Resources for the Future, Washington, pp 59–96
56. Lind RC (1995) Intergenerational equity, discounting, and the role of cost-benefit analysis in evaluating global climate policy. Energy Policy 23(4/5):379–389
57. Link PM, Tol RSJ (2004) Possible economic impacts of a shutdown of the thermohaline circulation: an application of FUND. Port Econ J 3:99–114
58. Long SP, Ainsworth EA, Leakey ADB, Noesberger J, Ort DR (2006) Food for thought: lower-than-expected crop yield stimulation with rising CO2 concentrations. Science 312: 1918–1921
59. Maddison DJ (2003) The amenity value of the climate: the household production function approach. Resour Energy Econ 25:155–175
60. Manne AS, Richels RG (2004) US rejection of the kyoto protocol: the impact on compliance costs and CO_2 emissions. Energy Policy 32:447–454
61. Marotzke J (2000) Abrupt climate change and thermohaline circulation: mechanisms and predictability. Proc Natl Acad Sci 97:1347–1350
62. Masters WA, McMillan MS (2001) Climate and scale in economic growth. J Econ Growth 6:167–186
63. McDonald RE, Bleaken DG, Cresswell DR, Pope VD, Senior CA (2005) Tropical storms: representation and diagnosis in climate models and the impacts of climate change. Clim Dyn 25(1):19–36
64. McLeman R, Smit B (2006) Migration as an adaptation to climate change. Clim Chang 76:31–53
65. Mendelsohn RO (2006) A critique of the stern report. Regulation (Winter 2006–2007):42–46
66. Mendelsohn RO, Morrison W, Schlesinger ME, Andronova NG (2000) Country-specific market impacts of climate change. Clim Chang 45:553–569
67. Mendelsohn RO, Nordhaus WD, Shaw D (1994) The impact of climate on agriculture: a ricardian analysis. Am Econ Rev 84(4):753–771
68. Mendelsohn RO, Schlesinger ME, Williams LJ (2000) Comparing impacts across climate models. Int Assess 1:37–48
69. Mulcahy M (2004) Weathering the storms: hurricanes and risk in the british greater caribbean. Bus Hist Rev 78(4):635–663
70. Nakicenovic N, Swart RJ (eds) (2001) IPCC special report on emissions scenarios. Cambrigde University Press, Cambridge
71. Newell RG, Pizer WA (2003) Discounting the distant future: how much do uncertain rates increase valuations? J Environ Econ Manag 46:52–71
72. Nicholls RJ, Tol RSJ (2006) Impacts and responses to sea-level rise: a global analysis of the SRES scenarios over the 21st Century. Phil Trans Royal Soc A Math Phys Eng Sci 361(1841):1073–1095
73. Nordhaus WD (1982) How fast should we graze the global commons? Am Econ Rev 72(2):242–246
74. Nordhaus WD (1991) To Slow or Not to Slow: The economics of the greenhouse effect. Econ J 101:920–937
75. Nordhaus WD (1993) Rolling the 'DICE': An optimal transition

path for controlling greenhouse gases. Resour Energy Econ 15:27–50

76. Nordhaus WD (1994) Managing the global commons: the economics of climate change. The MIT Press, Cambridge
77. Nordhaus WD (1994) Expert opinion on climate change. Am Sci 82(1):45–51
78. Nordhaus WD (2006) Geography and macroeconomics: new data and new findings. Proc Natl Acad Sci 103(10):3510–3517. www.pnas.org/cgi/doi/10.1073/pnas.0509842103
79. Nordhaus WD (2007) Critical assumptions in the stern review on climate change. Science 317:201–202
80. Nordhaus WD (2007) A review of the stern review on the economics of climate change. J Econ Lit 45(3):686–702
81. Nordhaus WD, Boyer JG (2000) Warming the world: economic models of global warming. The MIT Press, Cambridge
82. Nordhaus WD, Yang Z (1996) RICE: A regional dynamic general equilibrium model of optimal climate-change policy. Am Econ Rev 86(4):741–765
83. O'Brien KL, Sygna L, Haugen JE (2004) Vulnerable or resilient? A multi-scale assessment of climate impacts and vulnerability in norway. Clim Chang 64:193–225
84. Oppenheimer M, Alley RB (2005) Ice sheets, global warming, and article 2 of the UNFCCC. Clim Chang 68:257–267
85. Paavola J, Adger WN (2006) Fair adaptation to climate change. Ecol Econ 56:594–609
86. Pearce DW (2003) The social cost of carbon and its policy implications. Oxford Rev Econ Policy 19(3):1–32
87. Pearce DW, Cline WR, Achanta AN, Fankhauser S, Pachauri RK, Tol RSJ, Vellinga P (1996) The social costs of climate change: greenhouse damage and the benefits of control. In: Bruce JP, Lee H, Haites EF (eds) Climate Change 1995: Economic and Social Dimensions – Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 179–224
88. Pearce DW, Moran D (1994) The economic value of biodiversity. EarthScan, London
89. Pielke Jr RA, Landsea C, Mayfield M, Laver J, Pasch R (2005) Hurricanes and global warming. Bull Am Meteorol Soc 86(11):1571–1575
90. Plambeck EL, Hope CW (1996) PAGE95 – An updated valuation of the impacts of global warming. Energy Policy 24(9):783–793
91. Ready R, Navrud S, Day B, Dubourg R, Machado F, Mourato S, Spaninks F, Rodriguez MXV (2004) Benefit transfer in europe: How reliable are transfers between countries? Environ Resour Econ 29(1):67–82
92. Rehdanz K, Maddison DJ (2005) Climate and happiness. Ecol Econ 52:111–125
93. Schelling TC (1984) Choice and consequence. Harvard University Press, Cambridge
94. Schelling TC (1992) Some economics of global warming. Am Econ Rev 82:1–14
95. Schelling TC (1995) Intergenerational discounting. Energy Policy 23(4/5):395–401
96. Schneider SH, Semenov S, Patwardhan A, Burton I, Magadya CHD, Oppenheimer M, Pittock AB, Rahman A, Smith JB, Suarez A, Yamin F (2007) Assessing key vulnerability and the risk from climate change. In: Parry ML et al (eds) Climate Change 2007: Impacts, Adaptation and Vulnerability – Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 779–810
97. Scott D, McBoyle G, Mills B (2003) Climate change and the skiing industry in southern ontario (Canada): exploring the importance of snowmaking as a technical adaptation. Clim Res 23:171–181
98. Smith JB (1996) Standardized estimates of climate change damages for the united states. Clim Chang 32(3):313–326
99. Smith JB, Schellnhuber HJ, Mirza MMQ, Fankhauser S, Leemans R, Lin E, Ogallo L, Pittock B, Richels RG, Rosenzweig C, Tol RSJ, Weyant JP, Yohe GW (2001) Vulnerability to climate change and reasons for concern: a synthesis. In: Mccarthy JJ, Canziani OF, Leary NA, Dokken DJ, White KS (eds) Climate change 2001: Impacts, adaptation, and vulnerability, vol 19. Cambridge University Press, Cambridge, pp 913–967
100. Stern NH, Peters S, Bakhshi V, Bowen A, Cameron C, Catovsky S, Crane D, Cruickshank S, Dietz S, Edmonson N, Garbett SL, Hamid L, Hoffman G, Ingram D, Jones B, Patmore N, Radcliffe H, Sathiyarajah R, Stock M, Taylor C, Vernon T, Wanjie H, Zenghelis D (2006) Stern review: The economics of climate change. Cambridge University Press, Cambridge
101. Szolnoky C, Buzas K, Clement A (1997) Impacts of the climate change on the operation of a freshwater cooled electric power plant. Periodica Polytecnica: Civil Eng 41(2):71–94
102. Titus JG (1992) The costs of climate change to the united states. In: Majumdar SK et al (eds) Global climate change: implications, challenges and mitigation measures. Pennsylvania Academy of Science, Easton, pp 384–409
103. Tol RSJ (1995) The damage costs of climate change – towards more comprehensive calculations. Environ Resour Econ 5:353–374
104. Tol RSJ (2001) Equitable cost-benefit analysis of climate change. Ecol Econ 36(1):71–85
105. Tol RSJ (2002) New estimates of the damage costs of climate change, Part I: Benchmark Estimates. Environ Resour Econ 21(1):47–73
106. Tol RSJ (2003) Is the uncertainty about climate change too large for expected cost-benefit analysis? Clim Chang 56(3):265–289
107. Tol RSJ (2005) The marginal damage costs of carbon dioxide emissions: an assessment of the uncertainties. Energy Policy 33(16):2064–2074
108. Tol RSJ (2005) Emission abatement versus development as strategies to reduce vulnerability to climate change: an application of FUND. Environ Dev Econ 10:615–629
109. Tol RSJ (2005) Adaptation and mitigation: trade-offs in substance and methods. Environ Sci Policy 8:572–578
110. Tol RSJ (2007) The social cost of carbon: trends, outliers and catastrophes, research unit sustainability and global change FNU-144. Hamburg University and Centre for Marine and Atmospheric Science, Hamburg
111. Tol RSJ (2007) Europe's long-term climate target: A critical evaluation. Energy Policy 35:424–432
112. Tol RSJ, van der Grijp NM, Olsthoorn AA, van der Werff PE (2003) Adapting to climate change: A case study of riverine flood risks in the netherlands. Risk Analysis 23(3):575–583
113. Tol RSJ, Verheyen R (2004) State responsibility and compensation for climate change damages – a legal and economic assessment. Energy Policy 32:1109–1130
114. Tol RSJ, Yohe GW (2006) Of dangerous climate change and dangerous emission reduction. In: Schellnhuber HJ, Cramer

W, Nakicenovic N, Wigley T, Yohe G (eds) Avoiding dangerous climate change. Cambridge University Press, Cambridge, pp 291–298

115. Tol RSJ, Yohe GW (2006) A review of the stern review. World Econ 7(4):233–250

116. Toman M (2006) Values in the economics of climate change. Environ Values 15:365–379

117. Viscusi WK, Zeckhauser RJ (2006) The perception and valuation of the risks of climate change: A rational and behavioral blend. Clim Chang 77:151–177

118. Weitzman ML (2001) Gamma discounting. Am Econ Rev 91(1):260–271

119. Weitzman ML (2007) A review of the stern review on the economics of climate change. J Econ Lit 45(3):703–724

120. Weitzman ML (2008) On modeling and interpreting the economics of catastrophic climate change. Rev Econ Stat

121. Wilson KJ, Falkingham J, Melling H, de Abreu R (2004) Shipping in the canadian arctic: Other possible climate change scenarios. Int Geosci Remote Sens Symp 3:1853–1856

122. Yohe GW, Schlesinger ME (2002) The economic geography of the impacts of climate change. J Econ Geogr 2:311–341

123. Yohe GW, Tol RSJ (2002) Indicators for social and economic coping capacity – moving towards a working definition of adaptive capacity. Glob Environ Chang 12(1):25–40

124. Zhang DD, Jim CY, Lin GCS, He YQ, Wang JJ, Lee HF (2006) Climatic change, wars and dynastic cycles in China over the last millennium. Clim Chang 76:459–477

Climate Change and Human Health

HARTMUT GRASSL
Max Planck Institute for Meteorology,
Hamburg, Germany

Article Outline

Glossary

Health As defined by the World Health Organization (WHO) health is the state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.

Human bioclimate The fundamental issue in human biometeorology is the assessment of the direct health effects of the atmospheric environment from heat exchange, to solar radiation and air pollution.

Climate change related direct health effects Climate change always impacts on human bioclimate, presently it leads to increased heat stress and heat stress frequency, higher ultraviolet radiation doses especially in summer, longer allergic pollen seasons and new allergens as well as intensified photo-smog.

Thermal stress and mortality Summer heat waves in mid-latitudes and elsewhere increase without doubt mortality; hence also highlight lack of correct adaptive measures, i. e. heat waves impact most strongly in societies with lack of social cohesion.

Global expansion of tropical diseases The observed recent global warming has increased the incidence and enlarged the distribution of some tropical diseases due to the expansion of suitable conditions for both vectors and pathogens. A northward spread has been observed for West Nile fever, Leishmaniasis and Chikungunya fever and a climate-driven spread has in parts also been recorded for malaria, dengue fever and other vector-borne infectious diseases.

Vector-borne diseases In epidemiology a vector is an organism transmitting a pathogen from one of its reservoirs (e. g. ruminants, birds) to another one (e. g. human) without falling ill. Such vectors for tropical diseases are: mosquitoes, biting flies, bugs, lice, flea's and mites. Typical vector-borne diseases are malaria, yellow fever, dengue fever, West Nile fever, Leishmaniasis, Chikungunya fever. For some of these diseases global warming is the cause of the observed expansion or intensification. The complex web of reservoir organism, pathogens, vectors and infected organisms with different dependence on climate parameters often hinders a full understanding. Hence, surprises are and will be common.

Arbo viruses Arbo viruses are transmitted by arthropods (**ar**thropod-**bo**rne) to vertebrates and hence in parts also to humans. Besides yellow fever, tick borne encephalitis and dengue fever about 150 other diseases are due to virus infections by insects and spiders (arthropods). In very complex transmission cycles climatic conditions play a central role. The occurrence of unusual arbo virus infections is often related to changes in climatic conditions. Therefore, the partly dramatic global increase of some arbo virus infections is also driven at least in part by the ongoing global anthropogenic climate change.

Arbo viruses transmitted by Aedes mosquitoes Aedes mosquitoes and Aedes-transmitted arbo viruses such as the dengue and yellow fever viruses are a growing global threat. The primary vector of these diseases, *Aedes aegypti*, has re-emerged throughout the tropics, but also *Aedes albopictus* has emerged as one of the worst invasive species taking the role of *Aedes aegypti*. Direct human activities like global trade are mainly responsible for the spread of these vectors and global warming – indirectly anthropogenic as well – cannot be ruled out as a contributor. With further warming temperate regions like Central Europe could also become areas for *Aedes albopictus*.

Malaria and global warming Although the Anopheles mosquitoes transmitting the protozoae (e. g. *Plasmodium falciparum*) causing malaria are strongly dependent on temperature and suitable small water reservoirs for the larvae, the spread of malaria in recent years is more a consequence of deficiencies in public health systems of many countries rather than due to the observed global warming and concomitant precipitation changes.

Blue-tongue disease in Europe Since August 2006 the blue-tongue-disease of cattle and sheep (serotype 8 from South Africa) spread within months from the Netherlands to Belgium, Luxembourg, France and Germany alone at the end of 2006. The vector carrying the virus is the ceratopogonid, biting midge, *Culicoides obsoletus*. The new disease for ruminants in Western and Central Europe is primarily a consequence of globalization but the extremely warm winter 2006/2007 in Western and Central Europe supported further spreading.

Carbon dioxide fertilization and quality of food From field experiments at elevated carbon dioxide concentrations (close to a doubling of preindustrial values) it is known that agricultural yield increases for C3 plants by 10 to 30 percent; however, frequently also reduced nitrogen content in plant tissue including seeds is observed. Hence, food quality may be lowered.

Changes in the pollen season and new pollen The onset of flowering of plants in mid and high latitudes is mainly triggered by temperature. Therefore, warming in recent decades has caused an earlier start of pollen in the air, often also leading to a longer pollen season and for some pollen also to higher abundance stimulated by higher carbon dioxide concentration. Hence, susceptible individuals will suffer from pollen allergy longer and even perennial allergic symptoms may become possible.

Introduction

The basis of our life is energy from the sun, water from the skies and biomass production by plants on land and in the oceans. If we ask for the key climate parameters – given the size of the planet and its mean distance to the sun – we get a very similar answer: energy flux density of the sun, precipitation and land surface parameters, mostly determined by vegetation. Hence, climate is the key natural resource. If this resource changes rapidly, as it does now, life in all its forms is affected as well. Therefore, it is a must for decision makers to deal with climate change. Here only one facet of climate change is discussed: the health of humans. It is rather astonishing that it has not already been studied intensively since the beginning of the climate change debate, as other environmental policy decisions were nearly totally driven by health consequences for humans [12].

The consequences of climate change for the health of humans, animals and plants are complex with direct and indirect relations between causes and impacts. As any organism on land is in permanent struggle with the local weather a rather strong capacity for adaptation exists; however, the organisms are trained only with existing weather and climate variability during the life-time of an organism. For new weather extremes accompanying any climate change this adaptation capability is no longer given. Hence, a changing climate will often become a threat to health. As several health related factors may change simultaneously, "multiplying" impacts, which – if alone – would not have gone beyond existing adaptive capability, may do so. An example is heat stress during heat waves accompanied not only by high ultraviolet radiation levels but also by high near surface concentrations of the strong oxidant ozone, exacerbating the heat stress.

According to the World Health Organization (WHO) "Health" is defined as "the state of complete physical, mental and social well being and not merely the absence of disease or infirmity". This comprehensive and also ambitious understanding of the term health can also be applied to animals and plants as well as ecosystems (the terms ecosystem health and environmental health have been used very often in recent literature). Therefore, consequences of climate change on health are not confined to diseases but also include reduced well-being or reduced strength of an organism as well as weakened functions in an ecosystem or a socio-economic system.

This contribution to the Encyclopedia of Complexity does not focus exclusively on climate change and human health (see Sect. "Climate Change Impact on Human Health") because climate change impacts on animals and plants (see Sect. "Climate Change Impacts on Plants with Consequences for Human Health") also have consequences for human well-being and human health. An example of the complexity is the changed composition of nutrients in grains as a consequence of higher carbon dioxide levels in the atmosphere, in addition modulated by changed climate parameters that may in turn enhance infestations of plant diseases with consequences for the composition of our food. The section on Climate Change Impacts on Plants with Consequences for Human Health tries to collect known knowledge, reports on potential threats and chances for political reactions as well as pointing out major open research questions.

Climate Change Impact on Human Health

Whenever climate changes all ecosystems have to react, i. e. they try to adapt to changed climate that must also include new weather extremes. Typical reactions are shift of biomes, altered ecosystem composition, new geographical patterns of animal and plant diseases, new relations between predator and prey. As major so-called natural climate changes have occurred also in the recent few million years the first major question in the present anthropogenic climate change era is: Can earlier so-called abrupt climate changes be used as an example for projections into the future? The answer is no, because the mean global temperature change rate to occur in the 21st century exceeds by far even the most rapid natural ones. The largest and most rapid *global* mean temperature change rates in the recent few million years have been the collapses of the large northern hemisphere ice sheets that have led to a mean global warming of up to 5°C and a sea level rise of about 120 m in roughly 10,000 years. Projections for the 21st century without stringent climate policy [4] range between 1.5°C and about 4°C, i. e. an acceleration of at least a factor 30. Hence, the past cannot be used as an analogue. In other words: The key goal of the United Nations Framework Convention on Climate Change (UNFCCC) that speaks of avoidance of a dangerous interference with the climate system, cannot be met without globally coordinated climate policy. It requests, in addition, a climate policy helping to stabilize greenhouse gas concentrations within a time frame that firstly natural ecosystems are able to adapt to climate change, secondly food production is not generally threatened and a sustainable economic development remains possible.

In this situation with ill-adapted forests, rapidly changed patterns of infectious diseases for plants, animals and humans looking back through climatic history does also not help directly. Besides intensified research on changed disease patterns in the very recent past a stringent globally coordinated climate policy under UNFCCC is the best insurance against massively altered disease patterns.

This paper will concentrate on climate change impacts on human health but will not exclude totally impacts on food production. Major points will be "thermal stress" before "vector-borne diseases" and also prolonged allergen seasons and new allergens are discussed.

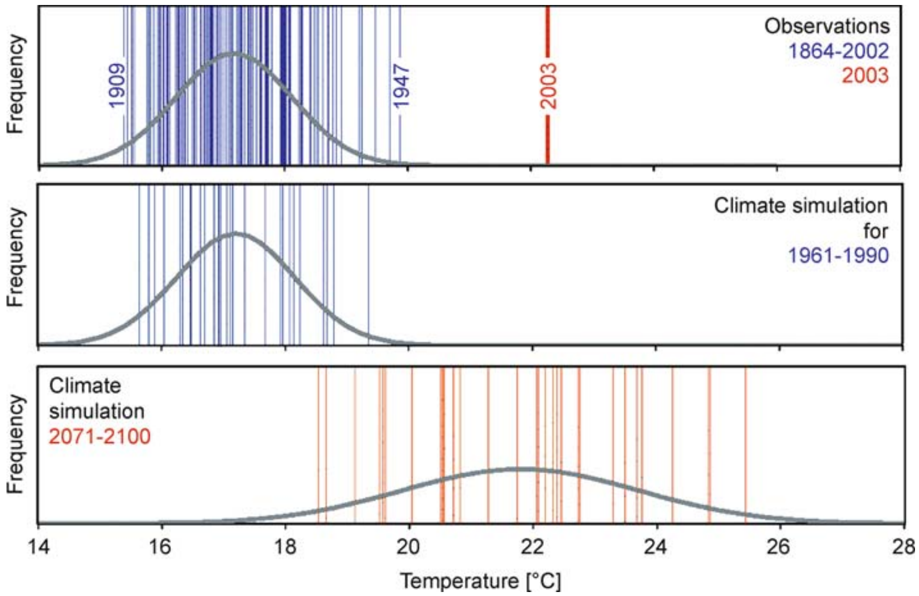
Changed Thermal Stress

Heat or cold stress forces our body to adapt to keep the core body temperature within a narrow temperature interval of about five degrees centigrade (35 to a maximum of slightly above 40°C). While extreme cold stress events have diminished and will further diminish as a consequence of ongoing global warming, extreme heat stress will increase dramatically (see Fig. 1), when the frequency distributions of temperature at a certain location are shifted by only a few degrees centigrade and may be broadened. Up to now only very few places on the Earth's surface exist where survival of a human being is nearly impossible. This would certainly happen if the wet bulb temperature (roughly equivalent to a ventilated sweating naked body) surmounts about 35°C. Under present climate conditions such areas do not exist, but coastal areas of the Red Sea come closest to it during on-shore winds after a sunny day that was heating surface waters to about 35°C. Hence, heat strokes can in principle be avoided by adequate behavior, if buildings are well insulated and properly ventilated and if an individual behaves. Therefore, the huge death toll caused by major heat waves, namely about 55,000 people that died during the heat wave in summer 2003 in Europe [5] is rather an indication for "social freeze" and ill-adapted buildings in our industrialized or developing societies than for really intolerable thermal conditions. Many have died because of lack of care. New weather extremes always demask weaknesses in our security-related infrastructure. A famous example is the large difference in the number of people dying during a summer heat wave in 1995 in two US cities (Chicago and Philadelphia) just because of intensified public care in Philadelphia where the weak and poor citizens where brought by the city administration to the cooled malls during daytime.

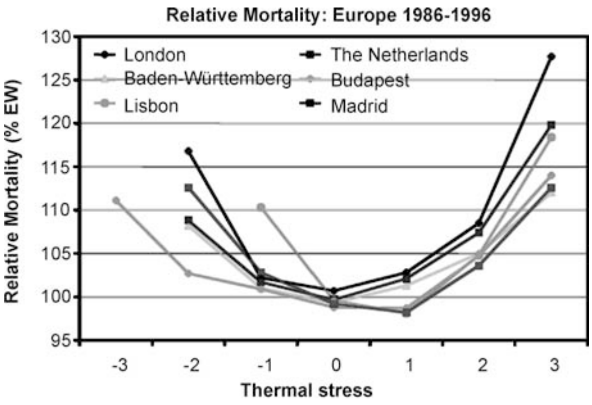
As Fig. 2 clearly demonstrates the mortality anomaly caused by heat waves is a fact observed over several decades (here in a developed country). Lowering the anomaly means not only investment in better warning systems but also enhanced social care in general.

Both heat and cold stress increase mortality as underlined by Fig. 3 for European countries. In the future heat stress category 3 will occur much more often and the lowering of the increased mortality will also be a sign of an improved public health system.

A further point to be made with respect to enhanced thermal stress as a consequence of climate change is the rapidly mounting heat stress in the inner tropics, where



Climate Change and Human Health, Figure 1
Observed (*upper panel*) and modeled summer mean temperatures for the Swiss Plateau both for present climate (1961 to 1990) (*central panel*) and the last three decades of the 21st century for scenario A2. Please note that the exceptional summer 2003 would occur every second year, if globally coordinated climate policy would not exist. From [13]



Climate Change and Human Health, Figure 2
Mean relative mortality in percent for different thermal stress categories observed in Europe during 1986 to 1996. Please note that the mortality increases by more than a factor 2 when heat stress category 2 is replaced by category 3. From [5]

dew points of about 25°C will more often occur, if global warming also continues there. The ability to work with high efficiency is shrinking there rapidly with rising temperatures and dew points. As is well known economic development of developing countries needs a cooled or well-ventilated work place.

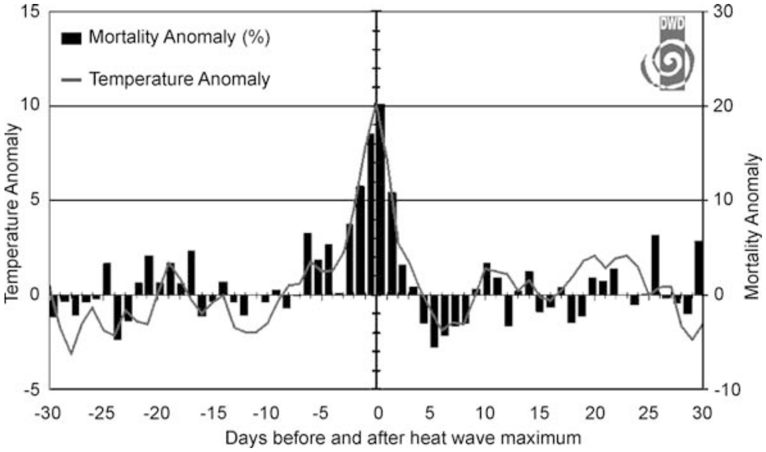
Impact on Photochemical Smog

Photochemical smog is formed if solar radiation stimulates chemical reactions in a polluted atmosphere. Emissions of non-methane hydrocarbons and nitrogen oxides ($NO + NO_2$) lead to the formation of ozone and other oxidizing toxic trace gases as well as aerosol particles. In mid-latitudes photochemical smog is typically strongest in late spring and summer and it has been the reason for some environmental policy making. Heat waves with intense solar radiation lead to major photochemical smog episodes. The higher frequency and longer duration of heat waves during recent and foreseen global warming will intensify health problems already existing during heat waves. As Tables 1 and 2 demonstrate the hot summer in Germany in 2003 has increased the number of days for both ozone and aerosol load (PM_{10}) where limit values have been surmounted in Germany dramatically [14,15], e. g. by a factor of 2 or more both for PM_{10} in rural areas and ozone in all areas.

Fighting against photochemical smog will become even more demanding in a warming world.

Changed Ultraviolet Radiation

Ozone protects us to a large extent from the dangerous part of ultraviolet solar radiation through absorp-



Climate Change and Human Health, Figure 3

Mean mortality anomaly during a heat wave (30 days before and after its maximum) resulting from nine observed heat waves in the state of Baden Württemberg (Germany) for the period 1968 to 1997. Source: [5]

Climate Change and Human Health, Table 1

Days with 8-hour mean ozone concentration above 120 µg m^{-3} since 1990 for all stations in the German network

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
22	22	28	23	32	29	20	22	19	21	19	21	19	51	19

Climate Change and Human Health, Table 2

Days with PM_{10} values above 50 µg m^{-3} for the years 2001 to 2004

2001	2002	2003	2004
Stations with highest values			
117	103	132	73
Stations in cities (average)			
65	75	83	55
Remote stations in cities (average)			
22	30	38	16
Rural stations (average)			
7	12	17	5

tion in the ultraviolet (UV-B) range from 0.28 to 0.32 µm wavelength. The observed latitude-dependent reduction of stratospheric ozone (from 8 to 50 km height in high latitudes), which constitutes about 90% of the total ozone column content and which is caused by chlorine compounds stemming from decay products of chlorofluorocarbons (CFCs) and other halocarbons, has increased UV-B radiation especially in spring of both hemispheres at higher latitudes. This stratospheric ozone decrease is strongest in Antarctic spring in the so-called ozone hole. This ozone depletion has certainly increased and still will increase the skin cancer and cataract incidence that has dramatically grown in recent decades. However, the ozone depletion contribution is largely buried in the variability and change of exposure of our body to UV radiation, which is due to changed behavior, especially in developed countries. The key question for the forthcoming decades is: How fast will ozone column content recovery be after the banning of CFCs and other chlorine and bromine containing compounds? Present knowledge says: Full recovery after several decades with a chance of a super-recovery caused by a further increase of the greenhouse effect of the atmosphere, which lowers stratospheric temperatures.

Health Effects Caused by Other Changed Climate Parameters

Climate change shifts and reshapes frequency distributions of meteorological and hydrological parameters, thereby multiplying the occurrence of known extremes and leading to new ones (see Fig. 1, where this is demonstrated for temperature). Therefore, the health of many more millions is affected by intensified flooding, higher storm surges, and many other weather-related disasters.

While the highest death toll of weather-related disasters was – for thousands of years – due to droughts, with up to 10 million people dying per decade around 1930, the highest death toll is now caused by flooding and wind damage due to tropical cyclones and hurricanes (Red Cross, …). The main reason for this change is – besides more dwellings in flood-prone areas, often already lying slightly below mean sea level – the international aid bringing food and seeds into the drought-affected areas; as long as civil war does not prevent this help.

Hence, climate change also calls – irrespective of globally coordinated climate change mitigation policies – for a coordinated climate change adaptation policy in the coming decades, because we have to adapt to the already unavoidable climate change. Mitigation measures meant to avoid the un-tolerable climate change will only become effective in decades due to the inertia of the climate system caused by the slow reaction of oceans and ice sheets. In other words: If flooding is prevented by strengthened dikes, anticipating shifted frequency distributions of precipitation, cholera epidemics will not occur. Diking has to become an international activity, as the emitters in industrialized countries are causing more flooding and sea level rise on a global scale, co-financed by the already existing but strongly to be increased adaptation fund under the UNFCCC, its Kyoto Protocol and the follow-up protocol envisaged to be signed in 2009.

As an aside I will here report on reactions of our body to high carbon dioxide levels in the atmosphere. Very often if many people are gathered in the same closed room and breathe the same air some will ask for fresh air (i. e. they require more oxygen). The need for fresh air is because of too high carbon dioxide levels. Regulations in some German states concerning ventilation in classrooms provide ventilation rules to avoid carbon dioxide concentrations above 1,500 ppm, a level after which concentration diminishes and some students may even develop signs of a beginning headache. The oxygen content of air has fallen by a bit more than a tenth of a percent only to about 20.84 percent at which point ventilation by opening windows becomes a must. Hence, ventilation of our living rooms means pushing out carbon dioxide.

Transmission of Infectious Diseases from Birds to Humans

In recent years migratory birds have been named as a cause for the transmission of bird flu (avian influenza) to humans because they can in principle transmit the influenza virus to chickens, geese, turkeys and ducks in our farms from where the virus infects humans that come into close contact with the fowl or their products. However, very often the cause for the long-range transmission is global trade and tourism on the one hand plus industrialized animal husbandry in developed and emerging countries on the other. The latter has been found as the principal cause for the comparably rapid mutation of slightly pathogenic bird flu viruses to highly pathogenic ones (Bairlein and Metzger, 2008). These in turn can reach wild bird species which then can rapidly transmit them to new areas, if the viruses are only slightly pathogenic for them. But also pathogens, vectors or reservoir species can inadvertently be introduced by globalization to regions where no longer-term co-evolution could have taken place. Further global warming will intensify the shift of pathogens with the shift of (migratory) birds, but will only add to the further rapid distribution of pathogens caused by globalization (Smith et al., 2007).

Allergies Caused by Pollen

In most countries allergies have recently increased dramatically. In Germany, for example, 20 to 30% of the population suffers from allergies. Most abundant is the allergic rhino conjunctivitis (hay fever), often turning into asthma bronchiale, caused by allergic pollen in air. Climate change has led to longer pollen seasons, in parts to more pollen, changed pollen spectrum and also new pollen [9].

From studies in Europe, North America and Japan an earlier flowering of 1 to 3 days per decade has been reported during the recent decades [11]. For some species, especially the late flowering ones a prolonged pollen season has been found [2,3], in parts caused also by long-range transport of the pollen. Consequently, for many people in mid-latitudes suffering from several pollen the pollen season became longer, sometimes already starting in December and ending only in October after the flowering of the neophyte ragweed (*Ambrosia artemisiifolia*) for Europe.

From differences between cities and rural areas as well as from laboratory studies it became known that pollen abundance increases with carbon dioxide concentration for some species.

In combination with air pollution allergic reactions have been shown to intensify [3].

Invasion of Allergenic Neophytes With ongoing global warming two processes combine for the spread of neophytes: Firstly, global trade and tourism transmitting plants and their seeds within days and weeks around the globe to all inhabited places and into ocean basins and secondly, increased temperatures allowing more and more often establishment of exotic species in new areas. A famous example where both mentioned processes work in combination is the invasion of the strongly allergenic ragweed (*Ambrosia artemisiifolia*) from North America to Europe in the 19th century with a large spread after the Second World War. Ragweed flowers from late August to September and its pollen also undergo long-range transport [1]. In recent decades it spread strongly in Central Europe facilitated by agricultural practice and by inadvertent transport with bird food as well as higher temperatures.

Indirect Health Effects Caused by Climate Change

Many infectious diseases are transmitted by vectors, i. e. by animals (very often insects) transferring a pathogen from another animal and/or human without suffering from the pathogen themselves. The complicated web of hosts, pathogens, reservoirs and vectors is dependent on many factors, among them climate variables, foremost temperature and precipitation. Therefore, distribution patterns and incidence of infectious diseases will also be modified by climate change; however, disentangling the cause and effect relationships is often extremely difficult. Since tropical infectious diseases are the most common of these diseases and especially temperature sensitive they will be a focus in this section. The World Health Organization (WHO) points to a highly probable increase of morbidity and mortality due to higher prevalence and a pole-ward shift of tropical infectious diseases due to further global warming.

Increase of Vector-Borne Tropical Infectious Diseases As always in scientific investigations it is especially difficult to derive long-term trends of certain variables influenced by many parameters like changes in land use, socio-economic conditions and climate or weather patterns. This is also true for tropical infectious diseases. Hence, only few trend analyses exist, e. g. for malaria reaching higher elevations in Africa.

The main vectors for infectious diseases are mosquitoes, biting flies, bugs, lice, fleas, and mites. Of about more than a million insect species roughly 17,000 have adapted to a blood-sucking mode of life. A small minority of these are vectors of pathogens. The pathogens are viruses, bacteria, protozoa or filarioses. Pathogens are either multiplied within the vectors without change of form (Arboviruses, Rickettsiae, Bacteria), multiplied with change of form (Protozoa) or changed without multiplication (Filaria). From uptake of a pathogen during a blood meal until the infective stage temperature is *the* factor determining duration, hence, higher temperatures can lead to enhanced spread of vector-borne diseases as for example observed for dengue fever.

According to number of people infected by blood-sucking insects *malaria* comes first with 300 to 500 million cases per year of which 1 to 2 million die, especially children. About 70 mosquito species of the genera Anopheles transmit four different protozoa (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*). The mosquito *Anopheles gambiae* in Africa is the most important vector. Because malaria has been eradicated in many developed countries in Europe an enhanced potential for malaria must not lead to a re-introduction there. However, in countries with a weak public health system higher prevalence and new malaria infected areas are highly probable.

About 120 million people in Asia, Africa, South and Central America are infected by *lymphatic filariasis* (elephantiasis) transmitted by different genera of mosquitoes. The larvae of the worm develop (from 0.4 to about 1.5 mm) in the vector and can be transmitted during the next blood meal to humans. The development period is temperature-dependent. Hence, a potential for a further spread exists, but the Global Program to Eliminate Lymphatic Filariasis (WHO, 2006) may reduce incidence strongly.

Dengue fever affects about 50 million persons per year and it is a clearly growing threat to human health in about 100 tropical and subtropical countries. The virus belonging to the flaviviruses is transmitted mainly by the *Aedes aegypti* mosquito. Vector and virus show strong temperature dependence in their development. But the disappearance of *Aedes aegypti* from Europe around 1950 points to the probably strong application of the insecticide DDT. As often, measures taken against infectious diseases can easily off-set climate influences.

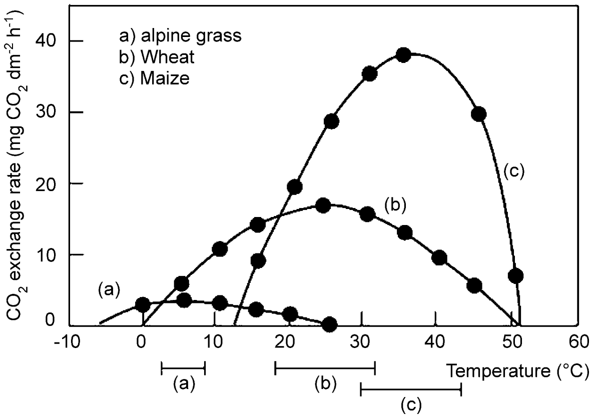
Onchocerciasis (in severe cases leading to river blindness) affects about 37 million persons, mainly in Africa, the microfilaria of the worm "*Onchocerca volvulus*" are taken up from the human skin by the vector, a simuliidae species, during a first blood meal, they develop through two skinnings within the vector and are transmitted back to a human being during a further blood meal after about 7 to 10 days. Whether the vector, developing in running water and sucking blood during daylight, is already reacting to climate change is not known although a temperature and precipitation dependence clearly exists.

Many other tropical vector-borne infectious diseases exist, like loiasis affecting about 13 million people in tropical Africa, "Schlafkrankheit" with about 60,000 new cases per year or West Nile fever that has reached the USA. In all cases it is a complex mix of influencing factors with positive and negative feedbacks, which inhibits a clear separation of a climate contribution to changed patterns and severity. Hence, there will be many surprises often based both on transport of vectors by growing trade and tourism and better survival conditions in higher latitudes due to higher temperatures and/or changed moisture conditions.

The best means to cope with changed disease distribution patterns is a strong public health system and links between the systems in different countries.

Climate Change Impacts on Plants with Consequences for Human Health

Our food is produced by plants on land and in the ocean, even if we eat meat or fish because animals also feed ultimately on plants. Hence, a climate change impact on plants may have strong indirect effects on human health. However, even if the climate change impact were small on certain plants, their way of reacting to enhanced CO_2 concentration could still have consequences for food production. In reality climate change and elevated CO_2 concentration act together and both have largely different impacts on plant types and species. The temperature dependence of photosynthesis rate as presented in Fig. 4 for very different plants, shows the rather steep decline of this rate for maize, a C4-plant, at high temperatures above about 40°C. Adding to this finding that there is nearly no positive feedback to higher CO_2 levels for C4-plants, yield in tropical areas would be reduced if plants have to assimilate at leave temperatures of about 40°C. Therefore, Working Group II of IPCC concluded [4]: "Crop productivity is projected to increase slightly at mid- to high latitudes for local mean temperature increases of up to 1–3°C depending on the crop, and then decrease beyond that in some regions. At lower latitudes, especially seasonally dry and tropical latitudes, crop productivity is projected to decrease for even small local temperature increases (1–2°C), which would increase the risk of hunger."



Climate Change and Human Health, Figure 4
Temperature dependence of photosynthesis rate for an alpine grass, wheat and maize. Please note the strongly differing optimal temperature ranges. From [10]

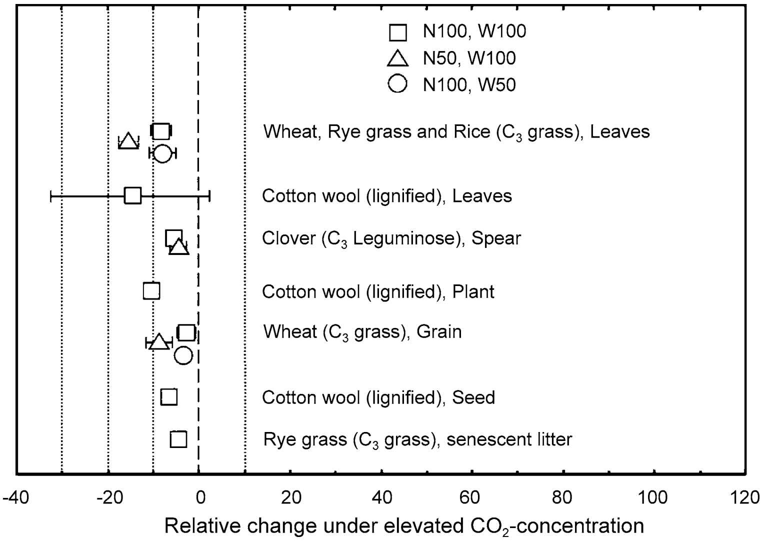
If C3-plants (e. g. wheat, rice, potato, sugar beet) live at higher CO_2-concentrations their photosynthesis rate increases rather linearly with CO_2-concentration, if water stress and nutrient scarcity are not limiting their photosynthesis rate. Consequently, the delay of decades between CO_2-concentration rise and full expression of the warming, to which we are already committed, is a "window of opportunity" for high crop yields of C3-plants. This may in the long-term also have global consequences, because the CO_2-concentration could rise additionally if the (high latitude) forests (C3-plants), acting as a sink for anthropogenic CO_2 presently, would lose this capacity under higher climate change stress. Whether and when this will occur is not yet known.

Changes in Food Composition for Main Crops at Elevated CO_2-Concentration

An important consequence of elevated CO_2-concentration would be changed composition of plant tissue, and especially of seeds, as it could have immediate health consequences for animals and humans eating them. The sparse body of published studies is nearly unanimously pointing to a loss of nitrogen content in leaves and stems as well as in seeds (see Fig. 5 and [6] for an overview). The results of so-called **F**ree **A**ir **C**arbon Dioxide **E**nrichment (FACE) studies at roughly doubled CO_2-concentrations ($\sim$ 550 ppm) are all reporting higher nitrogen content for plant tissue and a bit less for seeds. This negative consequence for our food is not yet fully acknowledged in the public, because of the very different impacts for different parts of society. For a grain-producing farmer the CO_2 fertilization effect leads to higher yields of C3-plants, a dampening of photo-smog yield reductions, higher water use efficiency for both C3- and C4-plants and thus less drought impact while for the baker the wheat quality for baking bread declines, and the cattle as well as the consumer get less healthy food.

Shift of Biomes and Migration

As already observed, precipitation is redistributed due to rather different warming patterns. And projections of precipitation changes in the 21st century, as published in the Fourth Assessment Report of [4], can be summarized in the following sentence: Rather humid areas will on average get more precipitation, especially in high latitudes, while semi-arid areas will get on average less (with strongly lowered soil moisture at higher temperatures). This is bad news for many millions of people in semi-arid zones of the tropics and subtropics. The main impact on plants is water stress, hence lower crop yields. But also less food for cat-



Climate Change and Human Health, Figure 5
Changes in nitrogen content in plant tissue under elevated CO_2-concentration (∼ 550 ppm) depending on nitrogen fertilizer and water availability (50 or 100%). From [16]

tle in pasture lands, continued or enhanced malnutrition for the poor, aggravation of existing deficiencies in public health care systems in these areas will be the negative consequences of climate change impact on the biomes of these areas. In other words: Undermined public health systems, loss of livelihoods and finally migration will be the dire consequences of anthropogenic climate change. The socio-economic and political reactions to this threat cannot be foreseen in detail. However, the international economic cooperation between developing and industrialized countries has to take as a priority adaptation to unavoidable climate change as the best means to lower impacts on public health.

Concluding Remarks

Although threat to human health was often the cause for environmental policy making, e. g. in the case of desulfurization of power plant exhaust, the manifold threats to human health as a consequence of global anthropogenic climate change have rarely been named as a key reason for climate change policy measures. What else is the partly still growing mal-nutrition of subsistence farming communities in the desertification-prone semi-arid tropics and subtropics than a threat to human health? On the other hand most new threats to our health caused by the spread of vector-borne infectious diseases due to higher temperatures or those caused by new weather extremes can be strongly reduced by proper health system up-grading and pre-cautionary measures that strengthen security-related infrastructure. However, this will probably not be the case in developing countries already suffering from (very) weak public health systems; unless the preliminary decisions of the 13th Conference of the Parties to the UNFCCC lead to a new international and binding protocol in 2009 as a follow-on to the Kyoto Protocol that then stipulates that a fixed portion of the revenues of international greenhouse gas emission trading should be used for adaptation measures in developing countries. A large share has to be invested in public health systems in poor developing countries in order to help the poorer parts of societies typically suffering most from epidemics and weather-related catastrophes.

If earlier tropical vector-borne infectious diseases, like the West Nile fever, reach developed countries, research to get proper vaccines will be stimulated within large pharmaceutical companies. If the threat remains confined to the poor South, this research effort will often not exist, because developing countries' normal population cannot afford the expensive new drugs or vaccines that remain property right protected for years. It is high time that political summits deal with this problem and WHO helps to circumvent this barrier as pointed out recently by the Nobel Prize laureate for economics (Stiglitz, 2006).

Bibliography

1. Bohren C, Memillod G, Delabays N (2006) Common ragweed in Switzerland: development of a nation-wide concerted action. J Plant Dis Prot, Special Issue XX, Ulmer, Stuttgart, pp 497–503

2. Confalioneri U, Menne B, Akhtar R, Ebi KL, Hanengue M, Kovats RS, Revich B, Woodward A (2007) Human health. In: Parry MO, Lanziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of IPCC. Cambridge University Press, Cambridge, pp 391–431
3. D'Amato G, Cecci L, Bonini S, Nunes C, Annesi-Maesano I, Behrendt H, Liccardi G, Popov T, van Canwenberge P (2007) Allergenic pollen and pollen allergy in Europe. Allergy 62:976–990
4. Intergovernmental Panel on Climate Change (IPCC) (2007) Science of Climate Change, Contribution of Working Group I to the Fourth Assessment Report. Cambridge University Press, Cambridge
5. Jendritzky G, Koppe C (2008) Die Auswirkungen von thermischen Belastungen auf die Mortalität. In: Lozán JL, Grassl H, Jendritzky G, Karbe L, Reise K (eds) Warnsignale Klima: Gesundheitsrisiken. Wissenschaftliche Auswertungen, Hamburg, ISBN 978-39809668-4-9
6. Kimball BA (2004): Global environmental change: implications for agricultural productivity. Crop Env Bioinform 1:251–263
7. Laschewski G, Jendritzky G (2002) Effects of the thermal environment on human health: an investigation on 30 years of daily mortality data from SW Germany. Clim Res 21:91–103
8. McMichael A, Campbell-Lendrum DH, Corvalán CF, Ebi KL, Githeko AK, Schraga ID, Woodward A (2003) Climate Change and Human Health: Risks and Responses. WHO, Geneva
9. Menzel A et al. (2006) European phonological response to climate change matches the warming pattern. Glob Chang Biol 12:1969–1976
10. Rosenzweig C, Hillel D (1998) Carbon Dioxide, Climate Change and Crop Yields. In: Rosenzweig D, Hillel D (eds) Climate Change and the Global Harvest. Potential Impacts of the Greenhouse Effect on Agriculture. Oxford University Press, Oxford, pp 70–100
11. Rosenzweig C, Cassassa G, Karoly DJ, Imeson A, Liu C, Menzel A, Rawlins S, Toot TL, Seguin B, Tryjanowski P (2007) Assessment of observed changes and responses innatural and managed systems. In: Parry MO, Lanziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of IPCC. Cambridge University Press, Cambridge, pp 79–131
12. Sauerborn R (2007) Climate change: an agenda for research and teaching in public health. Scand J Public Health 1–3
13. Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability for European summer heat waves. Nature 427:332–336
14. UBA (2007) www.env-it.de/luftdaten/download/public/docs/pollutant/03/Jahr/Ozberi06.pdf
15. UBA (2007) www.env-it.de/luftdaten/download/public/docs/pollutant/PM10_gesamt_2001-2006.pdf
16. Weigel H-J, Manderscheid R, Fangmeier A, Högy P (2008) Mehr Kohlendioxid in der Atmosphäre: Fluch oder Segen für die Landwirtschaft. In: Lozán JL, Grassl H, Jendritzky G, Karbe L und Reise K (eds) Warnsignale Klima: Gesundheitsrisiken. Wissenschaftliche Auswertungen, Hamburg, ISBN 978-39809668-4-9

Climate Modeling, Global Warming and Weather Prediction, Introduction to

HARTMUT GRASSL
Max Planck Institute for Meteorology,
Hamburg, Germany

All systems operating away from thermodynamic equilibrium develop structures. The planet Earth will always be far away from thermodynamic equilibrium because of the strongly differing solar radiation input as a function of latitude and season. Hence, the differential heating of the surface and also the atmosphere must lead to temperature gradients, in turn causing pressure gradients that create currents in the ocean and the wind in the atmosphere. On a rotating sphere (in reality the geoid, which is close to a rotational ellipsoid) these flows form low and high pressure systems both in the ocean and the atmosphere which are able to reduce latitudinal gradients but never come close to thermodynamic equilibrium because of continuing differential heating. The average temperatures and flow fields, as well as their strong spatial and temporal variability, are a function of land/sea distribution and atmospheric composition, especially depending on water and ice in clouds. The strongly climatically relevant gases in the atmosphere are, to a large extent, a consequence of life on Earth.

Ranking all radiatively active gases in the atmosphere according to their influence on weather and climate shows the exceptional composition of the atmosphere: Water vapor (H_2O) in all three phases but largely as a gas, carbon dioxide (CO_2), ozone (O_3), nitrous oxide (N_2O) and methane (CH_4) constitute only three thousandths of the atmospheric mass, yet they largely determine how much solar radiation reaches the surface, e. g., through clouds, and how much thermal or terrestrial radiation leaves from there to space, again a strong function of clouds and the above-mentioned gases. The average surface temperature is thus strongly depending on the concentration of the gases mentioned, which are all greenhouse gases. They do not strongly absorb solar radiation but do absorb terrestrial radiation, thereby forcing the surface and the lower atmosphere to warm in order to reach nearly equilibrium between absorbed and emitted energy. Any growth or reduction of greenhouse gas concentrations increases or decreases average surface temperature and thus changes climate.

The climate system, i. e., its interacting components atmosphere, ocean, land, vegetation, soils and crust, shows both a remarkable stability and high sensitivity. Over many million years the greenhouse effect of about 30 K has varied only by about +/ − 5 K with respect to present interglacial temperatures, thus has been stable in terms of temperature varying only from 283 to 293 K. +5 K however meant melting of all inland ice sheets and −5 K a new strong glaciation with major ice sheets reaching 40 to 50° N (see ► Cryosphere Models).

The observed strong increases of all long-lived naturally occurring greenhouse gases (+35% for CO_2, +120% for methane, and +10% for N_2O since 1750) have stimulated a mean global warming which now has emerged from strong climate variability. In 2007, the Fourth Assessment Report of the Intergovernmental Panel on Climate Change concluded: "The understanding of anthropogenic warming and cooling influences has improved ... leading to very high confidence that the global average net effect of human activities since 1750 has been one of warming". On the other hand, the high sensitivity of the climate system is demonstrated in so called abrupt climate change events (see ► Abrupt Climate Change Modeling).

The strongest global one of these – besides the impact of celestial bodies – is deglaciation after an intense glaciation in about 5000 to 10 000 years with a concomitant temperature increase of 4 to 5 K, caused by the slow latitudinal redistribution of solar radiation due to Earth orbit parameter changes as a consequence of slowly changing gravitational forcing by the neighboring planets, mainly Venus, Jupiter and Saturn. Because climate is a key natural resource for plants, animals and humans, any rapid climate change threatens life on Earth. Therefore agriculture (see ► Climate Change and Agriculture), forestry and all economic activities (see ► Climate Change, Economic Costs of) will be impacted by anthropogenic climate change in the 21st century, leading to strong consequences in societal behavior vis-a-vis this challenge, e. g., the one caused by growing inequity between those societies causing climate change, the industrialized countries, and those suffering first or more strongly such as subsistence farmers in semi-arid tropical areas. Climate Models developed so far include many physical processes (see ► Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface as an example), parts of atmospheric chemistry, and vegetated land surface atmosphere interactions, but still lack reaction of ocean biomass production to enhanced CO_2 levels. Finally, it is apparent that global warming will have an effect on human health and that this will include effects on food crops and animals (see ► Climate Change and Human Health).

The low horizontal resolution of global climate models has stimulated nested regional climate models (see ► Regional Climate Models: Linking Global Climate Change to Local Impacts) delivering enhanced output in areas with strong topography or sea/land contrasts. The slowly emerging Earth System Models are no longer driven by changed atmospheric composition alone but by emissions, i. e., they can calculate resulting greenhouse gas concentrations. However they are still not advanced enough to answer the question: When will growing climate change stress turn the present uptake of anthropogenic CO_2 into forests through the CO_2-fertilization into an additional CO_2 source for the atmosphere due to a generally weakened vegetation? It is common knowledge that weather, for example the passage of a coldfront at a certain location, can be forecast only for up to two weeks because of the intrinsically chaotic behavior of atmospheric flow. The accuracy of present day weather forecast models, which are very similar to the atmospheric component of climate models, has advanced strongly, driven by higher spatial resolutions and better parametrizations of sub-grid scale processes in the models, assimilation of more (and especially satellite data) into the models and ensemble forecasting. This has recently led to the same forecast accuracy up to about 10 days in the southern hemisphere, where an in situ observing system for the starting fields of the model is largely lacking.

The increased attention the topic climate change has finally attracted in the political arena will certainly accelerate progress in this field, despite the complex nature of the functioning of the Earth system.

There are two additional articles on climate change which were recruited for other sections. These articles are: ► Dynamic Games with an Application to Climate Change Models and ► System Dynamics Models of Environment, Energy and Climate Change.

Cognitive Robotics

Kazuhiko Kawamura[1], Will Browne[2]
[1] Center for Intelligent Systems, Vanderbilt University, Nashville, USA
[2] Cybernetics, The University of Reading, Reading, UK

Article Outline

Glossary

Amygdala Amygdala consists of almond-shaped groups of neurons located within the limbic lobe in the brain. The amygdale performs primary roles in the formation and storage of memories associated with emotions and is said to have a substantial role in mental states.

Basal ganglia Basal Ganglia are a collection of subcortical neuronal group and have a significant role in the control of movement.

Central executive agent (CEA) CEA is a cognitive or compound agent responsible for high-level executive control, such as reasoning, task switching and realization of internal rehearsal, e. g. within the ISAC cognitive architecture.

Chinese room argument
John Searle developed a thought experiment called the "Chinese Room" argument against what he calls "strong AI". Searle describes a scenario in which a person who knows no Chinese is locked in a room full of boxes of Chinese symbols together with a book of instructions for manipulating symbols. This person receives questions in Chinese from under the door. If the person in the room is able to pass out Chinese symbols using the instruction book to produce correct answers to the questions, he passed the Turing Test for intelligence in Chinese, but he does not understand a word of Chinese.

Connectionist models Connectionist models of cognition are structured on the concept of neural networks. Connectionist networks provide an account for the complex behavior in a way parallel distributed processing (PDP) does. There is no way to distinguish between simple and complex representations in connectionist models. In this sense, they are considered to be subsymbolic.

Cortex Cortex (or cerebral cortex) is a surface structure in the brain responsible for many brain functions including attention, sensory processing, motor functions, awareness, language processing and arguably consciousness. The human cortex is 2–4 mm thick and consists of large sheets of mostly layered neurons.

First-order cybernetics First-order cybernetics considers control and communication in the animal and machine, where the agent receives feedback, including utility of its actions, from the environment.

Cartesian theater A centered locus in the brain called *Cartesian materialism*, because it is the view one arrives at when one discards Descartes' dualism, but fails to discard the associated imagery of a central (but material) theater where it all comes together.

Global workspace Multiple parallel *specialist* processes compete and co-operate for access to a *global workspace*. If granted access to the global workspace, the information a process has to offer is *broadcast* back to the entire set of specialists.

Humanoid robots There is no universally accepted definition for a humanoid robot today. However, it is widely accepted that a humanoid robot must have a body somewhat resembled to a human body, exhibit human-like behavior, and be able to interact with humans using human-level intelligence. As of today, no existing humanoid robots satisfy all these requirements.

ISAC ISAC stands for Intelligent Soft Arm Control. The name arises from the fact that the arm is highly compliant and safe for working with and around people. In its multiagent architecture called the Intelligent Machine Architecture (IMA), human and many modules within the humanoid are represented as distinct agents within a common computational framework.

Minimum robust representationalism (MRR) MRR is a notion, rather than a formal definition, put forward by Clark and Grush that addresses the problem of internal representation when addressing cognitive phenomena. The emphasis on emulators differs from the classical ideas of cognitivism and representationalism. Transparent (i. e. analytically traceable) emulator circuitry is the minimal needed to usefully consider representations of external states.

Multiagent systems (MAS) A multiagent system (MAS) is a software system composed of multiple agents and collectively capable of reaching goals that are difficult to achieve by an individual agent. An agent within MAS can be autonomous in the sense that it has own decision-making capability or non-autonomous like a simple input-output device. MAS agents can include human agents like the case study in this section.

Neural networks Neural networks, or artificial neural networks, are a class of networks of simple processing units which can exhibit complex behavior. They were inspired by the way biological nerve systems, such as the brain, process information. Simple neural networks consist of three layers, input, hidden and output.

Production systems Production systems are symbolic artificial intelligence systems, i. e. they manipulate symbols, instead of numbers. Production systems are composed three parts: a global database, production rules and a control structure. Production rules (or productions) are called if-then rules.

Second-order cybernetics Second-order cybernetics recognizes that the agent has an important effect back on the environment and the two systems affect each other.

Definition of the Subject

Cognitive Robotics is an emerging field of robotics, which will continue to evolve for the years to come. The field of cognitive robotics generally comprises the design and use of robots with human-like intelligence in perception, motor control and high-level cognition. To realize cognitive robots many overlapping disciplines are needed, e. g. robotics, artificial intelligence, cognitive science, neuroscience, biology, philosophy, psychology, and cybernetics. Thus attempting to tightly define the subject is not constructive as often its nature is amorphous, growing and a strict definition could exclude future relevant work.

Work by Clark and Grush [1] towards a cognitive robot definition is well respected. We quote some important considerations below:

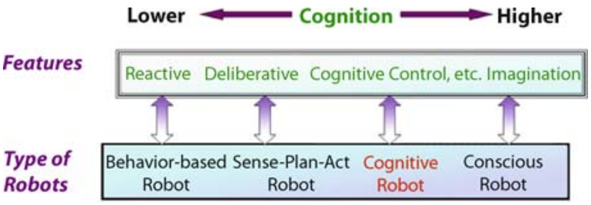
- We hold that fluent, coupled real-world action-taking is a necessary component of cognition.
- Cognition, we want to say, requires both fluent real-world coupling and the capacity to improve such engagements by the use of de-coupled, off-line reasoning.
- Cognizers, on our account, must display the capacity for environmentally decoupled thought and the contemplation of options. The cognizer is thus a being who can think or reason about its world without directly engaging those aspects of the world that its thoughts concern.

In 2006 international researchers representing many of the above disciplines attended the Cognitive Robotics, Intelligence and Control Workshop (COGRIC) [2], which discussed the future of cognitive robotics. Important concepts in cognitive robotics were held to be the ability to form internal states or models for reasoning and decision making (consequently planning), learning from experience, self reflection, embodiment and situatedness, perception to measure world, external behavior/internal representation and importing ideas from human cognition including perception and learning (the entire system) whilst understanding brains and minds. From discussions, it was concluded that it was not worth debating boundaries between consciousness, cognition and intelligence. Future opportunities included lifelong adaptability and developmental learning.

Other researchers have considered this topic, with different list of key features that a cognitive robot should/would possess. Key features that participants identified in the workshop are:

- Ability to perceive the world in a similar way to humans (or better) (e. g., "active perception", Dana Ballard [3], "ecological approach to perception", JJ Gibson [4])
- Ability to communicate with humans using natural language and mental models (robust human-robot interaction, such as overcoming the frame of reference problem, Alan Schultz [5])
- Ability to develop cognition through sensory-motor coordination (e. g., "morphological computation", Rolf Pfeifer [6])
- Ability to have a sense of self awareness (internal states, machine consciousness: Igor Alexander [7], Owen Holland [8] vs. Kevin O'Reagan [9])
- Ability to use attention and emotion (not in a sense of social robots) to control behaviors (K. Kawamura [10])

Robots may be justified in terms of ameliorating a labor shortage, working in hazardous environments, and undertaking repetitive tasks. However, the importance of cognitive robotics is in enabling robots to work autonomously in real-world environments. Currently, robots are not intelligent enough to operate without supervision in real-world situations, e. g. due to difficulties in applying existing knowledge to new situations, selecting between competing goals, coping with multiple sensory input and completing multiple tasks each with subtle variations. Deliberative robots may be able to weld a car, but higher order cognition (Fig. 1) is needed to care for the elderly as the environment is much more unstructured, dynamic, sensory rich, noisy, multitask, varied input/communication, emotive and interactive.



Cognitive Robotics, Figure 1
A spectrum of cognition in robotics. Adapted from [11]

Finally we quote Pentti Haikonen, Nokia Research Center:

"The next big push in robotics may well be machine cognition. Cognitive robotics, that are able to perceive and move as well as we humans do, are the focus of several national and international research programs in EU, Japan and USA. The emerging cognitive technology (as opposed to cognitive science) tries to emulate human information processing and in doing so utilizes experience, affordance (in terms of meaning), attention and even emotion" [2].

Introduction

Humans have sought to create human-like artificial creatures for menial tasks, for coworkers and for the intellectual challenge. Historical stories of Golems raised by religious rituals from mud have lead to the science-fiction of androids assisting or terrorizing the human race. Fiction has now been replaced by increasingly serious attempts by industry and academics, which is treated here as it relates to complexity.

Complexity covers systems where the interaction of multiple components leads to behaviors unachievable by any individual component alone. Cognitive robots require multiple components to function, often in parallel and in harmony, with the goal to produce intelligent human-like behavior and skills.

The origin of modern cognitive robotics comes from the field of cybernetics, the study of control and communication in living organisms, machines and organizations. The term cybernetics was popularized by Norbert Wiener in his 1948 book [12]. Cybernetics had a crucial influence on many important concepts, such as goal-directed behavior generation, self organization and situated nature of intelligence, which are now commonly used in the intelligent robotics community.

In the 1960s, researchers in artificial intelligence (AI) pursued the concept of intelligence from a more deterministic point of view. AI, or commonly known as strong AI, was predicated on the presumption that intelligence or knowledge can be represented as production systems and stored inside of a machine, i. e. computer. However, researchers in cybernetics claim that intelligence is an attribute of an interaction with the environment rather than a commodity stored in a computer and must be actively constructed by a machine itself (e. g., [13]).

The concept of strong AI is that a machine's processors can become a mind, exhibiting many aspects of intelligent behavior, especially sapience and insight into its inner workings. Weak AI, on the other hand, maintains that machines can only reproduce human attributes through the interaction with the environment, which sets constraints on cognitive robots. The best known example of a counter argument to claims made by the strong AI approach is the Chinese Room argument made by John Searle [14]. The debate Searle caused continues today.

In 1999, Clark and Grush [1] pointed out the problem with the notion of internal knowledge or world representation (a part of strong AI argument) and offered a solution called the Minimal Robust Representationalism (MRR). According to MRR, "Cognition, we want to say, requires both fluent real-world coupling and the capacity to improve such engagements by the use of de-coupled, off-line reasoning" ([1], p. 13). We will revisit this notion later in Sect. "Case Study".

As the opening paragraph stated, building a humanoid robot with which to embody human-level intelligence has been a dream of many AI and robotic researchers. In 1993, Rodney Brooks began to work on a first generation cognitive robot [15], which was an upper-torso humanoid and used to generate human-like behaviors. His effort led to a number of government and commercial humanoid development projects in Japan, such as Honda's ASIMO [16] and Toyota's personal assistant humanoid robot [17] during the 1990s and early 2000s. One of the strong arguments for humanoid robot development was that there is a strong desire to replicate human behavior within embedded artificial agents and humanoid robots are the best embedded system to do so. Other researchers consider the field too new to draw this conclusion, thus the jury is still out, e. g. no one knows whether an elderly person prefers a human-like robotic companion to a robotic wheelchair that provides specific needs for the person.

Recently, a number of international projects aimed to develop humanoid robot capabilities that can be used to test developmental theories of cognitive development and language acquisition were initiated (e. g. RoboGroup [18], RobotCub [19], Synergistic Intelligence [20]). The overall objectives of these projects are to study the perceptual, representational, reasoning and learning capabilities of embodied systems in human-centered environments. In particular, these projects propose to use humanoid robots to study mental development processes through robots' interaction with the environment similar to children's mental development processes. For example, RobotCub project, one of the most ambitious among them, aims to develop a set of fully functional humanoid cub robots the size of a 2-year old child.

The aim of this chapter is to provide an overview of the emerging field of cognitive robotics that will evolve over the years into a truly interdisciplinary field. The organization of the rest of the chapter is: Sect. "What Is Cognition?" covers the concept and definition of cognition. Section "Cognitive Architectures" covers major approaches to cognition: symbolic, connectionist and hybrid. Section "Cognitive Robotics" gives an overview of various cognitive robots in terms of important aspects and formative examples. Section "Case Study" summarizes Kawamura's attempt to realize a humanoid robot emphasizing the cognitive task of internal rehearsal. Finally Sect. "Future Directions" provides future directions for the field of cognitive robotics.

What Is Cognition?

This section explores cognition in its influence on constructing advanced robots, rather than debating philosophical insights into its nature. By investigating the brain's functionality, we can gain insight into how to form an artificial cognitive mind.

A goal of advanced/cognitive robots is to endow a level of intelligence and cognitive skills typically associated with people or animals. Although the goal may be based on nature, the methods do not have to be naturally inspired. Much of the early classical symbolic AI-type cognition work (details in Sect. "Cognitive Architectures") were based on computer science techniques, such as production rules, finite state machines and functional programming. Due to issues with scalability, generality, robustness, disturbance rejection and coping with environmental properties (such as uncertainty, noise, nonlinearities, non-determinism and epistasis), biologically inspired methods, such as connectionism became popular in the 1980s.

Although robotic cognition often takes inspiration from biological evidence, it does not guarantee that the outcomes will be equivalent due to discrepancies in evidence and implementation. Investigative techniques, from cell straining to functional magnetic resonance imaging or fMRIs, have been used to 'map' the brain. However, there is much debate on the accuracy and interpretation of the results, with the caveat that the brain's structure/functionality is "not as simple as it appears" often stated. It is argued that a boxology (where functional boxes are connected by lines), which may be easily represented in a computer program, could not replicate the biological evidence.

Mapping brain activity in awake, functioning animals is considered more valuable for cognitive research than when the animal is sleeping, e. g. sensory–motor coordination requires interaction with an environment. Insight can also be gained into self organizing and self reflection during sub-conscious processes using modern neuroscience methods.

Tests for cognition have a tradition in animal as well as human studies, e. g. maze exploration (path planning), T-maze latent learning, mirror test, and relational studies. These are being adapted for robots, e. g. recognition of own mirror image by a small robot [21].

The above tests identify key features and attributes for cognition, but individually do not represent a definition for recognizing cognition. Cognition may be considered as ill-defined due to many alternative definitions. A simple definition came from the foresight initiative by the United Kingdom Department of Trade and Industry (UKDTI) that states:

"Cognitive systems are natural or artificial information processing systems, including those responsible for perception, learning, reasoning and decision-making and for communication and action" [22].

Applying definitions is also controversial, e. g. at what point does something stop being cognitive? (e. g. are ants cognitive as an individual and/or do they display collective cognition?).

Cognitive Architectures

The task of developing a general cognitive architecture and subsequent computational models is greatly complicated by the lack of a clear definition of cognition as discussed in Sect. "What is Cognition?". Nevertheless, cognitive architectures are designed to propose structural properties of the modeled cognitive systems such as humans. Cognitive architectures are classified as symbolic, connectionist, or hybrid (i. e. symbolic-connectionist). As John Anderson, one of the pioneers of the unified theory of cognition, observed, "it is difficult to answer a question regarding which (architecture) might be the correct one or the most correct one since these architectures are often quite removed from empirical phenomena that they are supposed to account for" [23]. Nonetheless, cognitive architectures generally implement cognition as a whole as opposed to cognitive models (the topic of Sect. "Cognitive Robotics"), which focus on particular problems or applications. The aim of this section is to provide the reader a short summary and the important aspects of symbolic, connectionist and hybrid approaches using representative cognitive architectures in the fields of cognitive science, artificial intelligence, cognitive neuroscience, and machine cognition.

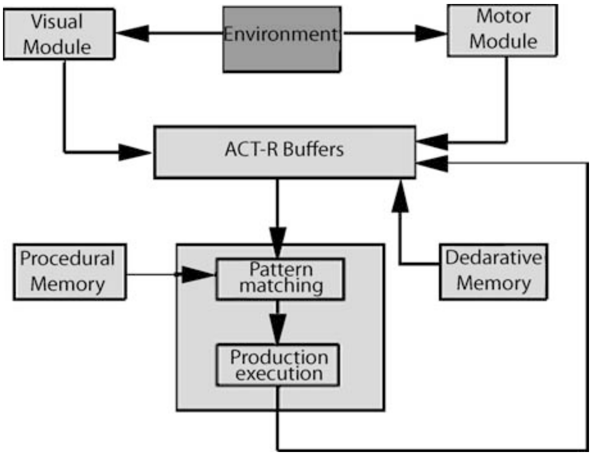
Symbolic Approach to Cognition

The symbolic approach to cognition can be best seen in Newell and Simon's physical-symbol system hypothesis [24] which states that "a physical symbol system has the necessary and sufficient means of general intelligent action". It means that any system that manipulates symbols is sufficient for producing intelligent behavior, and further that all intelligent systems are necessarily implementations of physical-symbol systems. The symbolic approach to cognition dominated the filed of cognitive science and artificial intelligence during the 1970s and 80s. Well-known examples of architectures that fall within this paradigm include ACT-R, Soar and EPIC.

ACT-R (Adaptive Character of Thought-Rational) ACT-R is a cognitive architecture for analyzing human cognitive performance which was developed by John R. Anderson and his group at Carnegie Mellon University [25]. ACT-R has been inspired by the work of Allen Newell and also inspired by the progresses of cognitive neuroscience. It is not a computational model, rather it is a framework for modeling specific human cognitive abilities whose predictions can be compared with human performance. That is, models of assumptions in cognitive processes can be created using ACT-R, allowing results to be easily observed, visualized, and compared with human performance data.

ACT-R makes use of symbolic processing and a sub-symbolic network. Symbolic processing is done through a set of production rules of the classic IF … THEN form. The sub-symbolic level runs concurrently with the symbolic level and consists of parallel processes that influence the performance of the symbolic system, such as conflict resolution, data retrieval, and rule execution. This allows the designer to have extra control over various conditions that affect the system performance without having to create explicit production rules. The central feature of the architecture is that all processing depends upon the current goal of the system. Functionally, ACT-R consists of three basic components: modules, buffers, and a pattern matcher as shown in Fig. 2.

There are two types of modules: perceptual-motor, and memory modules. Perceptual-motor modules (i. e., Visual and Motor Modules in Fig. 2 Declarative Memory) interface with the real world by perceiving environment and generating actions. Memory modules represent ACT-R's most important assumption, i. e. human knowledge can be divided into two kinds of representations: declarative and procedural. Declarative memory consists of facts which can be used to represent goals. Procedural memory describes tasks in the form of production rules. Each production rule consists of a condition in which it should be executed, and an instructional component describing what to do.

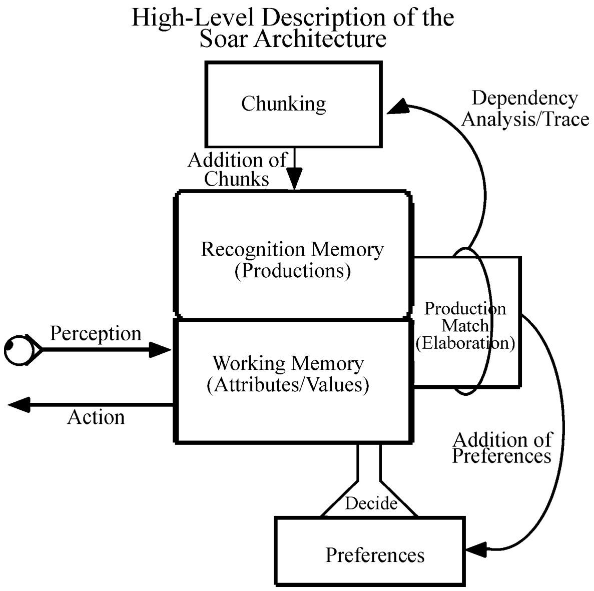


Cognitive Robotics, Figure 2
Functional structure of ACT-R. Adapted from [25]

Buffers hold declarative memory modules to be used by ACT-R and are somewhat analogous to the human working memory system [26]. Declarative information that is relevant to that goal is activated and considered to be in the focus of attention. The contents of the buffers are analyzed and a production rule is matched by the Pattern Matching module. A production rule is selected based on how well the conditions of execution associated with the production rule are met. The reader is referred to the ACT-R web site [25] for further information on the architecture and useful links.

SOAR (State, Operator And Result) Soar is another example of a production system-based cognitive architecture that implements goal-directed human behavior. It was created by John Laird, Allen Newell, and Paul Rosenbloom at Carnegie Mellon University [27]. Soar has been widely used by AI researchers to model different aspects of human behavior. Initially, Soar had been presented as a problem-solving architecture for artificial intelligence. Since then, Soar has been applied both within artificial intelligence and psychology for modeling human cognition.

Similar to ACT-R, underlying the Soar architecture is that knowledge representations can be made in forms of


High-Level Description of the Soar Architecture
Chunking
Dependency Analysis/Trace
Addition of Chunks
Recognition Memory (Productions)
Production Match (Elaboration)
Perception
Working Memory (Attributes/Values)
Action
Addition of Preferences
Decide
Preferences

Cognitive Robotics, Figure 3
Soar cognitive architecture. Adapted from [28]

procedural, declarative and episodic. Also Soar tries to address a collection of mechanisms of mind. Figure 3 illustrates a high-level Soar architecture showing major mechanisms.

Soar represents all tasks as collections of problem spaces, following Newell's problem space principle [29]. Problem spaces are made up of a set of states and operators that manipulate the states. Soar begins work on a task by choosing a problem space, then an initial state in the space. Soar represents the goal of the task as some final state in the problem state. There is a single current state that represents the information about the problem or task being processed, including information about the current situation, the current state of the model, the goal toward which the model is working, and information relevant to that goal. This information is brought into the Soar's dynamic memory (SDM). All productions that match the current contents of dynamic memory then fire in parallel. The results usually cause changes in the dynamic memory, so other productions may now match. Those productions are then allowed to fire causing further matching. This process continues until no new productions fire. At this time, there may be a set of possible actions available as results from the productions firing. Action decision is processed by the *preference system*. When an action is performed, this usually results in a new state, and thus the operation can be repeated.

All learning in Soar occurs through the "chunking" mechanism, which is a form of explanation-based learning and a by-product of impasse resolution [30]. When an impasse is resolved and results in a change in the contents of dynamic memory, information about the resolution is stored as a new production rule. Impasses occur frequently during problem solving in Soar therefore, so does learning. The reader is referred to the Soar web site [28] for complete documentations and useful inks.

EPIC (Executive-Process/Interactive-Control) EPIC is a cognitive architecture developed by David E. Kieras and David E. Meyer at the University of Michigan [31]. While ACT-R and Soar focus primarily on the internal information processing of humans, EPIC extends to include the perceptual and motor aspects of human cognition. A schematic overview of the EPIC architecture is shown in Fig. 4. It consists of a cognitive processor in the form of a production system, a set of perceptual and motor processors and a simulated task environment allowing interaction with the surroundings. EPIC has three types of simulated sensory organs: visual, auditory, and tactile. For each sensory organ, there is a perceptual processor that works in parallel to each other. These processors receive information from the sensory organs, convert it into a symbolic form, and send the symbolic representation into working memory.

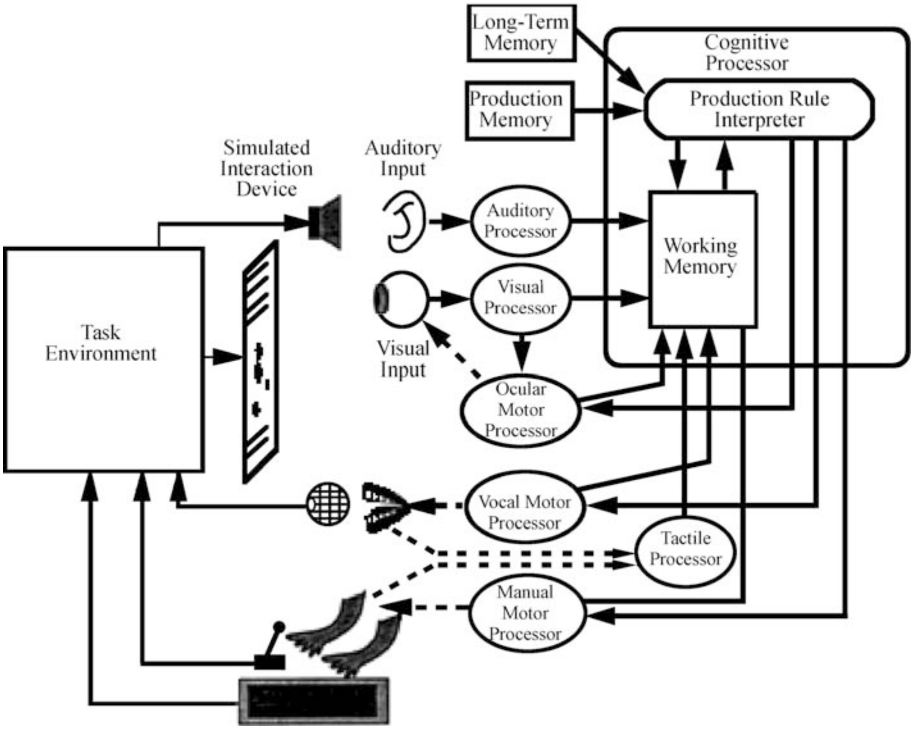
There are motor processors for each of the three sensory processors: ocular, vocal, and manual. The inputs to the motor processors are abstract symbolic representations of responses, which are then translated into explicit responses and passed to the motor processors. Information from the motor processors is passed to the working memory as a feedback.

Overall control of this architecture is attained through the use of task-specific control processes. There is no central control structure. Instead, production rules are used that are of the same form, but functionally separate from those used to encode information. Limitations in the EPIC architecture are postulated to be in the capacity of perceptual-motor processes.

The most important issue being pursued with EPIC is said to be the nature of human multiple-task performance. Major applications are user interface design to telephone operator workstations and cockpit systems, in which operation speed is critical and multiple perceptual and motor modalities are involved. The reader is referred to the EPIC web site [33] for complete documentations and useful inks.

Connectionist Approach

The connectionist approach to cognition provides an alternative theory of mind to the symbolic approach. It be-

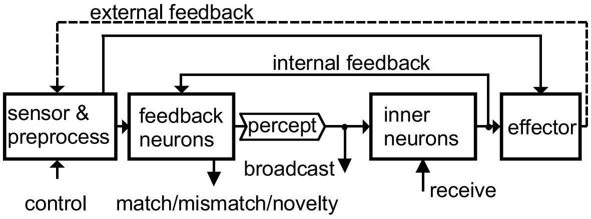


Cognitive Robotics, Figure 4
EPIC architecture. Adapted from [31,32] with kind permission from Professor David Meyer

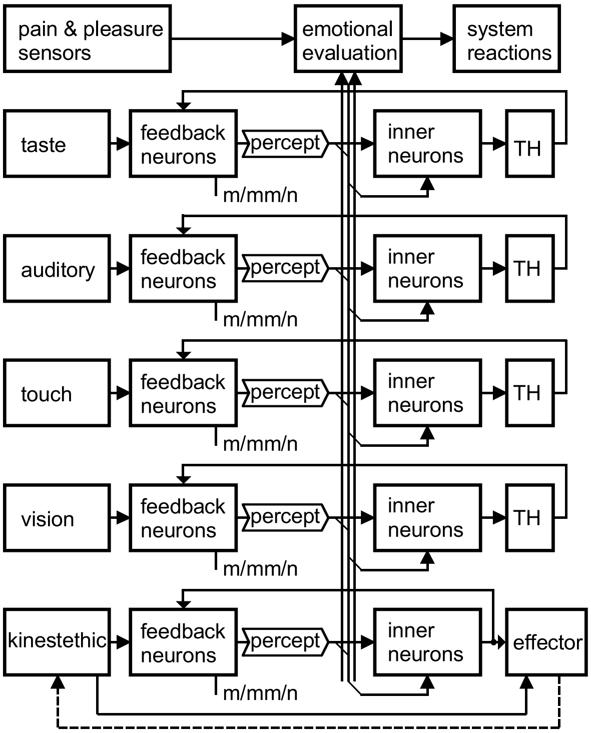
came popular in the 1980s and its popularity has not been diminished. The connectionist approach differs from the symbolic approach in almost all major dimensions. Connectionist (or parallel distributed processing) models offer many attractive features when compared with standard symbolic models. They include a level of biological plausibility, parallel distributed representations, pattern generalization performance, and adaptive learning.

Connectionist models represent information throughout a network of simple, but highly interconnected units or nodes. In some connectionist networks, each unit has a particular meaning. In others, such as neural networks, the nodes are individually meaningless and information is represented as a function of the simultaneous activation of multiple nodes. Neural networks [34] are one of the most popular connectionist models used in many fields, including artificial intelligence and robotics. The number of different connectionist architectures available today is large; to discuss them all is beyond the scope of this section. Instead, this section will focus on two unique architectures, which the authors believe generic enough to give the reader a good sense how to design connectionist-based cognitive robots. For more general discussion of connectionist approach, the reader is referred to an excellent review [35]. For more technical details, the reader is referred to, e. g. the Handbook on Parallel and Distributed Processing [36].

Connectionist Cognitive Machine Architecture Pentti Haikonen of Nokia Research Center takes a connectionist approach to build an embodied cognitive architecture. Instead of creating specialized models that reproduce cognitive processes directly at higher representational level like ACT-R, Haikonen proposes a cognitive architecture based on a distributed signal representation as a building block as shown in Fig. 5. In Fig. 5, “the preprocessed sensory information in the form of distributed signal representation is forwarded to feedback neurons, which also accept feed-



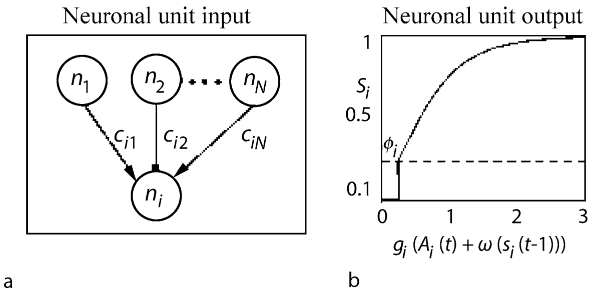
Cognitive Robotics, Figure 5
Perception/response loop module. Reprinted from [37] with kind permission from Dr. Pentti Haikonen



Cognitive Robotics, Figure 6
Haikonen's cognitive machine architecture. Reprinted from [37] with kind permission from Dr. Pentti Haikonen

back signals from the system". The feedback and the sensory information are compared with each other and the eventual match, mismatch and novelty conditions are indicated by respective signals. The internal feedback indicates also the intended act that is to be executed subsequently by the effector and will therefore be expected as the corresponding sensory percept. In that case, "the feedback signal is comparable to the so-called corollary discharge which is forwarded from the motor area to sensory areas in the brain" [37].

Figure 6 illustrates a schematic diagram of Haikonen's cognitive architecture. The cognitive architecture determines the cross-connections between the sensory perception/response loop modules of the Fig. 5. The architecture allows the association of complex additional meanings with percepts and also allows the generation of emotional significance, the "emotional soundtrack" to be associated with memorized entities and episodes. This emotional significance is then used to control attention. Each module works on its own and produces streams of percepts about environment and the module's own internal states. This output is then fed back to the feedback neurons and via this route becomes a percept.



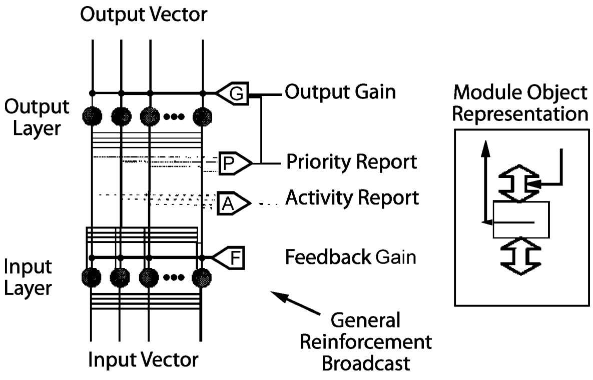
Cognitive Robotics, Figure 7
Typical neuronal unit model used in SNMs. Reprinted from [40] with kind permission from Dr. Jeffrey Krichmar

Synthetic Neural Modeling The synthetic neural modeling approach is a general method of testing theories of brain function at the system level [38]. It was developed by a group of researchers at the Neurosciences Institute over the last two decades. These models are designed to demonstrate the theory of neuronal group selection (TNGS) proposed by Edelman [39]. That is, it is based on the premise that neuronal circuits in the brain, formed with functional properties, the value of which is not knowable a priori, compete with one other to participate in the determination of behavior via strengthening of their connections with other brain circuits. Basic unit of selection in the TNGS are neuronal group, collections of dozens, or hundreds of neurons that are strongly connected to each other via synapses. A synthetic neural model (SNM) comprises a device, an environment and a neural system simulation [40]. Figure 7 shows a typical SNM unit (a) and its output (b). In (a), neuronal unit i receives input from N neuronal units via synaptic connections labeled c_{ij}, $j = 1, 2, \ldots, N$. Each connection has a relative strength that can be either excitatory (e. g. c_{i1} and c_{iN} are positive) or inhibitory (e. g. c_{i2} is negative). In (b), the output of a neuron is subject to a function based on its current activity and the input from other neurons. Below a specific threshold φ_i, the output is 0. SNMs have been embedded in a series of mobile robots called Darwin [38]. Section "Cognitive Robotics" includes a short description of such a Darwin model.

Hybrid Approach

The hybrid architecture is intended to capitalize on the complementary strengths of production-systems and connectionist architectures to implement a human dual processing theory of controlled and automatic processing.

CAP2 (Controlled and Automatic Processing 2) CAP2 is a hybrid cognitive architecture, incorporating both sym-



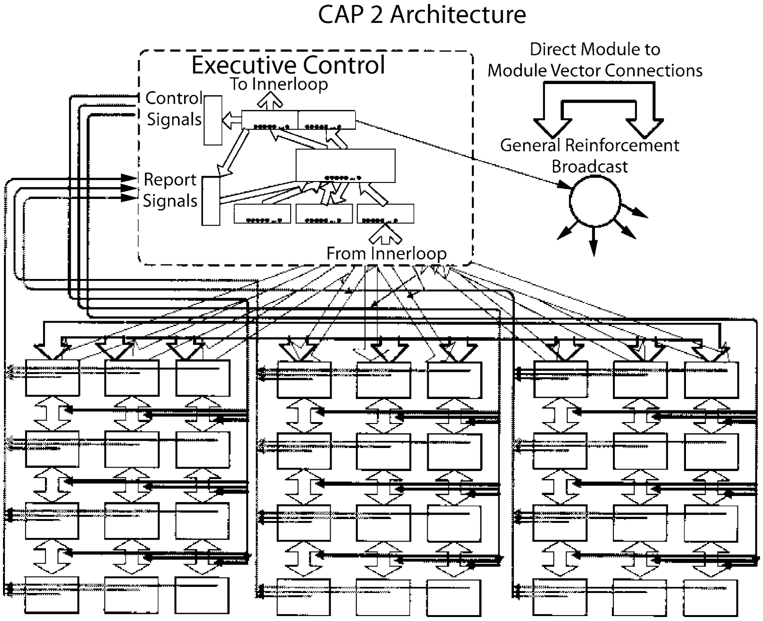
Cognitive Robotics, Figure 8
Microstructure of a CAP2 module. Reprinted from [41] with kind permission from Professor Walter Schneider

bolic (e. g., ACT-R) and connectionist (e. g., PDP) elements. It was developed by Walter Schneider and others at the University of Pittsburgh [41]. It is based on a human dual processing theory of automatic and controlled processing [42]. The CAP2 architecture is implemented with entirely connectionist components, but has networks that operate as sequential control structures to behave like a production system. The basic module of the CAP2 architecture is shown in Fig. 8. It involves two layers of units and five control signals. It is modeled after the cortical columns in the human cortex. An input vector of activation enters the module from below, passes through a connection matrix that reflects prior learning and evokes a new activation pattern in the input layer. The input layer then activates the output layer through a second connection matrix and sends a vector of activation to other modules.

A schematic overview of the CAP2 architecture, which implements a hierarchy of modules or modular processors is shown in the Fig. 9. At the top of this hierarchy is an executive control network, which is responsible for the control of the flow of information between modules. This control of information flow is conducted according to the current goal of the system, thus making it conceptually similar to the ARC-R architecture.

Cognitive Robotics

Now that the relevant aspects of cognition have been identified and the formative architectures in the field have been outlined, this section explores cognitive robotics. It does not attempt to provide necessary and sufficient measures for determining if a device is a cognitive robot as the field is still maturing and such definitions are likely to prove limiting when exploring the still somewhat unknown nature of cognition. Nor will this section present a history of cognitive robots as it is more insightful to detail important aspects of the field illustrating concepts with features of individual robots.



Cognitive Robotics, Figure 9
The CAP2 architecture showing three regions with four levels in each region. Reprinted from [41] with kind permission from Professor Walter Schneider

The first aspect is that a cognitive robot is embodied, which is commensurate with the situated nature of intelligence. It is noted that most robots have self-contained processing relating to a single environment often with a single objective. Increased wireless data transfer rates have resulted in the possibility of the processing being remote from the sensing and action. Furthermore, a single processor could now be embodied in multiple locations simultaneously, whilst communicating with other embodied and remotely embodied processes. Note: a disembodied processor has no online environmental interaction, which is employed occasionally where an offline agent is used to quickly develop promising strategies based on the model of the real environment built or learnt from embodied interactions.

Arguably the first robots exploring cognition were built by the cyberneticist Grey Walter in the late 1940s. The idea of cognitive robotics had not been developed in 1948 when the first tortoise was designed. The legacy of Grey Walter [43] is such that some of his ideas have influenced biologically inspired robotics and are still relevant today. The tortoises were based on analogue electronics as digital electronics barely existed at the time. They relied on relays – threshold devices – to implement behavioral transitions. Phototaxis behaviors were obtained as well as touch-controlled behavior and results were reported on the mirror test [44].

Walter was also an eminent neuroscientist, a field that continues to have much input and importance into cognitive robotics. This highlights the equal support of all the associated fields. This support is a two-way process, e. g. the intellectual challenge of creating cognitive robotics includes testing insights into natural cognition. One unique robot, or a design study, along this line is CRONOS robot (Fig. 10) being developed by Owen Holland [45]. Holland calls his approach *anthropomimetric robotics*.

Cybernetics arose from the realization that many seemingly diverse fields share similar properties; namely the importance of feedback from the environment to a system. This feedback is central to communication and control in order to achieve efficacy of action. The theoretical insight from cybernetics into dynamical systems is very relevant to cognitive robots.

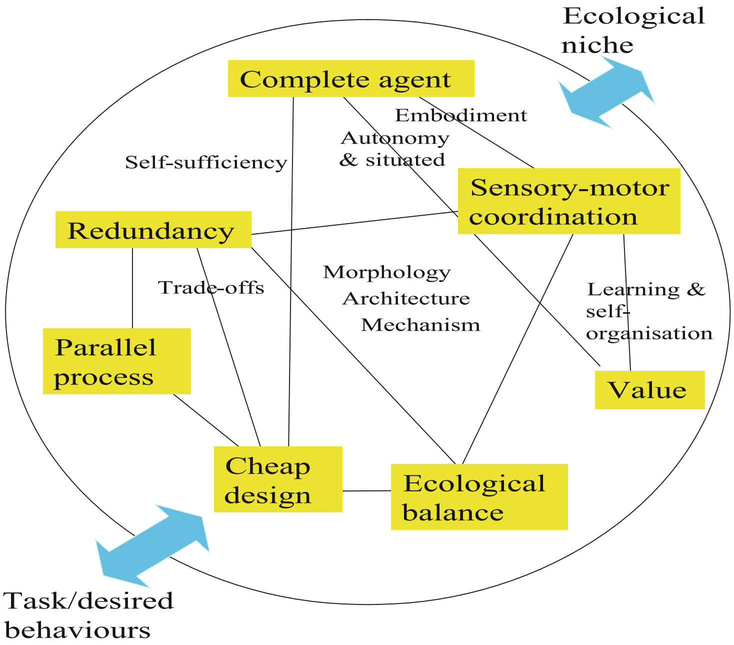
First-order cybernetics considers the effect of the environment through feedback on the robot, whilst second-order cybernetics acknowledges that the robot can also affect the environment and thus the meta-system must be taken into account. Consequently, the environment can determine the level of cognition that any robot can display. An example is where an exploratory robot has to learn only not to collide with obstacles, so the robot learns to stay away from obstacles by staying still. Braitenberg, a cybernetician and a neuroanatomist, originated the Braitenberg Vehicles as a thought experiment to "illuminate the key issues of what we may call cybernetics or artificial intelligence or cognitive science" [46] based on adjusting the morphology of a vehicle in order to respond differently to similar environmental stimuli [47].

Cognitive Robotics, Figure 10
CRONOS robot. Reprinted with kind permission from Professor Owen Holland

Morphological computation is an important modern research field where the robot is designed to complement the environment. Whereas evolution has had millennia to adapt creatures to their environment in order to exhibit cognition, robots must be designed to fit their environment. An example is the robot Puppy designed by Iida and Pfeifer [48], where stable gates are obtained through efficient mechanical design and simple processing rather than convoluted cognitive learning. A morphological design guide, adapted from Pfeifer and Scheier [49], is shown in Fig. 11.

Perception interfaces with the environment, so the sensor set-up of the robot is core to cognition. Human eyes are tuned to the wavelengths of primary colors so will have a stronger reaction to these colors (e. g. red) than a robot with a CCD camera that treats visible frequencies equally. Humans have multiple types of sensors (more than the anecdotal five), whilst robots may have many more (such as laser range finding). Thus robot cognition of an environment will differ, subtly or significantly, from human cognition.

Even if a human and a robot could be instantaneously given the same perceptions there is the *explanatory gap*, where the quality of the experience is important. Smelling



Cognitive Robotics, Figure 11
Morphological design guide. Adapted from [49]

a red rose will bring a deeper quality of experience to a human due to past associations. Humans also utilize *active perception*, where the perception of an object is deliberately manipulated in order to experience it to a greater extent. Thus a cognitive robot will need to perceive by exercising a skill, rather than passively receiving information.

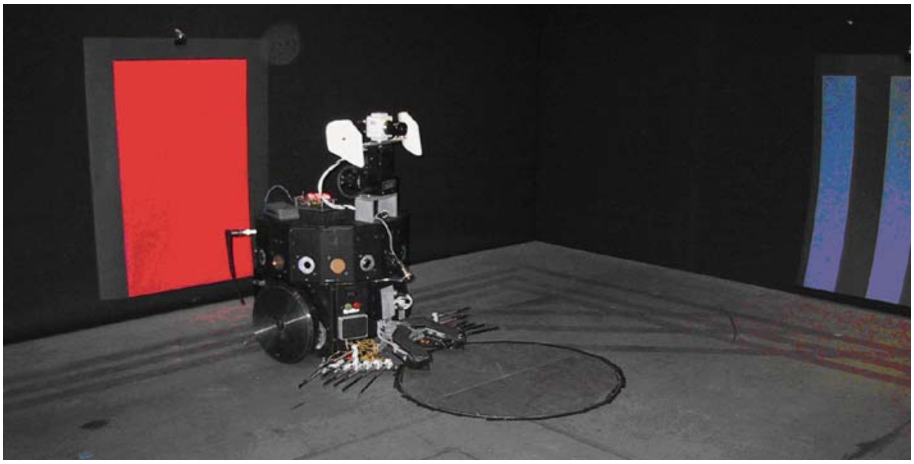
The perceptions need to be represented internally within a robot in order to manipulate them to determine the appropriate action. Either connectionist or symbolic representations may be utilized. Connectionist approaches reputedly mimic the representation in the human brain, which is difficult to understand. Symbols have proved useful when transparency for human readability and understanding of cognitive processing is required. The *symbolic grounding problem* [50] addresses how symbols get their meaning and in a deeper sense, what it is to experience the symbol. The word 'red' is very emotive to humans, but storing the letter sequence 'r e d' in a memory address does not have the same connotations. The more symbols are experienced, the greater the grounding.

A connectionist architecture based on the biological evidence has been successfully used by Krichmar [40] to replicate the behavior of rats in a Morris Water Maze (a simple water-based maze where a rat is required to use memory and path planning in order to locate a safe platform) (Fig. 12). Through much post processing of the learned artificial neuronal links an additional artificial pathway was deemed necessary, which was subsequently found to exist in nature. This is an example where cognitive studies can be enhanced through cognitive robotics.

Although it is possible to perceive and represent a multitude of external information it would be too slow and take too much memory if every detail was processed and stored. Thus a cognitive robot must have *attention* on the salient features of the domain, whilst ignoring others. This reasoning continues as the actions are available to the agent (*affordances*) within a domain must be determined for analysis/selection. Classical search is too slow so key feature selection is required. Once an action has been selected the *frame problem* (Frame Problem in Artificial Intelligence) occurs as the cognitive robot must determine what to update dependent on the results of effecting this action.

It has been argued that the mind has a central executive that acts as a librarian for memories [51]. Although this may provide a practical solution, it is undetermined what tells the central executive how to act. As an alternative approach, Noelle proposes that:

"Since working memory is employed extensively during cognitive processing in humans and animals, it is reasonable to conjecture that robots could benefit from the capabilities provided by a working memory system. Benefits could include the focusing of attention on the most relevant features of the current tasks" [52].



Cognitive Robotics, Figure 12
Darling robot exploring dry version of Morris Water Maze. Reprinted with kind permission from Dr. Jeffrey Krichmar

It is worth noting here that memory is not considered uniform in humans, with types such as episodic, procedural and semantic being identified [53]. Episodic memory encodes individual events, whilst semantic memory expresses the essence of memories. It is argued that a mechanism, such as abstraction, is required to transfer between episodic and semantic memory. Memory is also temporal in nature, which has caused much debate into its categorization. Long-term, short-term and working memory are hypothesized, but it is the ability to recall appropriate past events/lessons in a timely manner and compare these with prescient events that is important to cognitive robotics.

Learning from feedback encompasses one of the most active areas of AI research, so it will not be detailed here. Instead, a few important aspects for cognitive robotics are highlighted. To map the vast world into memory requires compactness as well as the recall of salient features. States, actions and feedback can be *imprinted*, which although reliable is memory intensive. The ability to *generalize*, i. e. remove irrelevant information, improves performance but more steps are necessary. *Abstraction* is where patterns, both spatial and temporal, are identified independent of their associations, e. g. the concept of a tree without recalling specific instances. The ability to scale learning is a main challenge in the field of AI.

Many advanced robotic platforms exist or are emerging, such as ASIMO [16] and iCub [19]. Unfortunately, many modern actuators require much power, space and advanced control. Lack of compliance, which is inherent in most animals and other material properties, can limit the behavior in a cognitive robot. CRONOS [45] is an example of a platform with natural actions to mimic human cognition. Thus a new 'breed' of robots will require new control methods as their nonlinear nature results in many existing control methods being invalid.

Now that we have considered an embodied robot with the rudiments of cognition, there are aspects of cognition, behavior and social interaction that need to be considered.

A cognitive robot will need to plan a series of actions over time. This will require a series of state-action links, an internal world model, or a method that utilizes the world itself as a model. Cartesian theater, global workspace [54] and other architectures have been proposed for such a model.

Despite the ability to memorize and learn from situations, a robot will need to react to unknown/unfamiliar situations. Interpolation and extrapolation are popular techniques, but are often hard to compute and cannot be guaranteed to be effective. Instead, artificial emotions are being considered as action modifiers, goal setters and decision makers. Emotions are also useful to aid human robotic interaction, but their cognitive uses are much wider [55].

Most previous robots were designed to achieve a single or limited number of similar tasks and goals. Increasing the range of competencies of a robot will require it to select between goals, so a value system will be needed on which to base decision-making. Science fiction may suggest hard-coded laws for robotics [56], but this could lead to conflicts and contradictions. Debates have begun on the ethics and legal aspects of giving robots autonomy in goal setting, especially if this enables new goals to be determined.

Robots operating in a domestic or social setting will need to assess the intentions of others to aid collaboration and prevent destructions/accidents. Humans have inherited neural circuitry to recognize faces and use body language/gestures as important communication tools. Much work is required in this field and it is unproven on how humans will react to robots in a social setting [5].

Work in the 1970s by Masahiro Mori on the Uncanny Valley hypothesizes that a robot must appear and behave very humanlike to be acceptable [57]. Alternatively it is better to make the robot mechanical in appearance, so it would be acceptable as a mechanoid.

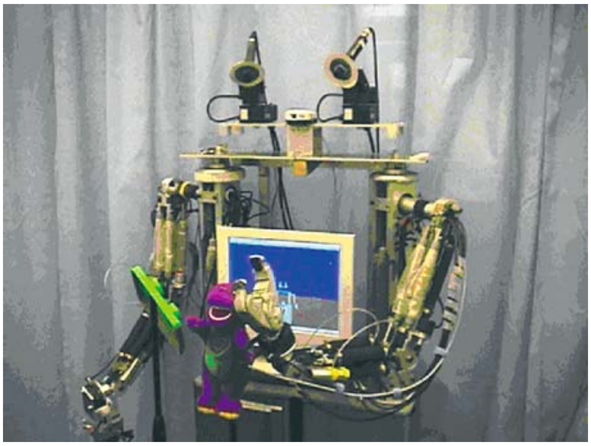
After considering what future cognitive robots may need it is worth noting that their processing is unlikely to be serial, silicon-based digital systems. Parallel systems are more timely when processing large quantities of data. Analog systems could encompass non-linear and possibly nondeterministic behavior. Finally, biological substrates, which exhibit both the above properties, are being developed for use in robot control [58].

Case Study

Introduction

This section presents a case study on the humanoid robot ISAC [59] that one of the authors (Kawamura) is developing. Specifically, the task of reaching for a named object involving internal rehearsal is explored as sense, memory and planning are all required and the implementation of an 'inner sense' where sensory experiences and consequences of different behaviors may be anticipated is developed.

The first lesson is that humanoid robots are indeed complex systems due to the number of components that must interact. Figure 13 shows ISAC that is an upper bodied humanoid robot incorporating stereo vision, audio and touch sensors with pneumatic actuation for multiple arm and hand movement. Sensing inputs are mapped to non-linear and flexible actuators through parallel, distributed processing that represent perception, reasoning and learning.



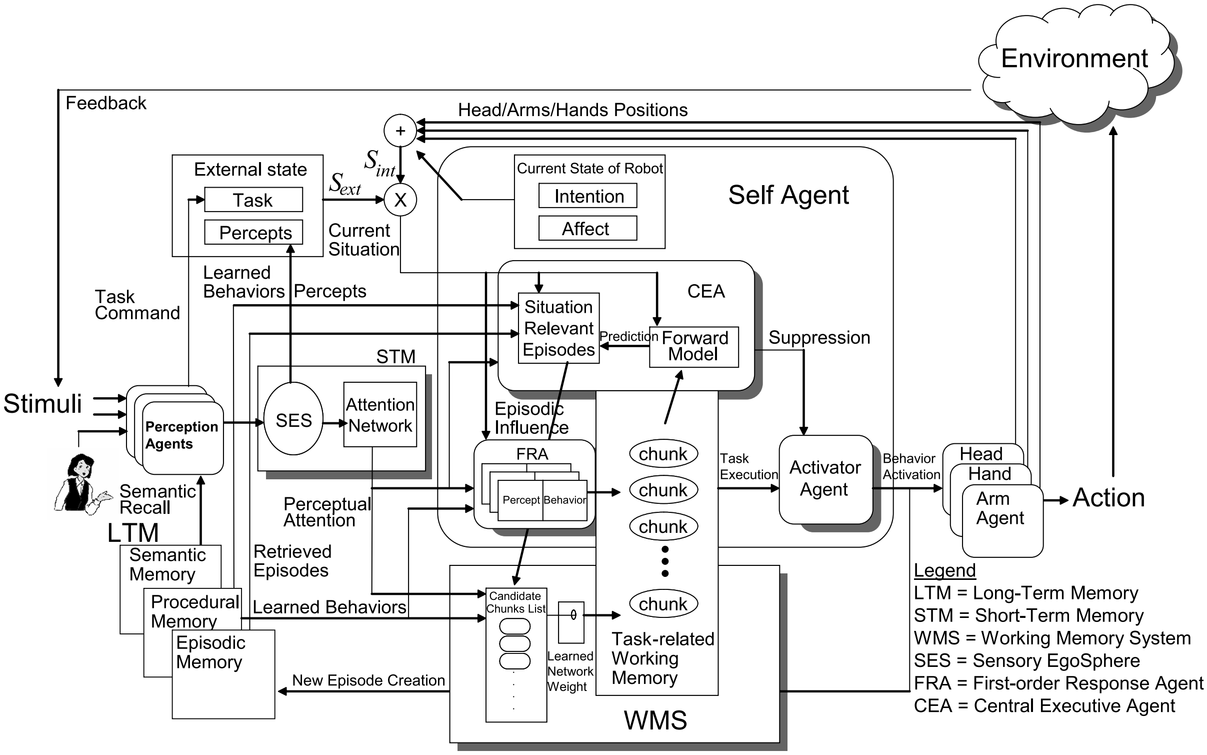
Cognitive Robotics, Figure 13
ISAC reaching to Barney with left arm

Architecture

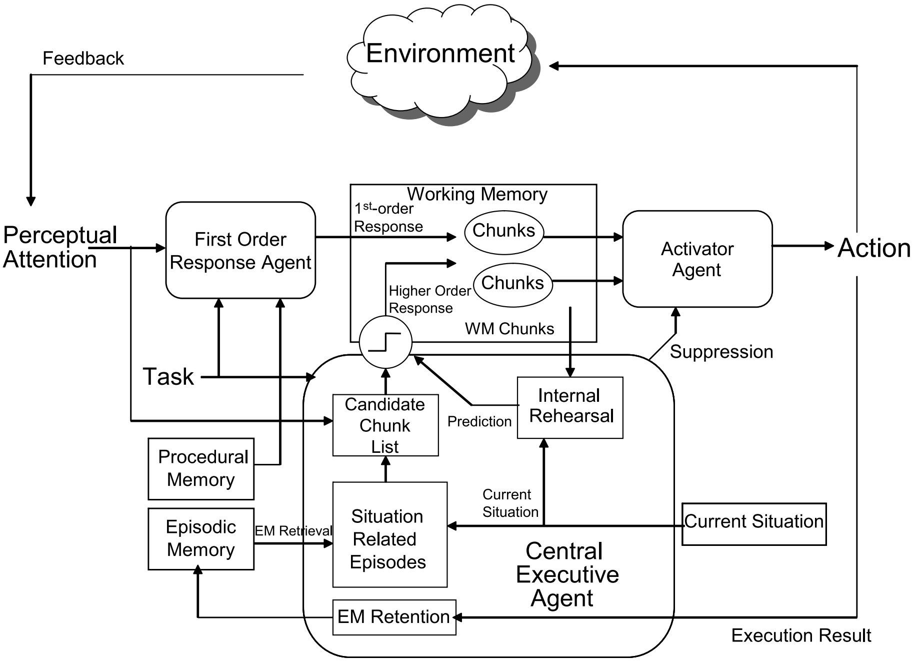
During earlier development it was realized that enhancement and maintenance of such robotic software systems could benefit from domain-specific guidelines that promote code reuse and integration through software agents. This led his group to develop a multiagent-based robot control architecture based on the Intelligent Machine Architecture (IMA) [60,61]. IMA allows for modular design and the development of subsystems from perception modeling to behavior control through the collections of software agents and associated memories. Figure 14 was configured using IMA agents and associated memory structures. It consists of a number of IMA agents and a set of memory structures. Within the cognitive architecture, the Self Agent represents the robot's self [62] and is responsible for multiple aspects of cognition including internal rehearsal.

In the ISAC architecture, the Self Agent handles the dual sensory-motor loops as shown in Fig. 15. The First–Order Response Agent (FRA) is responsible for the reactive and routine responses of the system while the Central Executive Agent (CEA) is responsible for the cognitive response. The Internal Rehearsal System (IRS) takes the working memory chunks as the motor commands, the current situation as the external state and sends a rehearsed result to CEA. If IRS produces a poor prediction, CEA will suppress the Activator Agent, replace the working memory chunks, and tell the Activator Agent to switch action.

It is known that humans are able to have sensory experiences in the absences of external stimuli as illustrated by experimental results of, e. g. Lee and Thompson [63]. It thus seemed reasonable to assume the existence of an 'inner sense' where sensory experiences and consequences of different behaviors may be anticipated. The idea of the existence of such an inner sense (or model) does not necessary go against the theory of embedded intelligence advocated by a number of researchers e. g., Brooks [64], Clancy [65], Clark [66], Pfeifer [49] who de-emphasize the role of internal world models and instead emphasize the situated and embodied nature of intelligence. An alternative to internal world models is the 'simulation hypothesis' by Hesslow [67] which accounts for the 'inner world' in terms of internal simulation of perception and behavior. Kawamura's approach may be termed as a "grounded



Cognitive Robotics, Figure 14
IMA-based cognitive robot architecture

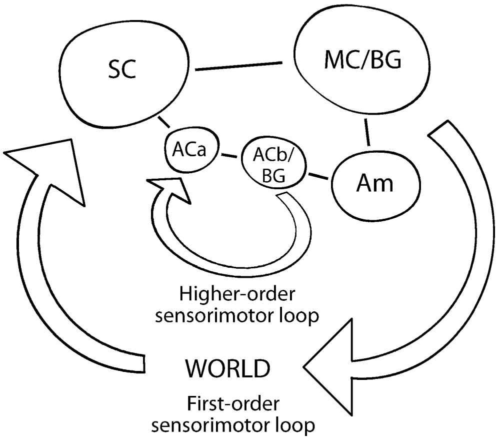


Cognitive Robotics, Figure 15
ISAC self agent cognitive cycle

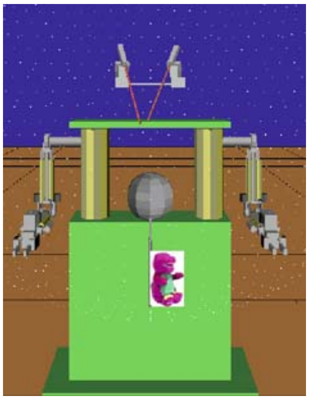
internal simulation" utilizing one type of internal representation of perception and behavior.

Design of the Internal Rehearsal System (IRS)

Brain-inspired internal simulation research has now moved into the robotics field. For example, Shanahan [68] proposed a cognitive architecture for a mobile robot, shown in Fig. 16, involving two sensory-motor loops: the reactive or first-order loop and the cognitive or higher-order loop. The first-order loop involves the sensory cortex (SC), motor cortex (MC), and basal ganglia (BG). This loop directly maps sensory input to motor actuation. The higher-order loop internally rehearses the decision from the first-order loop and changes the output of the system based on the observation of this rehearsal through the Amygdala (Am) or the emotion system.



Cognitive Robotics, Figure 16
A top-level schematic diagram of two interacting sensory-motor loops in the brain. Adapted from [68]



Cognitive Robotics, Figure 17
ISAC simulator displaying a Barney doll as a sphere

When IRS is invoked by CEA, it takes the current behavior chunk as the motor command and the current environment ISAC is in as the current state. After CEA selects a behavior to perform the skill described by the task, IRS internally rehearses the behavior with the percept corresponding to the current. If a collision occurs with the percept during the rehearsal, IRS returns the percept, the step in interpolation where the collision occurred, and the total number of joint steps in the interpolated motion to CEA.

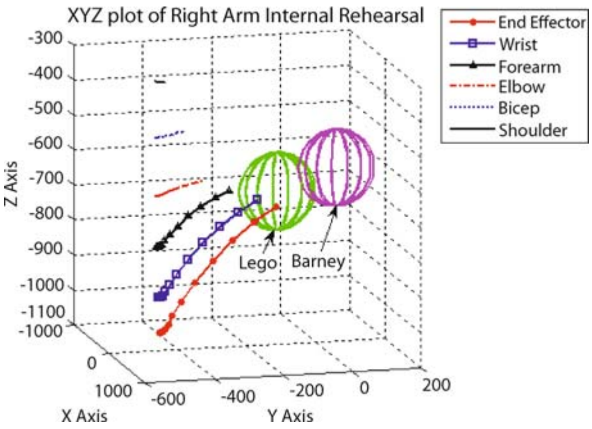
The following experiment is designed to evaluate how FRA, CEA, and IRS work together. The experiment involves two percepts: Barney (target) and a Lego toy (obstacle) (Fig. 17).

1. A task to *reach-to-Barney* is given to ISAC. FRA immediately places ReachRight and Barney into the working memory (WM) as chunks.
2. Using the chunks, IRS will try to reach to the Barney with the right arm, but predicts a collision with the Lego toy.
3. CEA will suppress the Activator Agent based on this prediction from IRS.
4. CEA will use the episodic retrieval technique and replace the chunk ReachRight to ReachLeft.
5. IRS will reach to the Barney with the virtual left arm. This reach will be successful.
6. CEA will let the Activator Agent proceed to reach to the Barney using the left arm.

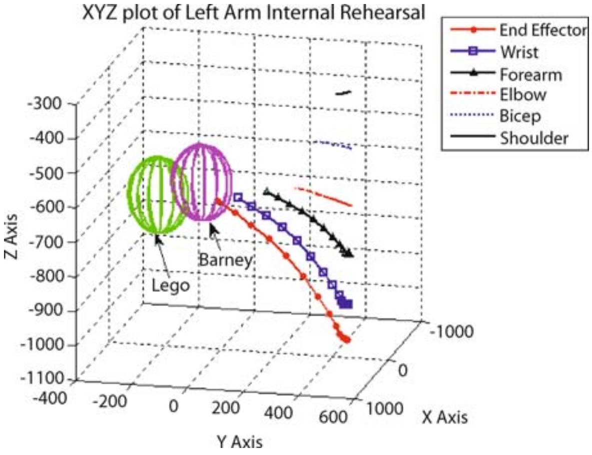
Performance

When ISAC was given a command to reach to the Barney, FRA placed two chunks "ReachRight" and "Barney" into the working memory (Note: ISAC normally begins with its right arm). Both the Activator Agent and IRS began to process these chunks. IRS completed the computation within 3.2 s and sent its results to CEA. At the same time, the Activator Agent sent a motion command to the Right Arm Agent to perform the reaching motion. The Right Arm Agent would take 11 s to perform this type of reach if no obstacle exists. When IRS finished, the following output was sent to CEA: [15 68 lego_toy]. This means that during the simulation, IRS determined a collision with the Lego toy in the fifteenth step of the interpolated *reach* behavior out of the total of 68 interpolated steps. Figure 18 shows the trajectories of the right arm collision points during the rehearsal. CEA took this result and determined that it did not reach to the Barney. CEA then suppressed the Activator Agent and prevented the right arm from further action.

CEA then decided to use the left arm and replaces the working memory chunks with "ReachLeft" and "Barney".



Cognitive Robotics, Figure 18
Right arm internal rehearsal



Cognitive Robotics, Figure 19
Left arm internal rehearsal

IRS and the Activator Agent were once again initiated, and IRS internally rehearsed the reach skill and determined no collision with the Lego toy (Fig. 19). Both the wrist and end effector points entered the Barney percept sphere on the sixteenth step out of the total of 69 interpolation steps. CEA determines this as a success and did not impede the Activator Agent thus allowing ISAC to reach to Barney using the left arm (Fig. 13).

Summary

This section illustrated how dual mechanisms of coupled real world of action-taking loop and off-line inner reasoning loop can work together to improve cognitive skills for a robot using internal rehearsal.

Future Directions

Cognitive robotics is still an evolving field with many possible and exciting future directions. To date, much progress has been made in architecture, perception, mobility, reasoning, learning from environmental embodiment, and advanced actuators. However, it will take years before we will see truly integrated cognitive robots in everyday life.

A few of the challenges to cognitive robots are listed below:

- Developing flexible systems, with ability to cope with multiple tasks, environments and disturbances
- Design strategies for learning by taking into account morphological and material constraints
- Cognitive skills development through social interaction
- Robots with mental states and emotions
- Modeling consciousness and its interaction with cognition
- Built in sensation and communication.

As robotic technology continues to penetrate every aspect of human society, the importance of social acceptance, such as trustworthiness and ethics, will become important. The field of cognitive robotics is expected to play a leading role in this area in the future. It is our hope that within decades, we will see true cognitive robots that will be accepted by the general public.

Bibliography

Primary Literature

1. Clark A, Grush R (1999) Towards a cognitive robotics. Adapt Behav 7(1):5–16
2. COGRIC (2006) Cognitive Robotics, Intelligence and Control. http://www.cogric.reading.uk/. Accessed 16–18 Aug 2006
3. Ballard DH (1991) Animal Vision. Artif Intell 48(1):1–27
4. Gibson JJ (1979) The Ecological Approach to Visual Perception. Houghton Mifflin, Boston
5. Trafton JG, Schultz AC, Bugajska M, Mintz F (2005) Perspective-taking with Robots: Experiments and models. In: IEEE International Workshop on Robots and Human Interactive Communication, Nashville, pp 580–584
6. Pfeifer R, Bongard J (2007) How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, MA
7. Alexander I (2005) The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines. Imprint Academic, UK
8. Holland O (ed) (2003) Machine Consciousness. Imprint Academic, UK
9. O'Reagan JK http://nivea.psycho.univ-paris5.fr/. Accessed 28 Aug 2008
10. Kawamura K et al (2006) From Intelligent Control to Cognitive Control. In: 11th International Symposium on Robotics and Applications (ISORA), Budapest

11. Kawamura K, Browne WN (2006) Tutorial on Cognitive Robots. IEEE RO-MAN, Hertfordshire
12. Wiener N (1948) Cybernetics, or Control and Communication in the Animal and the Machine. MIT Press, MA
13. Heylighen F, Joslyn C (2001) Cybernetics and Second-Order Cybernetics. In: Meyers RA (ed) Encyclopedia of Physical Science and Technology, 3rd edn. Academic Press, New York
14. Searle J (1999) The China Room. In: Wilson RA, Kei F (eds) The MIT Encyclopedia of the Cognitive Science. MIT Press, MA
15. COG www.ai.mit/edu/projects/humanoid-robotics-group/cog/
16. Honda http://world.honda.com/ASIMO/. Accessed 28 Aug 2008
17. Toyota http://www.toyota.co.jp/en/special/robot/. Accessed 28 Aug 2008
18. RoboGroup http://www.isc.cnrs.fr/dom/RoboGroup/htm/. Accessed 28 Aug 2008
19. RobotCub http://www.robotcub.org/. Accessed 28 Aug 2008
20. Synergistic Inteligence http://www.jeap.org/humanoids/pdf/MAsada.pdf. Accessed 28 Aug 2008
21. Takeno J et al (2005) Experiments and examination of mirror image cognition using a small robot. In: IEEE International Symposium on Computational Intelligence in Robotics and Automation. Espo, Finland, pp 493–498
22. Foresight http://www.foresight.gov.uk/index.asp. Accessed 28 Aug 2008
23. Anderson JR (1983) The Architecture of Cognition. Harvard University Press, MA
24. Newell A, Simon HA (1972) Human Problem Solving. Prentice-Hall, NJ
25. ACT-R http://act-r.psy.cmu.edu/about/. Accessed 28 Aug 2008
26. Anderson JR, Bothell D, Byme MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111(4):1036–1060
27. Laird J, Newell A, Rosenbloom P (1987) Soar – An architecture for general intelligence. Artif Intell 33:1–64
28. Soar http://sitemaker.umich.edu/soar/home. Accessed 28 Aug 2008
29. Newell A (1990) Unified Theories of Cognition. Harvard University Press, MA
30. Young RM, Lewis RL (1999) The Soar cognitive architecture and human working memory. In: Miyake A, Shah P (eds) Models of Working Memory: Mechanisms of active maintenance and executive control. Cambridge University Press, NY, pp 224–256
31. Kieras DE, Meyer DE (1997) An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Hum-Comput Interact 12:391–438
32. Meyer DE, Kieras DE (1997) A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic Mechanisms. Psychol Rev 104:3–65
33. EPIC http://www.umich.edu/~bcalab/epic.html. Accessed 28 Aug 2008
34. Abdi H, Valentin D, Edelman BE (1999) Neural Networks. Sage Publications, CA
35. Connectionism http://plato.stanford.edu/entries/connectionism/. Accessed 28 Aug 2008
36. Blazewicz J, Ecker K, Plateau B, Trystram D (eds) (2000) Handbook on Parallel and Distributed Processing. Springer, Germany
37. Haikonen PO (2006) Towards the times of miracles and wonder; a model for a conscious machine. Brain Inspired Cognitive Systems (BICS), Athens
38. Krichmar JL, Edelman GM (2003) Brain-Based Devices: Intelligent systems based on principles of the nervous system. In: IEEE/RSJ Int Conf on Intelligent Robotics and Systems. Las Vegas, pp 940–945
39. Edelman GM (1987) Neural Darwinism: The Theory of Neuronal Group Selection. Basic Books, NY
40. Krichmar JL, Reeke GN (2005) The Darwin Brain-Based Automata: Synthetic Neural Models and Real-World Devices. In: Reeke GN, Poznanski RR, Lindsay KA, Rosenberg JR, Sporns O (eds) Modeling in the Neurosciences: From Biological Systems to Neuromimetic Robotics. Taylor and Francis F, London
41. Schneider W (2000) Working Memory in a Multilevel Hybrid Connectionist Control Architecture (CAP2). In: Miyake A, Shah P (eds) Models of Working Memory: Mechanisms of active maintenance and executive control. Cambridge University Press, NY, pp 340–374
42. Schneider W, Shffrin RM (1977) Controlled and automatic human information processing: Detection, search, and attention. Psychol Rev 84:1–66
43. Holland O (2003) Exploration and high adventure: the legacy of Grey Walter. Phil Trans R Soc Lond A 361:2085–2121
44. Gallup GG Jr (1970) Chimpanzees: Self recognition. Sci 167: 86–7
45. CRONOS http://cswww.essex.ac.uk/staff/owen/machine/cronos.html. Accessed 28 Aug 2008
46. Arbib MA (1984) Forward. In: Braitenberg V (ed) Vehicles: Experiments in Systhetic Psychology. MIT Press, NY
47. Braitenberg V (1965) Taxis, kinesis, and decussation. Prog Brain Res 17:210–222
48. Iida F, Pfeifer R (2004) Self-Stabilization and Behavioural Diversity of Embodied Adaptive Locomotion. In: Iida F, Pfeifer R, Steels L, Kuniyoshi Y (eds) Embodied artificial intelligence. Springer, Berlin, pp 119–129
49. Pfeifer R, Scheier C (1999) Understanding Intelligence. MIT Press, MA
50. Mayo MJ (2003) Symbol Grounding and its Implications for Artificial Intelligence. In: Proc. 26th Australas Comput Sci Conf, vol 16. Adelaide, pp 55–60
51. Rubinstein JS, Meyer DE, Evans JE (2001) Executive control of cognitive processes in task switching. J Exp Psychol Hum Percept Perform 27(4):763–797
52. Phillips JL, Noelle DC (2005) A biologically inspired working memory framework for robots. In: IEEE International Workshop on Robots and Human Interactive Communication, Nashville, TN, pp 599–604
53. Gazzaniga MS (ed) (2004) The Cognitive Neurosciences III. MIT Press, Boston
54. Baars BJ (1997) In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press, Oxford
55. Browne WN, Tingley C (2006) Developing an Emotion-Based Architecture for Autonomous Agents. In: Third International Conference on Autonomous Robots and Agents (ICARA 2006), Palmerston North, pp 225–230
56. Asimov I (1950) I, Robot. Fawcett Crest, NY
57. Gee FC, Browne WN, Kawamura K (2005) Uncanny Valley revisited. In: Proc IEEE Robot and Human Interactive Communication, Nashville, pp 151–157

58. Potter SM (2001) Distributed Processing in Cultured Neuronal Networks. Prog Brain Res 130:1–14
59. CIS http://eecs.vanderbilt.edu/CIS/CRL. Accessed 28 Aug 2008
60. Pack T, Wilkes DM, Kawamura K (1997) A software architecture for integrated service robot development. In: IEEE Trans on Systems, Man and Cybernetics October, pp 3774–3779
61. IMA http://eecs.vuse.vanderbilt.edu/cis/concepts/ima.shtml. Accessed 28 Aug 2008
62. Kawamura K, Dodd W, Ratanaswasd P, Gutierrez R (2005) Development of a robot with a sense of self. In: 6th IEEE Int Symposium on Computational Intelligence in Robotics and Automation, Espoo
63. Lee DN, Thompson AI (1982) Vision in Action: The Control of Locomotion. In: Ingle D, Gooddale MA, Mansfield RJW (eds) Analysis of Visual Behavior. MIT Press, Cambridge, pp 411–433
64. Brooks RA (1986) A robust layered control system for a mobile robot. IEEE J Robot Autom RA-2(1):14–23
65. Clancey WJ (1997) Situated Cognition. Cambridge University Press, NY
66. Clark A (1997) Being There: Putting Brains, Body, and World Together Again. MIT Press, Cambridge
67. Hesslow G (2002) Conscious thought as simulation of behavior and perception. Trends Cogn Sci 6(6):242–247
68. Shanahan MP (2006) A cognitive architecture that combines internal simulation with a global workspace. Conscious Cogn 15:433–449

Books and Reviews

Bekey G (2005) Autonomous Robots: From Biological Inspiration to Implementation and Control. MIT Press, MA

Brooks RA (1999) Cambrian Intelligence: The Early History of the New AI. MIT Press, MA

Dennett DC (1991) Consciousness Explained. Little, MA

Gabriel M, Moore J (eds) (1990) Learning and Computational Neuroscience: Foundations of Adaptive Networks. MIT Press, MA

Haikonen PO (2003) The Cognitive Approach to Conscious Machines. Imprint Academic, UK

Hauser MD (2000) Wild Minds. Henry Holt and Company, NY

Meystel AM, Albus JS (2002) Intelligent Systems: Architecture, Design, and Control. Wiley, NY

Nolfi S, Floreano D (2004) Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-organizing Machines. Bradford Book, MIT Press, MA

Picard RW (1997) Affective Computing. MIT Press, MA

Siegwart R, Nourbakhsh IR (2004) Introduction to Autonomous Mobile Robots. Bradford Book, MIT Press, MA

Collective Transport and Depinning

Lei-Han Tang
Department of Physics, Hong Kong Baptist University,
Kowloon Tong, Hong Kong SAR, China

Article Outline

Glossary

Elastic manifold An elastic manifold is a spatially extended object whose energy is given by a quadratic function of the gradients of its transverse displacements. In statistical physics, it is used as a coarse-grained description of the low-energy modes in an ordered structure.

Pinning Pinning is a common phenomenon in condensed systems where the relevant degrees of freedom are trapped to an energy minimum and hence respond dynamically only when the external driving exceeds a certain threshold. Pinning is often caused by defects or impurities in the system but it may also be due to intrinsic properties such as the existence of a periodic lattice that breaks the translational symmetry of space.

Scaling Scaling describes power-law relationships among two or more physical quantities. Physical systems at continuous phase transitions are often found to exhibit scale invariance, i. e., structural and dynamic properties on different length and time scales can be mapped onto each other through suitable scale transformations. Such properties are characterized by a set of scaling exponents. The renormalization group theory, through its flow equations under a scale transformation, provides a systematic method to compute these exponents.

Universality Universality is a key concept in the classification of systems which exhibit scale invariance. Models in the same universality class have identical scaling properties and are described by the same set of scaling exponents. Therefore the identification of possible universality classes is one of the key issues in the study of, e. g., critical phenomenon at continuous phase transitions. The origin of universality is best illustrated by the fixed-point structure of the renormalization group flow equations. Studies have shown that dimensionality, symmetry, and conservation laws are the key factors that determine a particular universality class. There is, however, a great deal of theoretical interest to identify new principles that determine universality classes, or exceptions to the above rule, particularly in driven nonequilibrium systems.

Definition of the Subject

Collective transport takes place in systems which exhibit highly correlated response to external driving. This is in contrast to, e. g., electrical conduction in a normal metal, where free electrons drift independently under an applied electric field, leading to Ohm's law. Collective motion of microscopic degrees of freedom, on the other hand, often yields a nonlinear response or even threshold behavior, where steady-state transport sets in only when the driving exceeds a certain critical strength. In the subcritical regime, the collective modes are pinned by defects or impurities in the system. An increasing degree of correlation and corporation takes place as the threshold is approached. The subject plays an important role in the study of a broad class of solid state phenomena, including charge and spin density wave transport, hysteresis in dirty magnets, and nonlinear current-voltage characteristics in type-II superconductors. It has also found interesting applications outside physics, such as in crack propagation and earthquake modeling. Important theoretical developments that took place in the early 1990s culminated in the formulation of a functional renormalization group theory for the nonequilibrium depinning transition. The analytical framework enables systematic computation of the critical properties and, perhaps more importantly, elucidation of universality through its fixed point structures. Subsequent work by many research groups have established a close link between driven depinning and the sandpile models of self-organized criticality (SOC). This connection has been fruitfully explored to gain a better understanding of the long-range spatial correlations and intermittent temporal activities in the two classes of problems.

Introduction

The motion of a water drop down a glass window under gravity illustrates many salient features of collective transport and depinning, although the phenomenon itself is surprisingly rich in physics and chemistry when examined in microscopic detail [18,51,73]. The size of the drop is governed by the so called capillary length (typically of the order of a few millimeters) from elementary physics: On this scale the surface tension that makes a water drop spherical competes with the gravity which acts to deform the droplet. The actual descent of the droplet, which usually follows a zigzag path with an ever changing speed, is a result of many factors, not least the random force set by the local wettability of the glass surface. With a microscope, one may observe the hysteric advancement of the contact line separating wet and dry regions on the substrate. The interplay between the surface tension, the pinning force on the contact line, and the driving force provided by gravity gives rise to a complex and intermittent dynamical behavior encompassing a wide range of length and time scales.

Similar types of collective and intermittent transport exist in solids. In certain low-dimensional materials such as $NbSe_3$, periodic modulation of the electronic density, known as the charge-density-wave (CDW) [33], develops spontaneously at sufficiently low temperatures through the Peierls instability. An applied electric field in the direction of the charge modulation exerts a body force on the CDW much the same way as gravity does on the water droplet. However, in the presence of impurity atoms or crystal defects, the CDW does not move unless the electric field exceeds a certain threshold value, as seen in experiments [33,85]. The vortex lattice in type-II superconductors is another example of modulated electronic structures within a solid [44]. Pinning of the vortex lattice by intrinsic or artificial defects is essential for achieving a high critical supercurrent in these materials.

Yet another class of collective pinning phenomena in solids involves the dynamics of topological defects such as crystal dislocations [31] or magnetic domain walls [10]. These objects have internal dimensions lower than that of the embedding medium, and hence can explore inhomogeneities in the surrounding environment. The driven motion of these objects has a great effect on the physical properties of their host system, e. g. plastic deformation of a solid due to glide and climb motion of dislocations, and hysteresis effect [78,80,88] related to the pinning of domain walls separating regions of opposite magnetization.

The development of quantitative theories for impurity pinning and the driven depinning transition began in the 1970s after fundamental breakthroughs in many-body physics and equilibrium critical phenomena [29,44,45,50]. Research in this area has traditionally followed two separate approaches: The microscopic approach that attempts to explain the observed behavior starting from the fundamental laws of physics, and the phenomenological approach that focuses on the large-scale properties using the simplest models possible. The second approach, which is popular in the field of statistical physics, has the advantage of mathematical simplicity and clarity. It facilitates identification of the underlying symmetries of the original problem, and the establishment of universality classes through which model systems can be classified and compared. However, real physical systems often contain complications that prevent a direct comparison between model predictions and experimental observations (see, e. g. [85]). This is where the microscopic approach, popular among condensed matter theorists, comes to aid. In the best studied cases, a microscopic theory allows one to derive or estimate system-specific parameters and properties, and to suggest correct phenomenological models and possible improvements when discrepancies are found.

The present article focuses on the recent theoretical developments in the statistical physics of this class of problems. This is partly motivated by the significant progress that has been achieved in a quantitative characterization of the depinning transition over the past two decades. Similar to the critical phenomena in thermal equilibrium systems, static and dynamic fluctuations exhibit scaling properties with exponents that fall into well-defined universality classes determined by the symmetry and dimensionality of the problem. These concepts allow one to establish precise relationships among models proposed under different physical contexts. The successful application of the renormalization group methods to the depinning transition has led to a deeper analytical understanding of the observed critical phenomena. Our choice is also motivated by the close connection between depinning and the subject of self-organized criticality [5], which enjoys broad interest in the complexity community.

We shall start with a brief review of systems, mostly from solid state physics, where collective modes and depinning play an important role in the interpretation of the observed transport phenomenon. A generic mathematical formulation of this class of problems, known as the elastic manifold in a disordered potential, is introduced, along with a discussion of the complex energy landscape underlying its equilibrium and dynamic properties. This is followed by a description of the critical properties at the driven depinning transition and numerical results. The basic analytic structure of the threshold solution can be seen in a mean-field theory. We then summarize the main results of functional renormalization group calculations which provide a systematic and quantitative characterization of the depinning transition. The relationship between driven depinning and self-organized criticality is explained. The influence of medium anisotropy on the depinning transition is briefly discussed. Finally, we mention a few open problems in the field for future work.

Elastic Manifolds and Impurity Pinning

Elasticity, Order, and Symmetry Breaking

Although long-range spatial correlations that underlie collective transport can be generated dynamically under certain conditions, we shall focus on systems for which such correlations are already present before external driving is applied. In particular, we shall assume that the system is in a low temperature ordered state and responds elastically to external perturbations. This assumption allows us to concentrate on the large scale properties of the system in the depinning process, while leaving the microscopic, system-specific behavior to a separate discussion.

Elasticity is a familiar concept in macroscopic physics. An object is said to be *elastic* if it deforms under an applied force in a continuous and reversible manner. Microscopically, elasticity is intimately related to the existence of an ordered state that breaks a continuous symmetry [4]. This point can be appreciated with the example of a crystalline solid. Atoms in the solid form a periodic structure in space with energetically preferred unit cells and lattice constants. This state is said to break the translational symmetry of the particle system as the density of atoms is no longer uniform in space. A uniform translation of the structure does not lead to a change in the system energy. On the other hand, relative displacement of atoms distorts the unit cells and leads to an increase in energy which grows quadratically with the displacement. Thus, the system has acquired a form of rigidity by breaking a continuous symmetry to gain order. Elasticity is a manifestation of this rigidity when the ordered state is perturbed by external forces or internal impurities and defects.

Both charge-density waves and vortex lattices are periodic structures in space, and hence naturally their behavior is similar to a crystalline solid. There are, however, differences in the number of components needed to describe a generic deformation in each case. A CDW with a single modulation wave vector $\mathbf{Q}$ corresponds to an electron density $\rho(\mathbf{r}) = \rho_0 + \rho_1 \cos(\mathbf{Q} \cdot \mathbf{r} + \varphi)$, where ρ_0 is the average density and ρ_1 the modulation amplitude due to charge ordering. A uniform phase shift $\varphi \to \varphi + c$ moves the CDW uniformly against the underlying lattice. Weak deformation of the CDW is described by a spatially varying phase $\varphi(\mathbf{r})$. The elastic energy of a CDW depends quadratically on the phase gradient $\nabla\varphi$, with different elastic constants along and perpendicular to the modulation vector $\mathbf{Q}$ [29]. In the case of the vortex lattice, a two-component vector field $\mathbf{u}(\mathbf{r})$ perpendicular to the vortex lines is needed to describe a general distortion of the ideal structure. The construction of the elastic energy parallels that of a crystalline solid. In an isotropic medium, three elastic constants are needed to describe the energetics of a vortex line array [44,45].

Topological defects in an ordered structure also behave elastically when deformed from their ideal positions under a great variety of circumstances, though this form of elasticity has a different microscopic origin. A dislocation in a solid, for example, has a core where the atomic arrangement differs from elsewhere in the crystal. This gives rise to an excess amount of energy proportional to the length of the line. The interface between two bulk phases has an excess free energy (commonly known as the surface tension) proportional to the surface area for the same reason. This applies also to domain walls in a magnet.

Elastic Manifolds in a Disordered Potential

The energy of a topological defect is affected by the presence of impurities or point defects in the system. For example, an impurity atom sitting in the core of a dislocation has a different atomic environment and hence a different energy than when it resides in a normal region. This effect gives rise to a short-range and often attractive force between the impurity and the dislocation core. The tendency for the dislocation to distort itself in order to adapt to the impurity configuration is countered by the elastic energy cost. In general, the two competing forces lead to a large number of metastable conformations of the dislocation core, separated by energy barriers. The effect of impurities on interfaces and magnetic domain walls can be considered in a similar way.

A unified statistical mechanical description of impurity pinning can be formulated in terms of a D-dimensional elastic manifold embedded in d spatial dimensions [28,34,49]. Let $\mathbf{r} \in R^D$ be the internal coordinates of the manifold and $\mathbf{u}(\mathbf{r}) \in R^{d-D}$ be the transverse displacement of the manifold at point $\mathbf{r}$ with respect to a flat configuration. When the deformation is small, the excess volume of the manifold is a quadratic function of the gradient of $\mathbf{u}(\mathbf{r})$. The following energy function includes effects of both elasticity and impurities,

$$E(\{\mathbf{u}\}) = \int \mathrm{d}^D r \left[\frac{\gamma}{2} |\nabla \mathbf{u}|^2 + V_R(\mathbf{r}, \mathbf{u})\right] . \qquad (1)$$

Here γ is the stiffness constant of the manifold, and $V_R(\mathbf{r}, \mathbf{u})$ is a random potential arising from the interaction between the manifold and the impurities in the medium. In most theoretical treatments, the set of random variables $\{V_R(\mathbf{r}, \mathbf{u})\}$, which take particular values in a given sample, is assumed to be Gaussian distributed with the mean $\langle V_R(\mathbf{r}, \mathbf{u})\rangle = 0$ and the second moment $\langle V_R(\mathbf{r}_1, \mathbf{u}_1) V_R(\mathbf{r}_2, \mathbf{u}_2)\rangle = R(\mathbf{u}_1 - \mathbf{u}_2)\,\delta(\mathbf{r}_1 - \mathbf{r}_2)$. Here $\delta(\mathbf{r})$ is a short-ranged function which vanishes beyond a coarse-graining length $a_\parallel$.

The form of the disorder correlator $R(\mathbf{u})$ in the transverse directions is dictated by the symmetries of the original physical problem [28]. For contact interactions and randomly distributed impurities in the embedding space, $R(\mathbf{u})$ is short-ranged with a characteristic decay length $a_\perp$. A different situation is encountered when the manifold represents an interface that separates two bulk phases, each affected differently by the impurities which serve as a "random field" (i. e., the impurity atom has different chemical potential in the two bulk phases). The potential $V_R(\mathbf{r}, u)$ represents the cumulative effect of impurities swept by the interface when it is displaced to $u(\mathbf{r})$ from a reference configuration. Consequently $R(u) \sim |u|$ for large u.

A CDW in the presence of impurities can also be described by Eq. (1), where $u(\mathbf{r}) = \varphi(\mathbf{r})$ is a scalar field and $\mathrm{d} = D$. Since the impurity interacts with the local charge density which is a periodic function of the local phase, the potential $V_R(\mathbf{r}, \varphi)$ is also periodic in φ. Consequently, $R(\varphi)$ is periodic in φ as well.

Application of Eq. (1) to the vortex lattice requires special care. On scales smaller than the spacing between vortex lines, each vortex line behaves as an independent object. However, this description fails on large scales, where the periodicity of the lattice changes the nature of the problem [30,68].

Rugged Energy Landscape, Critical Dimension, and the Pinning Length

The two terms in Eq. (1) represent competing effects on the manifold: Elasticity favors a flat manifold with a constant $\mathbf{u}$, while the attractive force from impurities gives rise to a spatially varying $\mathbf{u}(\mathbf{r})$. The configuration $\mathbf{u}(\mathbf{r})$ which minimizes Eq. (1) satisfies the following force-equilibrium condition:

$$\gamma \nabla^2 \mathbf{u} - \nabla_{\mathbf{u}} V_R(\mathbf{r}, \mathbf{u}) = 0 . \qquad (2)$$

Complications arise when Eq. (2) has many solutions. Each solution corresponds to a local minimum of the energy function (1), separated by energy barriers from other local energy minima. The resulting energy surface (or landscape) is known as *rugged*.

Proper characterization of the energy landscape defined by Eq. (1) has been a long-standing problem in the statistical mechanics of disordered systems. For $D < 4$ and weak disorder, the ruggedness appears when the system size is greater than a characteristic length L_c along the manifold, known as the *pinning length*. This important observation was due to Larkin [44] for flux lines, Fukuyama, Lee and Rice [29,50] for CDWs, and Imry and Ma [37] for magnetic domain walls. On scales $L < L_c$, elasticity limits the transverse displacement $\mathbf{u}$ (known as roughness) to be within the respective correlation length $a_\perp$ of the impurity potential $V_R(\mathbf{r}, \mathbf{u})$, so that the manifold lies within a single minimum of the energy surface. On the scale L_c, the typical strength of the random force $\nabla_{\mathbf{u}} V_R$, averaged over a volume L_c^D, is estimated to be $\left|\overline{\nabla_{\mathbf{u}} V_R}\right| \simeq \frac{R^{1/2}(0)}{a_\perp} L_c^{-D/2}$. Balancing it with the elastic force $\gamma a_\perp / L_c^2$, one obtains,

$$L_c = \left(\frac{\gamma^2 a_\perp^4}{R(0)}\right)^{1/(4-D)} . \tag{3}$$

For $L > L_c$, Eq. (2) has an exponentially increasing number of solutions.

For $D > 4$, the elastic term in (2) dominates over the random force term on large length scales. Consequently, the rugged energy landscape is a small scale phenomenon which occurs when L_c is larger than the correlation length $a_\parallel$ of the disorder potential $V_R(\mathbf{r}, \mathbf{u})$ parallel to the manifold. From Eq. (3) we see that this condition requires the disorder strength to be sufficiently strong. Weak disorder is not able to produce pinning for manifolds with internal dimension greater than 4.

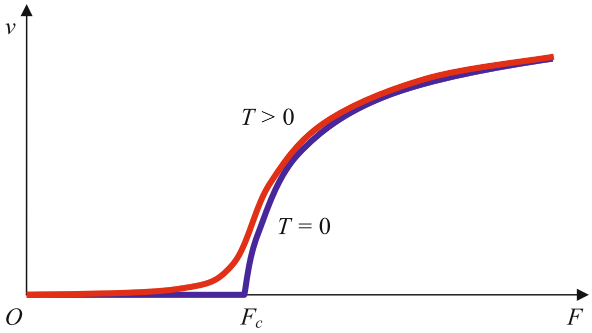
The qualitative analysis above shows that the energetics of the manifold problem are qualitatively different below and above four dimensions. The existence of a *critical dimension* $D_c = 4$ suggests systematic renormalization group approach to the problem. However, earlier attempts based on a perturbative treatment of the disorder potential failed to produce a renormalizable theory [1,22,32]. This was first achieved successfully by Daniel Fisher [28] in 1986 in the equilibrium case. Extension of the scheme to the driven depinning transition is discussed in Sect. "Analytical Treatments of the Depinning Transition and Universality".

Driven Depinning, Critical Properties, and Scaling

The Driven Depinning Transition

An applied force **F** coupled linearly to the displacement field $\mathbf{u}(\mathbf{r})$ tilts the equilibrium energy landscape as defined by Eq. (1) towards a particular direction. For small **F**, the manifold makes small adjustments locally to reach a new stationary state where force equilibrium is re-established, and this happens for the majority of solutions to Eq. (2). The number of such solutions, however, continues to decrease as **F** increases. When the magnitude of **F** exceeds a certain critical value, all stationary states disappear, and the manifold enters a running state. The transition from stationary to running states with increasing **F** is known as the *driven depinning transition*.

Figure 1 illustrates the dependence of the steady-state velocity v of the manifold against $F = |\mathbf{F}|$. In the absence of thermal fluctuations, there is a well-defined threshold F_c that separates the pinned from moving regimes. At finite temperatures, the transition from a pinned to a moving manifold is smeared out by thermally activated creep motion. At low temperatures, the rounding effect is weak except in a very small region around F_c.



Collective Transport and Depinning, Figure 1
Schematic plot showing the manifold velocity *v* against the driving force *F* at zero (*blue*) and finite (*red*) temperatures. Below the threshold F_c, the manifold is pinned by impurities in the medium, but thermal activation may generate a small but finite velocity

The discussion in Subsect. "Rugged Energy Landscape, Critical Dimension, and the Pinning Length" on the pinning length can be used to estimate the critical force needed to depin the manifold. For $D < 4$, the maximum pinning effect is seen on the scale L_c, where the typical strength of the first two terms in (5) is given by [10,26,50]

$$F_c \simeq \gamma a_\perp / L_c^2 = \frac{R^{2/(4-D)}(0)}{\gamma^{D/(4-D)} a_\perp^{(4+D)/(4-D)}} . \tag{4}$$

This is also the force needed to depin the manifold. For $D > 4$, pinning is possible only if $L_c > a_\parallel$ or $R^{1/2}(0) a_\perp^{-1} a_\parallel^{-D/4} > \gamma a_\perp / a_\parallel^2$, i. e., the strength of the pinning force is stronger than that of the elastic force on the minimal scale $a_\parallel$.

Continuum Model of the Manifold Dynamics

A dynamical model for the manifold can be constructed by assuming the motion to be completely overdamped. This is quite reasonable for the applications mentioned in Sect. "Introduction", where the displacement $\mathbf{u}(\mathbf{r}, t)$ represents a course-grained variable which changes on a time scale much longer than the relaxation time of underlying microscopic processes. With this assumption, the equation of motion for the manifold takes the form,

$$\mu^{-1} \frac{\partial \mathbf{u}}{\partial t} = \gamma \nabla^2 \mathbf{u} + \boldsymbol{\eta}(\mathbf{r}, \mathbf{u}) + \mathbf{F} .$$

Here μ is known as the mobility of the manifold, and $\boldsymbol{\eta}(\mathbf{r}, \mathbf{u}) = -\nabla_{\mathbf{u}} V_R(\mathbf{r}, \mathbf{u})$ is the random pinning force.

An interface is described by a height function $u(\mathbf{r},t)$, so that the dynamical equation, which defines the so-called *linear interface model* (LIM), takes a scalar form [10,26],

$$\mu^{-1}\frac{\partial u}{\partial t} = \gamma\nabla^2 u + \eta(\mathbf{r},u) + F\,. \tag{5}$$

The same equation applies to a CDW, but with $\eta(\mathbf{r},u+2\pi) = \eta(\mathbf{r},u)$ periodic in the phase variable [27]. Using the example of a single vortex line in three dimensions, Ertas and Kardar [23] have shown that the extra transverse dimensions do not change the main features of the depinning process. Hence Eq. (5) can be considered as the generic description of the driven manifold problem, where $u(\mathbf{r},t)$ stands for the component of the transverse displacement in the driven direction.

In Layman's terms, Eq. (5) may be viewed as describing the advance of a military front where the attacking side has more ammunition and manpower F but the defending side is able to exploit the hilly ground positions $\eta(\mathbf{r},u)$ to pose an effective resistance. In addition, stretching the front line into a convoluted form results in a decrease in attacking power and is hence discouraged!

For the analysis of Eq. (5), it is convenient to describe pinning effects in terms of the random force $\eta(\mathbf{r},u)$ instead of the random potential $V_R(\mathbf{r},u)$. Without loss of generality, one may assume the mean value of $\eta(\mathbf{r},u)$ to be zero. For Gaussian distributed random forces, it is suffice to specify the statistics of $\eta(\mathbf{r},u)$ with the correlator,

$$\langle\eta(\mathbf{r}_1,u_1)\,\eta(\mathbf{r}_2,u_2)\rangle = \Delta(u_1-u_2)\,\delta(\mathbf{r}_1-\mathbf{r}_2)\ . \tag{6}$$

On the "bare" scale, $\Delta(u) = -\partial^2 R/\partial u^2$ but as shown in [66] and [48], this relation breaks down in the driven case upon coarse-graining of the original degrees of freedom.

Critical Properties and Scaling Laws

Let us first consider the interface depinning problem as described by Eq. (5). This model has been studied extensively both analytically [14,66,69] and numerically [53,76] in recent years, and the main characteristics of the solution have been well-understood. These findings are summarized below.

The form of Eq. (5) suggests a "non-crossing condition" as first noted by Middleton [58]. Consider two interface configurations $u_1(\mathbf{r},t_0)$ and $u_2(\mathbf{r},t_0)$ at some initial time t_0. If $u_1(\mathbf{r},t_0) < u_2(\mathbf{r},t_0)$ for all $\mathbf{r}$ on the interface, one can easily show that $u_1(\mathbf{r},t) < u_2(\mathbf{r},t)$ at any later time $t > t_0$, i.e., the two solutions never touch if initially one is completely behind the other. This property in particular implies that the interface velocity $v = \overline{\mathrm{d}u/\mathrm{d}t}$ (averaged over all sites $\mathbf{r}$) in the steady state is a unique and continuous function of F, ruling out first order phase transition in this class of models.

On the moving side but close to the depinning threshold, advancement of the interface can be described with the help of Fig. 2. The thick line illustrates the interface position at a given time t_0. Pinning yields a roughness which grows as a power-law of the distance between two points on the interface, i.e.,

$$\left\langle\left[u(\mathbf{r},t_0)-u(\mathbf{r}',t_0)\right]^2\right\rangle \sim |\mathbf{r}-\mathbf{r}'|^{2\zeta}\ , \tag{7}$$

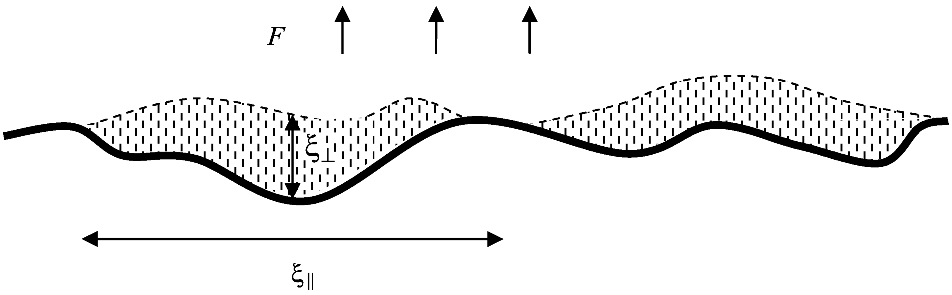
where ζ is known as the *roughness exponent*. This behavior holds on a range of length scales from the pinning length L_c to a correlation length $\xi_\parallel$ along the interface. The transverse displacement of the interface on scale $\xi_\parallel$ is given by $\xi_\perp \sim \xi_\parallel^\zeta$.

Motion of the interface within a correlation time τ also obeys scaling. On time scales shorter than τ, the interface advances through a sequence of rapid, localized movements (known as avalanches) of varying size up to the scale $\xi_\parallel$. The average time for a given site to move by a distance Δu grows as a power law: $\Delta t \sim \Delta u^{1/\beta}$. During this time, activities within a distance $l \sim \Delta t^{1/z}$ along the interface are correlated. The exponent $z = \zeta/\beta$ is known as the *dynamical exponent*. This scaling terminates when l reaches the correlation length $\xi_\parallel$, or $\Delta t = \tau \sim \xi_\parallel^z$, where the interface as a whole advances to a new disorder environment with a different distribution of pinning sites and pinning forces.

As F decreases towards F_c, the size of each correlated domain grows to infinity in a power-law fashion as well, e.g., $\xi_\parallel \sim |F-F_c|^{-\nu}$, where ν is known as the *correlation length exponent*. The three exponents ζ, z and ν together characterize the critical properties of the interface at the depinning transition. Through dimensional arguments, one may determine the critical behavior of quantities other than those discussed above. For example, the interface velocity near the transition can be estimated from $v \simeq \xi_\perp/\tau \sim \xi_\parallel^\zeta/\xi_\parallel^z \sim |F-F_c|^{\nu(z-\zeta)}$, hence the corresponding *velocity exponent* is given by,

$$\theta = \nu(z-\zeta) \tag{8}$$

The region between two interface configurations separated by time τ, as shown in Fig. 2, can be described as a set of "bubbles", each representing a correlated volume of base area $\xi_\parallel^D$ and height $\xi_\perp$. The average strength of the pinning forces within each bubble has a variation of the order $\xi_\parallel^{-D/2}\xi_\perp^{-1/2}$ from the system-wide average. Variation of this quantity among bubbles should be smaller than the excess driving force $F-F_c$ so that depinning can take place



Collective Transport and Depinning, Figure 2
An interface above but close to the depinning threshold. Depinning events within a correlation length $\xi_\parallel$ and correlation time τ are highly correlated and obey scaling at the transition. Shaded area represents the volume swept by the interface over the time interval τ

Collective Transport and Depinning, Table 1
Scaling exponents at the depinning transition. Numbers in parentheses indicate uncertainties in the last digit

	Interface[a]			CDW			MF $D \geq 4$
	$D = 1$	$D = 2$	$D = 3$	$D = 1$	$D = 2$	$D = 3$	
ζ (roughness)	1.25(1)	0.75(2)	0.35(1)	0			0
z (dynamic)	1.42(4)	1.56(6)	1.75(15)	1	1.32(4)[b]	1.65(6)[b]	2
ν, ν_f (correlation length)	1.33(2)	0.80(1)	0.606(4)	0.4(1)[c] 2.01(2)	0.5(1)[c] 0.98(3)[b]	0.5(1)[c] 0.68(4)[b]	$1/2$
θ (velocity)	0.25(3)	0.64(2)	0.84(2)	0.45(5)[c]	0.64(3)[b,c]	0.81(3)[b,c]	1

[a] Leschhorn H [53]; [b] Middleton AA, Fisher DS [61]; Narayan O, Middleton AA [67]; [c] Mayer CR, Sethna JP [63].

uniformly across the system. This condition is fulfilled if the correlation length exponent satisfies the following inequality [35,66,69],

$$\nu \geq \frac{2}{D + \zeta} \tag{9}$$

The above description of interface depinning applies also to the CDW with one important caveat. Due to the noncrossing condition, steady-state motion of the CDW at $F > F_c$ is periodic in time, i. e., $\varphi(\mathbf{r}, t + \tau) = \varphi(\mathbf{r}, t) + 2\pi$ for all $\mathbf{r}$, with τ being the period of the attractor [58]. The dynamic phase advance at different sites has thus bounded variations described by $\zeta = 0$. Above but close to F_c, the activity at a given site, as measured by the phase velocity $\dot{\varphi}(\mathbf{r}, t)$, is typically concentrated in time windows much shorter than T, but acquires long-ranged spatial correlations up to a correlation length $\xi_\parallel \sim (F - F_c)^{-\nu}$. Both analytical [65] and numerical calculations [63] indicate that $\nu = 1/2$ in all dimensions, therefore violating Eq. (9). Although the origin of this behavior has not been settled completely [67], a plausible explanation is that fluctuations of the pinning force in a given region of the system are compensated by a static phase distortion $\varphi_0(\mathbf{r}) = \varphi(\mathbf{r}, t = 0)$, so that the system behaves much more homogeneous than the naïve estimate used to obtain Eq. (9). The sample-to-sample fluctuations of F_c, on the other hand, have been shown recently [7,25] to obey a Gaussian distribution with a width proportional to $L^{-D/2}$as predicted, where L stands for the linear system size.

Numerical Results for the Critical Exponents

The depinning transition of the elastic manifolds and the CDW has been studied extensively using various lattice models. Middleton and Fisher [60,61], Myers and Sethna [63], and Narayan and Middleton [67] simulated a discretized version of Eq. (5) for CDW depinning in one to three dimensions. The random force is given by $\eta(\mathbf{r}, \varphi) = V \sin(\varphi - \beta(\mathbf{r}))$, where V is the strength of the pinning force, and $\beta(\mathbf{r})$ is the preferred phase at site $\mathbf{r}$, chosen randomly from site to site. Values of critical exponents as determined in their numerical work are given in Table 1, which show good agreement with the analytic results in Sect. "Analytical Treatments of the Depinning Transition and Universality". The exponent ν_f (second row) is determined from quantities related to the avalanche propagation below the depinning threshold.

Leschorn [53] carried out large-scale simulations of a discretized version of the LIM. Exponents he obtained are also given in Table 1. These values are in good agreement with more recent studies [76]. Note that the roughness exponent ζ for a one-dimensional interface at the depinning threshold is greater than one. This gives rise to a subtlety in using Eq. (7) to measure ζ as discussed in [56].

Analytical Treatments of the Depinning Transition and Universality

The depinning transition differs from the usual critical phenomena in thermal equilibrium systems in terms of the vast separation of fast and slow time scales. Therefore the successful development of a renormalization group theory in the early 1990s to effectively capture this unique feature was an important milestone in the analytical studies of nonequilibrium phase transitions. The work acquired broader significance due to the later discovered correspondence between depinning and sand pile models of self-organized criticality (SOC), which we discuss in Sect. "Self-Organized Criticality".

Mean-Field Theory

Fisher [27] and others [43,52,55,65] considered a mean-field approximation to Eq. (5) which can be treated analytically. It is instructive to examine the calculation in some detail here as the solution reveals several important features of the depinning transition which extend to lower dimensions, while the mathematical manipulations can be kept at an elementary level.

In sufficiently high dimensions, one may approximate the Laplacian in Eq. (5) by a spring force, whose equilibrium point is set by the mean interface position $\bar{u}(t)$. The resulting dynamical equation for a given site on the interface takes the form,

$$\mathrm{d}u/\mathrm{d}t = \gamma(\bar{u} - u) + \eta(u) + F\,. \tag{10}$$

For simplicity we have dropped μ in (5) through a redefinition of t. In the steady-state, $\bar{u} = vt$, where v is the interface velocity to be determined self-consistently from the solution of Eq. (10). Since the disorder $\eta(u)$ is uncorrelated along the interface, the system-wide average $\bar{u}$ can be replaced by an average over the distribution of $\eta(u)$.

Following the discussion in [55], we adopt a moving frame and define $x = u - vt - (F - v)/\gamma$ to be the displacement away from the equilibrium position when the random force $\eta(u)$ is absent. In terms of x, Eq. (10) reads (after a rigid translation of $\eta(u)$),

$$\mathrm{d}x/\mathrm{d}t = -\gamma x + \eta(vt + x)\,. \tag{11}$$

The self-consistency condition becomes,

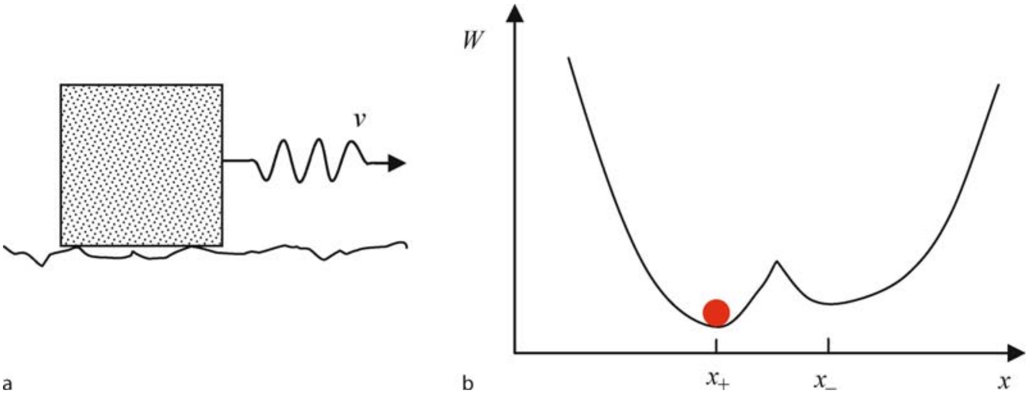
$$\gamma\,\langle x\rangle = -F + v\,. \tag{12}$$

The driving force F now disappears from the equation of motion (11) and can be computed from (12) as a function of v, once a solution to (11) is found.

As illustrated in Fig. 3, Eq. (11) describes the dynamics of an overdamped particle in a moving potential,

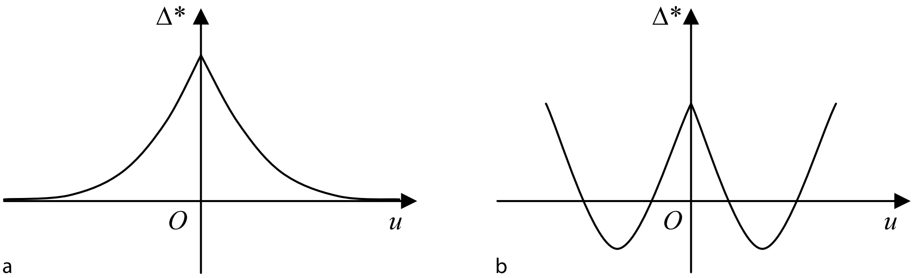
$$W(x,t) = \frac{\gamma}{2}x^2 + U(vt + x)\,. \tag{13}$$

The random part $U(vt + x) = -\int_{x_0}^{vt+x} \eta(u)\mathrm{d}u$ travels at a constant velocity v to the left or to the right depending



Collective Transport and Depinning, Figure 3

The mean-field model of depinning. a Block (representing the interface) dragged by a spring whose right end moves at a slow but constant velocity *v*. The block is sitting on a rough surface and immersed in a very viscous fluid, so that inertia effects can be neglected. b The effective potential *W* on the block at a particular instant when viewed in a co-moving frame. *W* is a sum of two parts: A quadratic spring potential plus a random potential *U* which travels slowly to the left. In the absence of thermal motion, the block (represented by the red ball in the figure) traces the left most minima x_+ of *W* until it disappears, followed by an "avalanche" into the next available minimum



Collective Transport and Depinning, Figure 4
Fixed point solutions to Eq. (14) for a the linear interface models and b the charge-density wave depinning

on the sign of v. In the quasi-static limit $v \to 0$, the particle stays in a local minimum of the potential W for the whole time, and is interrupted occasionally by jumps when the minimum it follows disappears. When the potential travels to the left ($v > 0$), the particle traces the leftmost minimum x_+ of W. The opposite occurs when the potential travels to the right.

With the above picture in mind, we may use Eq. (12) to calculate the thresholds $F_c^+ = -\gamma \lim_{v\to 0+}\langle x\rangle$ and $F_c^- = -\lim_{v\to 0-}\langle x\rangle$ for forward and backward depinning, respectively. For a continuous, slow-varying function $\eta(u)$ with a sufficiently small amplitude, the potential W has a unique minimum all the time, and hence $F_c^+ = F_c^- = 0$, i. e., the system is never pinned. However, for a discontinuous function $\eta(u)$ (or for sufficiently strong disorder), the potential W has, from time to time, more than one minimum, and hence $F_c^+ = -F_c^- > 0$. Note that only upward jumps in η (i. e. a sudden weakening of the pinning force) give rise to upward cusps in the potential which generate double minima in W.

When the potential $W(x,t)$ travels at a small but finite velocity v, the ball shown in Fig. 3b is displaced to the leftmost minimum of W by an amount proportional to v due to viscosity. From Eq. (12) and the value of F_c determined above, one finds $v = a(F - F_c)$, where a is a numerical constant. This behavior is confirmed by exact solution of a "two-state" model for the pinning force [52]. The mean-field velocity exponent is thus given by $\theta_{\text{MF}} = 1$.

Functional Renormalization Group

Equation (5) has been studied analytically since the 1970s. Earlier attempts by Efetov and Larkin [22] and others [26,43] based on a perturbative expansion of the noise term around a flat reference state ran into divergences for $D < D_c = 4$, which can not be cured with the usual renormalization group (RG) procedure. Breakthroughs were achieved in 1992 by Narayan and Fisher [65] for the CDW depinning and by Nattermann et al. [69] for the driven interface. In the field theory jargon, the model has an infinite number of relevant operators that correspond to the coefficients in a power-law expansion of the random force correlator $\Delta(u)$. Upon momentum shell integration on a running scale L, the one-loop corrections to these coefficients can be summarized in terms of a functional RG flow for $\Delta(u)$,

$$\frac{\mathrm{d}\Delta(u)}{\mathrm{d}\ln L} = -cL^{\varepsilon}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\left[\frac{1}{2}\Delta^2(u) - \Delta(u)\Delta(0)\right] . \qquad (14)$$

Here $\varepsilon = 4 - D$ and c is a constant. Dependence on L can be absorbed with the scaling transformation $u \to L^{\zeta}u, \Delta \to L^{2\zeta-\varepsilon}\Delta$. The resulting flow equation has an infinite number of fixed points, which are distinguished by the number of nodes of the function Δ. Figure 4 shows two such solutions, the first with no node at $\zeta = \varepsilon/3$, and the second with infinitely many nodes (periodic) at $\zeta = 0$. The periodic fixed point is associated with the CDW depinning transition. The one with no node describes the interface depinning transition. The most prominent feature of the fixed point function Δ^* is the cusp singularity at $u = 0$, which is responsible for the failure of previous RG treatments.

Equation (14) is supplemented with a RG equation for the mobility μ that determines the dynamical response on scale L,

$$\frac{\mathrm{d}\ln\mu}{\mathrm{d}\ln L} = -c\Delta''(0+)L^{\varepsilon} . \qquad (15)$$

The elastic constant γ, on the other hand, is not renormalized. The latter property yields the scaling relation

$$\nu(2-\zeta) = 1 . \qquad (16)$$

The functional RG analysis has been carried out to two-loop order by Le Doussal and collaborators [14,48]. This study is particularly useful as it helps to clear up several subtleties which can not be easily resolved in the one-loop calculation. For interface depinning, exponents up to the second order in ε are given by,

$$\begin{aligned}\zeta &= \frac{1}{3}\varepsilon + 0.047771\varepsilon^2\\ z &= 2 - \frac{2}{9}\varepsilon - 0.0432087\varepsilon^2\\ \nu &= \frac{1}{2-\zeta} = \frac{1}{2} + \frac{1}{12}\varepsilon + 0.0258316\varepsilon^2\\ \theta &= \nu(z-\zeta) = 1 - \frac{1}{9}\varepsilon + 0.040123\varepsilon^2\,.\end{aligned} \tag{17}$$

Good agreement is seen with the numerical values given in Table 1.

For CDW depinning, the two-loop analysis yields the following exponents,

$$\begin{aligned}\zeta &= 0\\ z &= 2 - \frac{1}{3}\varepsilon - \frac{1}{9}\varepsilon^2\\ \nu &= \frac{1}{2-\zeta} = \frac{1}{2}\\ \theta &= \nu(z-\zeta) = 1 - \frac{1}{6}\varepsilon - \frac{1}{18}\varepsilon^2\,.\end{aligned} \tag{18}$$

Interestingly, here the expression for the dynamical exponent z truncated at first order in ε appears to be in better agreement with the simulation results given in Table 1.

Very recently, Rosso, Le Doussal and Wiese [77] proposed and implemented a scheme to measure numerically the effective pinning force and the renormalized correlator Δ on large distances. The idea is similar to the mean-field model discussed above, with the particle replaced by the center of mass of the manifold in a weak confining potential that introduces a running cut-off to spatial correlations. Their numerical results provide direct confirmation of the cusp singularity in Δ, which has become the hallmark of the RG theories for the manifold problem.

Contact Line Depinning

The above analysis has been extended by Ertas and Kardar [24] to the contact line depinning problem. The increase in surface free energy (or surface area) due to a deformation of the contact line from an ideal straight configuration has been worked out previously by Joanny and de Gennes [39]. Result of this analysis, generalized to a "contact manifold" with D internal dimensions, is that the Laplace term in Eq. (5) is replaced by a nonlocal term, $\gamma\nabla^2 u \to \int \mathrm{d}^D\mathbf{r}' K_D(\mathbf{r}-\mathbf{r}')[u(\mathbf{r}',t)-\bar{u}(t)]$. Here $\bar{u}(t)$ denotes the mean displacement. The function $K_D(\mathbf{r}) \sim |\mathbf{r}|^{-D-1}$ describes the decay of the long-ranged elastic coupling along the contact manifold, assuming the modes in the "bulk" have relaxed. This form of elastic coupling changes the critical dimension to $D_c = 2$. In terms of $\varepsilon = 2 - D$, the critical exponents of the depinning transition from a two-loop renormalization group calculation by Le Doussal et al. [48] are given by,

$$\begin{aligned}\zeta &= \frac{1}{3}\varepsilon + 0.13245\varepsilon^2\\ z &= 1 - \frac{2}{9}\varepsilon - 0.1132997\varepsilon^2\\ \nu &= \frac{1}{2-\zeta} = 1 + \frac{1}{3}\varepsilon + 0.24356\varepsilon^2\\ \theta &= \nu(z-\zeta) = 1 - \frac{2}{9}\varepsilon - 0.1873737\varepsilon^2\,.\end{aligned} \tag{19}$$

Interestingly, the same long-range coupling is found to describe tensile crack propagation [75]. Numerical simulations [21] of a discretized long-range model in $D = 1$ ($\varepsilon = 1$) yield $\zeta = 0.385(5)$, $z = 0.770(5)$, $\nu = 1.625(10)$, $\theta = 0.625(5)$, in good agreement with the theory.

Self-Organized Criticality

Close to the depinning threshold, motion of the system is intermittent. The growth activities are highly localized and exhibit strong correlations over time. The natural separation of slow and fast events at the onset of depinning invites comparison to models of self-organized criticality (SOC) and of earthquakes, which were made popular through the seminal work of Bak, Tang and Wissenfeld (BTW) [5,6] on the "sandpile" model. This observation prompted a number of detailed investigations of the microscopic processes leading to the build-up of spatiotemporal correlations as the depinning threshold is approached from below, as summarized in [72]. Subsequent work by several groups [67,71] established a direct link between these two classes of models. This development has also contributed to a better understanding of SOC in automaton models. It opens the door to a systematic exploration of their scaling properties which were hitherto hindered by the lack of a suitable continuum representation.

The mapping between depinning and SOC automaton models is most easily understood by considering $u(\mathbf{r},t)$ in Eq. (5) as the accumulated activity of a given lattice site $\mathbf{r}$ after time t [71]. Take for example the original sandpile model introduced by BTW on a square lattice. The system is specified by an integer height $z(x,y)$ on each lattice site (x,y). Starting from a random set of values $z_0(x,y)$, "sand" is gradually added through a sequence of events $z(x, y) \to z(x, y) + 1$, each time only a single site is chosen randomly from all available sites in the system for updating. If the operation leads to an overflow, i. e., $z(x, y) > K$, then toppling takes place according to the rule,

$$\begin{aligned} z(x, y) &\to z(x, y) - 4 \\ z(x \pm 1, y) &\to z(x \pm 1, y) + 1 \\ z(x, y \pm 1) &\to z(x, y \pm 1) + 1 \,. \end{aligned} \tag{20}$$

This process is continued until no site has a height exceeding K. Let $u(x, y; t)$ be the total number of toppling events at site (x, y) from the beginning of the process. Counting the number of sand grains received and given away by the site yields immediately

$$\begin{aligned} z(x, y; t) = z_0(x, y) &- 4u(x, y; t) + u(x + 1, y; t) \\ &+ u(x - 1, y; t) + u(x, y + 1; t) \\ &+ u(x, y - 1; t) \\ = z_0(x, y) &+ \nabla^2 u \,, \end{aligned} \tag{21}$$

where ∇^2 is the Laplace operator on the lattice. The condition $z(x, y; t) > K$ for toppling $u \to u + 1$ is thus equivalent to

$$\nabla^2 u + z_0(x, y) - K > 0 \,. \tag{22}$$

The above growth rule is almost identical to a lattice automaton model introduced by Narayan and Middleton [67] for CDW depinning. Comparing (22) to Eq. (5), we see that here (i) u is restricted to take only integer values plus a random, site dependent shift $\beta(\mathbf{r})$, which can be considered as the preferred phase set by the local disorder; (ii) The dynamical rules obey a global symmetry $u \to u + 1$, as in CDW depinning. The usual wisdom in critical phenomena suggests that the two models are in the same universality class due to property (ii), while property (i) is a microscopic detail which does not affect the scaling properties at the depinning transition. Indeed, various scaling exponents determined through analytical and numerical studies of the two models are shown to be consistent with each other [67,82]. Subtle differences exist, however, in the way the system is driven to criticality [67,72]. The effect of boundary conditions also need to be treated with care [67].

To eliminate inertia effects in real sand pile avalanches [64] which prevented direct comparison with the BTW model, the Oslo group [15] designed a two-dimensional rice pile experiment that displays power-law scaling and self-organized criticality. The group also proposed an automaton model, known as the Oslo model, which yields critical exponents in good agreement with the experiment. The Oslo model differs from the BTW model only in the choice of the threshold value K for toppling: Instead of taking a constant value, K is site-dependent. Its value is updated after each toppling event at the site. Consequently, condition (22) changes to,

$$\nabla^2 u + z_0(\mathbf{r}) - K(\mathbf{r}, u) > 0 \,. \tag{23}$$

The symmetry $u \to u + 1$ is no longer present. A similar discussion as above links the Oslo rice-pile model to the LIM of depinning [8,71]. Other SOC models, such as the Manna model [57] and the Zaitsev model [87], have similar updating rules as the Oslo model and hence belong to the LIM universality class as well. The mechanism of self-organization and the development of spatiotemporal correlations in the SOC and depinning models have been discussed in detail by Paczuski, Maslov and Bak [72], who have also identified a large number of scaling relations that give all other exponents in terms of two basic ones.

As an example, let us consider the power-law distribution of the avalanche size which can be measured, e. g., by the total number of toppling events s when a new grain is added. For a system of linear size L, the distribution satisfies the finite-size scaling,

$$P(s, L) = s^{-\tau} F(s/L^{\delta}) \,. \tag{24}$$

The mapping discussed above identifies $\delta = D + \zeta$, i. e., the dimension of the total volume swept by the interface in a system-wide avalanche. On the other hand, a sum rule yields the following expression,

$$\tau = 2 - \frac{\zeta + \nu^{-1}}{\zeta + D} \tag{25}$$

For example, the exponents $\zeta = 0$ and $\nu = 1/2$ for the CDW depinning predicts $\tau = 2\,(D - 1)/D$, in excellent agreement with simulation results on the sandpile model in two and three dimensions [82]. Care must be taken, however, in applying Eq. (25) to the $D = 1$ rice pile model in the boundary driven case, as noted by Paczuski and Boettcher [71].

Interface Depinning in Anisotropic Medium

Easy and Hard Directions of Depinning

The linear interface model defined by Eq. (5) together with Eq. (6) for the random force is statistically invariant under the *tilt transformation* $u \to u + \mathbf{s} \cdot \mathbf{r}$, where $\mathbf{s}$ is the slope of the interface. Hence the depinning threshold F_c, as well as other critical properties, are independent of the orientation of the interface. This property holds for an interface moving in an isotropic medium.

When the medium is anisotropic, one or more of the model parameters that enter the estimate Eq. (4) for F_c may change with **s**. Consequently, the depinning threshold $F_c(\mathbf{s})$ becomes orientation-dependent [83]. An interface oriented in the "easy" direction enters the moving phase at the weakest driving. However, a moving interface in this direction is generically unstable against faceting towards directions of slower growth, particularly near the depinning threshold. To illustrate this point, we may consider a part of the interface that is slowed down by an exceptionally strong pin. As the rest of the interface moves forward at a higher velocity, a local tilt develops that moves the interface away from the easy direction. This further strengthens the pinning effect, producing a cone-like structure. Indeed, the depinning transition in the easy direction has been shown to be of first order with a velocity jump [38].

The story is different for an interface oriented in the "hard" direction. Both numerical [2,3,11,84] and experimental [2,11,12] studies have shown that the depinning transition here is continuous, but the critical exponents take different values from those of the LIM. Yet another set of critical exponents are encountered when the interface is forced to tilt away from the hard direction by, say boundary conditions [83]. The resulting depinning problem is related to *directed percolation*, which we consider in some detail below.

A Lattice Automaton

Tang and Leschhorn [84] introduced a lattice automaton for imbibition in a two-dimensional porous medium. A very similar model was studied by Buldyrev et al. [11] around the same time and reported together with experimental results. The automaton model is defined as follows. On a square lattice, each site (i, j) is assigned a random pinning force $\eta(i, j)$, uniformly distributed in the interval [0,1). At $t = 0$, the interface h_i is completely flat. In each time step, growth $h_i \to h_i + 1$ is performed in parallel when either of the following two conditions is satisfied: (i) the random force $\eta(i, h_i)$ on the interface at column i is less than a pre-specified driving force f; or (ii) $h_i < h_{i-1} - 1$ or $h_i < h_{i+1} - 1$. Thus to obtain a completely pinned interface, we must have $|h_i - h_{i-1}| \leq 1$ and $\eta(i, h_i) > f$ (called a blocking site) for all i. This condition requires the existence of a *directed percolating path* through blocking sites. Such paths exist if the density p of blocking sites exceeds a critical value $p_c \simeq 0.539$. Consequently, the depinning threshold is given by $f_c = 1 - p_c \simeq 0.461$. It is possible to show that, for $f < f_c$, the interface stops at the first directed percolating path that lies fully above its initial position.

The roughness of the incipient directed percolating path at $p = p_c$ is $\zeta = \nu_\perp / \nu_\parallel \simeq 0.63$, where $\nu_\parallel = 1.733$ and $\nu_\perp = 1.097$ are exponents that characterize the divergence of parallel and perpendicular correlation lengths, $\xi_\parallel \sim |p - p_c|^{-\nu_\parallel}$, $\xi_\perp \sim |p - p_c|^{-\nu_\perp}$, respectively [41]. This is also the roughness exponent assumed by the interface at the depinning threshold $f = f_c$. Simulations [84] have shown that the dynamic exponent $z = 1$ at the depinning transition of the above model, while the crossover length exponent $\nu = \nu_\parallel$. From Eq. (8) one obtains $\theta = \nu_\parallel - \nu_\perp \simeq 0.63$.

The orientational dependence of the depinning threshold in the automaton model has been studied by applying the helical boundary condition $h_{i+L} = h_i + sL$ which forces a global tilt [83]. The threshold force is indeed found to be lower. This behavior matches well with the well-known property of directed percolation: For $p > p_c$, percolating paths fall within a cone of opening angle $\varphi(p)$ around the symmetry direction [41]. If the average tilt s of the interface exceeds this angle, no spanning percolating path exists and the interface is free to move. The roughness exponent of the cone boundary is $\zeta = 1/2$, which belongs to yet another universality class of interface depinning away from a symmetry direction [83].

The above automaton model has been generalized by Buldyrev et al. [12] to higher dimensions. The directed percolating strings are replaced by directed percolating surfaces which have been considered in the context of resistor-diode percolation [20]. The roughness exponent of such a surface at the percolation threshold in $D + 1$ dimensions is given in Table 2. Havlin et al. [36] investigated in detail the temporal sequence of growth activities at the depinning transition, which exhibits an interesting super-diffusive behavior. They argued that the dynamic exponent z that describes the growth in size of the affected region after an initial depinning event is related to the scaling of minimal path length with Euclidean distance on the critical (isotropic) percolation cluster in D dimensions. The values of z from their work are also given in Table 2.

As in the $1 + 1$ dimensional case, a tilted interface has a lower depinning threshold. Just above the depinning threshold, the interface decomposes into stripes of inactive regions perpendicular to the tilt, separated by active fronts that propagate towards the "easy" direction at a finite velocity. Based on this phenomenology, Tang, Kardar and Dhar [83] determined exponents characterizing the depinning transition of a tilted interface. In particular, the velocity exponent $\theta = 1$ in all dimensions.

Collective Transport and Depinning, Table 2

Summary of the exponents for directed percolation depinning along the hard direction. After [2]

	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$	MF
ζ (roughness)	0.63(1)	0.48(3)	0.35(1)	0.27(5)	0.25(5)	0.2(2)	0
z (dynamic)	1	1.15(5)	1.36 (5)	1.58(5)	1.7(1)	1.8(2)	2
ν (correlation length)	1.73(2)	1.16(5)	0.95(10)	0.66(10)	0.6(1)	0.5(1)	$1/2$
θ (velocity)	0.58(7)	0.8(2)	1.0(2)	1.0(2)			1

Quenched KPZ Equation

The slope-dependent depinning threshold can be modeled explicitly by adding the Kardar–Parisi–Zhang (KPZ) term to Eq. (5). The resulting equation of motion takes the form,

$$\mu^{-1}\frac{\partial u}{\partial t} = \gamma\nabla^2 u + \frac{1}{2}\lambda(\nabla u)^2 + \eta(\mathbf{r}, u) + F\,. \qquad (26)$$

In the original proposal [40], the λ-term is due to a kinematic effect known as lateral growth, and hence its strength is proportional to the interface velocity v. At the depinning threshold, however, this term is present only when the medium is anisotropic. A positive λ describes depinning along a hard direction, while a negative λ corresponds to growth along an easy direction. Directed numerical integrations [17,54] of Eq. (26) in $(1+1)$ dimensions yielded exponents consistent with the directed percolation models. The relevance of the KPZ nonlinearity in modifying the critical behavior of the LIM has also been confirmed by renormalization group calculations [47,81].

Future Directions

Through the intensive work by many groups in the past fifteen years, the depinning transition of elastic manifolds in a disordered medium has emerged as one of the best understood nonequilibrium critical phenomena with nontrivial scaling properties. The discovery of a renormalizable continuum theory provided the much needed theoretical foundation for the identification of universality classes and symmetry principles. Owing to this remarkable development, models and experimental systems that differ in microscopic details can be compared and classified with regard to their threshold behavior. More recent refinements of the theory by the Paris group (e. g., [46]) have yielded deeper insights on the role of the cusp singularity of the random force correlator in capturing the multiple-minima energy landscape of the disordered manifold problem.

Within the class of models describe by Eq. (5), there are still a few unresolved issues. For example, the roughness exponent ζ of the $(1+1)$-dimensional LIM has been found numerically to be indistinguishable to $5/4$, suggesting the possibility of an exact derivation. The effect of a thermal noise term added to the deterministic model has not been well-understood [13,59,86]. There now exist very good numerical estimates of the exponent ψ that describes the vanishing of the interface velocity $v \sim T^{\psi}$ with temperature T at the depinning threshold, but its value does not agree with any of the existing theoretical proposals [13]. Another puzzle is the closeness between the numerical estimates for the dynamical exponent z given in Table 1 and the one-loop RG prediction for the CDW depinning.

The mapping between the CDW depinning and the sandpile models offers a very powerful tool for the development of RG theories for systems that exhibit SOC. To our knowledge, this connection has not been fully explored so far. It would be interesting to see if SOC models other than the sandpile and the rice-pile, such as the Olami–Feder–Christensen model [70] for earthquakes, can also be mapped to some type of depinning model. Conversely, it would be nice to find out whether the lessons learned from the detailed characterization of avalanche dynamics [19] can contribute to a better understanding of critical correlations in the LIM at the depinning threshold from small to large scales.

On the experimental side, perhaps the best studied system is the CDW depinning, yet the CDW velocity generally grows faster than linear above the depinning threshold [85], while the elastic depinning theory predicts $\theta < 1$. It has been suggested that the apparent discrepancy could be due to a combination of factors: Nonuniform driving, thermal activation across energy barriers, and plasticity when dislocations in the phase field is present. Simultaneous presence of strong pinning sites and weak collective pinning in a finite system may also give rise to a complex I–V curve in experimental systems. While some of these effects have been investigated theoretically [79] (see also the review by Brazovskii and Nattermann), a clear picture is yet to emerge.

Moulinet et al. [62] designed an experimental system to study the roughness and dynamics of the contact line of a viscous fluid. The roughness exponent ζ obtained from experimental measurements is 0.5, much bigger than the value 0.385 obtained from numerical simulations [21] of the model considered by Ertas and Kardar [24]. Similar discrepancies have been reported in the experimental studies [9,74] of the roughness of crack fronts. Modifications of the LIM to include nonlinear couplings and wave-like dynamics [9] for avalanche propagation have been suggested. However, many unresolved issues remain.

It is perhaps not surprising to see that real physical systems almost always contain complications that invalidate direct comparison with the linear elasticity theory of depinning. The pinned state just below the transition is highly metastable and hence is susceptible to various relaxational processes that could possibly invalidate our simple assumptions (see, however [42] for a discussion of correlation lengths below the elastic depinning threshold). Inhomogeneities inside the sample, e. g., macroscopic variations in the concentration of impurity atoms, may also have a dramatic effect on the critical properties associated with the depinning transition. Incorporating the relevant microscopic processes into the generic description discussed here is expected to yield a more complete theory of depinning. It is hoped that future work in this direction will unleash the full impact of the theoretical progress that have been made in the past two decades, which in itself has been intellectually extremely stimulating.

Bibliography

Primary Literature

1. Aharony A, Imry Y, Ma SK (1976) Lowering of dimensionality in phase transitions with random fields. Phys Rev Lett 37:1364–1367
3. Amaral LAN, Barabasi AL, Stanley HE (1994) Universality classes for interface growth with quenched disorder. Phys Rev Lett 73:62–65
2. Amaral LAN, Barabasi AL, Buldyrev SV, Harrington ST, Havlin S, Lahijany RS, Stanley HE (1995) Avalanches and the directed percolation depinning model: Experiments, simulations, and theory. Phys Rev E 51:4655–4673
4. Anderson PW (1984) Basic notions of condensed matter physics. Benjamin/Cummings, Menlo Park
5. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation for $1/f$ noise. Phys Rev Lett 59:381–384
6. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38:364–374
7. Bolech CJ, Rosso A (2004) Universal statistics of the critical depinning force of elastic systems in random media. Phys Rev Lett 93:125701
8. Bonachela JA, Chaté H, Dornic I, Munoz MA (2007) Absorbing states and elastic interfaces in random media: Two equivalent descriptions of self-organized criticality. Phys Rev Lett 98:155702
9. Bouchaud E, Bouchaud JP, Fisher DS, Ramanathan S, Rice JR (2002) Can crack front waves explain the roughness of cracks? J Mech Phys Solids 50:1703–1725
10. Bruinsma R, Aeppli G (1984) Interface motion and nonequilibrium properties of the random-field ising model. Phys Rev Lett 52:1547–1550
11. Buldyrev SV, Barabasi AL, Caserta F, Havlin S, Stanley HE, Vicsek T (1992) Anomalous interface roughening in porous media: Experiment and model. Phys Rev A 45:R8313–R8316
12. Buldyrev SV, Barabasi AL, Havlin S, Kertesz J, Stanley HE, Xenias HS (1992) Anomalous interface roughening in 3D porous media: Experiment and model. Physica A 191:220–226
13. Bustingorry S, Kolton AB, Giamarchi T (2008) Thermal rounding of the depinning transition. Europhys Lett 81:26005
14. Chauve P, Le Doussal P, Wiese KJ (2001) Renormalization of pinned elastic systems: How does it work beyond one loop? Phys Rev Lett 86:1785–1788
15. Christensen K, Corral A, Frette V, Feder J, Jossang T (1996) Tracer dispersion in a self-organized critical system. Phys Rev Lett 77:107–110
16. Coppersmith SN, Millis AJ (1991) Diverging strains in the phase-deformation model of sliding charge-density waves. Phys Rev B 44:7799–7807
17. Csahók Z, Honda K, Vicsek T (1993) Dynamics of surface roughening in disordered media. J Phys A 26:L171–L178
18. de Gennes PG (1985) Wetting: Statics and dynamics. Rev Mod Phys 57:827–863
19. Dhar D (2006) Theoretical studies of self-organized criticality. Physica A 369:29–70
20. Dhar D, Barma M, Phani MK (1981) Duality transformations for two-dimensional directed percolation and resistance problems. Phys Rev Lett 47:1238–1241
21. Duemmer O, Krauth W (2007) Depinning exponents of the driven long-range elastic string. J Stat Mech 2007:P01019
22. Efetov KB, Larkin AI (1977) Charge-density wave in a random potential. Sov Phys JETP 45:1236–1241
23. Ertas D, Kardar M (1994) Anisotropic scaling in depinning of a flux line. Phys Rev Lett 73:1703–1706
24. Ertas D, Kardar M (1994) Critical dynamics of contact line depinning. Phys Rev E 49:R2532–R2535
25. Fedorenko AA, Doussal PL, Wiese KJ (2006) Universal distribution of threshold forces at the depinning transition. Phys Rev B 74:041110
26. Feigel'man MV (1983) Propagation of a plane front in an inhomogeneous medium. Sov Phys JETP 58:1076–1077
27. Fisher DS (1985) Sliding charge-density waves as a dynamic critical phenomenon. Phys Rev B 31:1396–1427
28. Fisher DS (1986) Interface fluctuations in disordered systems: $5-\varepsilon$ expansion and failure of dimensional reduction. Phys Rev Lett 56:1964–1967
29. Fukuyama H, Lee PA (1978) Dynamics of the charge-density wave. I. Impurity pinning in a single chain. Phys Rev B 17:535–541
30. Giamarchi T, Le Doussal P (1994) Elastic theory of pinned flux lattices. Phys Rev Lett 72:1530–1533
31. Granato A, Lüke K (1956) Theory of mechanical damping due to dislocations. J Appl Phys 27:583–593
32. Grinstein G (1976) Ferromagnetic phase transitions in random fields: The breakdown of scaling laws. Phys Rev Lett 37:944–947
33. Grüner G (1988) The dynamics of charge-density waves. Rev Mod Phys 60:1129–1181
34. Halpin-Healy T, Zhang YC (1995) Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Aspects of multidisciplinary statistical mechanics. Phys Rep 254:215–414

35. Harris AB (1974) Effect of random defects on the critical behaviour of Ising models. J Phys C 7:1671–1692
36. Havlin S, Amaral LAN, Buldyrev SV, Harrington ST, Stanley HE (1995) Dynamics of surface roughening with quenched disorder. Phys Rev Lett 74:4205–4208
37. Imry Y, Ma SK (1975) Random-field instability of the ordered state of continuous symmetry. Phys Rev Lett 35:1399–1401
38. Jeong H, Kahng B, Kim D (1996) Anisotropic surface growth model in disordered media. Phys Rev Lett 77:5094–5097
39. Joanny JF, de Gennes PG (1984) A model for contact angle hysteresis. J Chem Phys 81:552–562
40. Kardar M, Parisi G, Zhang YC (1986) Dynamic scaling of growing interfaces. Phys Rev Lett 56:889–892
41. Kinzel W (1982) Directed Percolation. In: Deutscher G, Zallen R, Adler J (eds) Percolation structures and processes. Annals of the Israel Physical Society, vol 5. Hilger, Bristol, p 425
42. Kolton AB, Rosso A, Giamarchi T, Krauth W (2006) Dynamics below the depinning threshold in disordered elastic systems. Phys Rev Lett 97:057001
43. Koplik J, Levine H (1985) Interface moving through a random background. Phys Rev B 32:280–292
44. Larkin AI (1970) Effect of inhomogeneities on the structure of the mixed state of superconductors. Sov Phys JETP 31:784–786
45. Larkin AI, Ovchinnikov YN (1979) Pinning in type-II superconductors. J Low Temp Phys 34:409–428
47. Le Doussal P, Wiese KJ (2003) Functional renormalization group for anisotropic depinning and relation to branching processes. Phys Rev E 67:016121
48. Le Doussal P, Wiese KJ, Chauve P (2002) Two-loop functional renormalization group theory of the depinning transition. Phys Rev B 66:174201
49. Le Doussal P, Wiese KJ, Chauve P (2004) Functional renormalization group and the field theory of disordered elastic systems. Phys Rev E 69:026112
46. Le Doussal P, Muller M, Wiese KJ (2008) Cusps and shocks in the renormalized potential of glassy random manifolds: How functional renormalization group and replica symmetry breaking fit together. Phys Rev B 77:064203
50. Lee PA, Rice TM (1979) Electric field depinning of charge density waves. Phys Rev B 19:3970–3980
51. Leger L, Joanny JF (1992) Liquid spreading. Rep Prog Phys 55:431–486
52. Leschhorn H (1992) Interface motion in a random medium: Mean field theory. J Phys A 25:L555–L560
53. Leschhorn H (1993) Interface depinning in a disordered medium? Numerical results. Physica A 195:324–335
54. Leschhorn H (1996) Anisotropic interface depinning: Numerical results. Phys Rev E 54:1313–1320
56. Leschhorn H, Tang LH (1993) Comment on "Elastic string in a random potential". Phys Rev Lett 70:2973
55. Leschhorn H, Nattermann T, Stepanow S, Tang LH (1997) Driven interface depinning in a disordered medium. Ann Phys (Leipzig) 6:1–34
57. Manna S (1991) Two-state model of self-organized criticality. J Phys A 24:L363–L369
58. Middleton AA (1992) Asymptotic uniqueness of the sliding state for charge-density waves. Phys Rev Lett 68:670–673
59. Middleton AA (1992) Thermal rounding of the charge-density-wave depinning transition. Phys Rev B 45:9465–9468
60. Middleton AA, Fisher DS (1991) Critical behavior of pinned charge-density waves below the threshold for sliding. Phys Rev Lett 66:92–95
61. Middleton AA, Fisher DS (1993) Critical behavior of charge-density waves below threshold: Numerical and scaling analysis. Phys Rev B 47:3530–3552
62. Moulinet S, Guthmann C, Rolley E (2002) Roughness and dynamics of a contact line of a viscous fluid on a disordered substrate. Eur Phys J E 8:437–443
63. Myers CR, Sethna JP (1993) Collective dynamics in a model of sliding charge-density waves. I. Critical behavior. Phys Rev B 47:11171–11192
64. Nagel SR (1992) Instabilities in a sandpile. Rev Mod Phys 64:321–325
65. Narayan O, Fisher DS (1992) Critical behavior of sliding charge-density waves in $4-\varepsilon$ dimensions. Phys Rev B 46:11520–11549
66. Narayan O, Fisher DS (1993) Threshold critical dynamics of driven interfaces in random media. Phys Rev B 48:7030–7042
67. Narayan O, Middleton AA (1994) Avalanches and the renormalization group for pinned charge-density waves. Phys Rev B 49:244–256
68. Nattermann T (1990) Scaling approach to pinning: Charge density waves and giant flux creep in superconductors. Phys Rev Lett 64:2454–2457
69. Nattermann T, Stepanow S, Tang LH, Leschhorn H (1992) Dynamics of interface depinning in a disordered medium. J Phys II (Paris) 2:1483–1488
70. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68:1244–1247
71. Paczuski M, Boettcher S (1996) Universality in sandpiles, interface depinning, and earthquake models. Phys Rev Lett 77:111–114
72. Paczuski M, Maslov S, Bak P (1996) Avalanche dynamics in evolution, growth, and depinning models. Phys Rev E 53:414–443
73. Podgorski T, Flesselles JM, Limat L (2001) Corners, cusps, and pearls in running drops. Phys Rev Lett 87:036102
74. Ponson L, Bonamy D, Bouchaud E (2006) Two-dimensional scaling properties of experimental fracture surfaces. Phys Rev Lett 96:035506
75. Ramanathan S, Fisher DS (1998) Onset of propagation of planar cracks in heterogeneous media. Phys Rev B 58:6026–6046
76. Rosso A, Hartmann AK, Krauth W (2003) Depinning of elastic manifolds. Phys Rev E 67:021602
77. Rosso A, Le Doussal P, Wiese KJ (2007) Numerical calculation of the funcitonal renormalization group fixed-point functions at the depinning transition. Phys Rev B 75:220201(R)
78. Ryu KS, Akinaga H, Shin SC (2007) Tunable scaling behaviour observed in Barkhausen criticality of a ferromagnetic film. Nature Physics 3:547–550
79. Saunders K, Schwarz JM, Marchetti MC, Middleton AA (2004) Mean-field theory of collective transport with phase slips. Phys Rev B 70:024205
80. Sethna JP, Dahmen K, Kartha S, Krumhansl JA, Roberts BW, Shore JD (1993) Hysteresis and hierarchies: Dynamics of disorder-driven first-order phase transformations. Phys Rev Lett 70:3347–3350
81. Stepanow S (1995) Dynamics of growing interfaces in disordered medium: The effect of lateral growth. J Phys II (France) 5:11–18

82. Tang C, Bak P (1988) Critical exponents and scaling relations for self-organized criticality phenomena. Phys Rev Lett 60:2347–2350
84. Tang LH, Leschhorn H (1992) Pinning by directed percolation. Phys Rev A 45:R8309–R8312
83. Tang LH, Kardar M, Dhar D (1995) Driven depinning in anisotropic media. Phys Rev Lett 74:920–923
85. Thorne RE (2005) A history of the I–V characteristic of CDW conductors. J Phys IV (France) 131:89–94
86. Vandembroucq D, Skoe R, Roux S (2004) Universal depinning force fluctuations of an elastic line: Application to finite temperature behavior. Phys Rev E 70:051101
87. Zaitsev SI (1992) Robin Hood as self-organized criticality. Physica A 189:411–416
88. Zapperi S, Cizeau P, Durin G, Stanley HE (1998) Dynamics of a ferromagnetic domain wall: Avalanches, depinning transition, and the Barkhausen effect. Phys Rev B 58:6353–6366

Books and Reviews

Alava M, Dubé M, Rost M (2004) Imbibition in disordered media. Adv Phys 53:83–175

Blatter G, Feigelman MV, Geshkenbein VB, Larkin AI, Vinokur VM (1994) Vortices in high-temperature superconductors. Rev Mod Phys 66:1125–1388

Brazovskii S, Nattermann T (2004) Pinning and sliding of driven elastic systems: From domain walls to charge density waves. Adv Phys 53:177–252

de Gennes PG, Brochard-Wyart F, Quéré D (2003) Capillarity and Wetting Phenomena: Drops, Bubbles, Pearls, Waves. Springer, New York

Fisher DS (1998) Collective transport in random media: From superconductors to earthquakes. Phys Rep 301:113–150

Kardar M (1998) Nonequilibrium dynamics of interfaces and lines. Phys Rep 301:85–112

Quéré D (2005) Non-sticking drops. Rep Prog Phys 68:2495–2532

Sethna JP, Dahmen KA, Myers CR (2001) Crackling noise. Nature 410:242–250

Turcotte DL (1999) Self-organized criticality. Rep Prog Phys 62:1377–1429

Community Structure in Graphs

Santo Fortunato[1], Claudio Castellano[2]
[1] Complex Networks Lagrange Laboratory (CNLL),
ISI Foundation, Torino, Italy
[2] SMC, INFM-CNR and Dipartimento di Fisica,
“Sapienza” Università di Roma, Roma, Italy

Article Outline

Glossary

Graph A graph is a set of elements, called *vertices* or *nodes*, where pairs of vertices are connected by relational links, or *edges*. A graph can be considered as the simplest representation of a complex system, where the vertices are the elementary units of the system and the edges represent their mutual interactions.

Community A community is a group of graph vertices that “belong together” according to some precisely defined criteria which can be measured. Many definitions have been proposed. A common approach is to define a community as a group of vertices such that the density of edges between vertices of the group is higher than the average edge density in the graph. In the text also the terms *module* or *cluster* are used when referring to a community.

Partition A partition is a split of a graph in subsets with each vertex assigned to only one of them. This last condition may be relaxed to include the case of overlapping communities, imposing that each vertex is assigned to at least one subset.

Dendrogram A dendrogram, or hierarchical tree, is a branching diagram representing successive divisions of a graph into communities. Dendrograms are frequently used in social network analysis and computational biology, especially in biological taxonomy.

Scalability Scalability expresses the computational complexity of an algorithm. If the running time of a community detection algorithm, working on a graph with n vertices and m edges, is proportional to the product $n^{\alpha} m^{\beta}$, one says that the algorithm scales as $O(n^{\alpha} m^{\beta})$. Knowing the scalability allows to estimate the range of applicability of an algorithm.

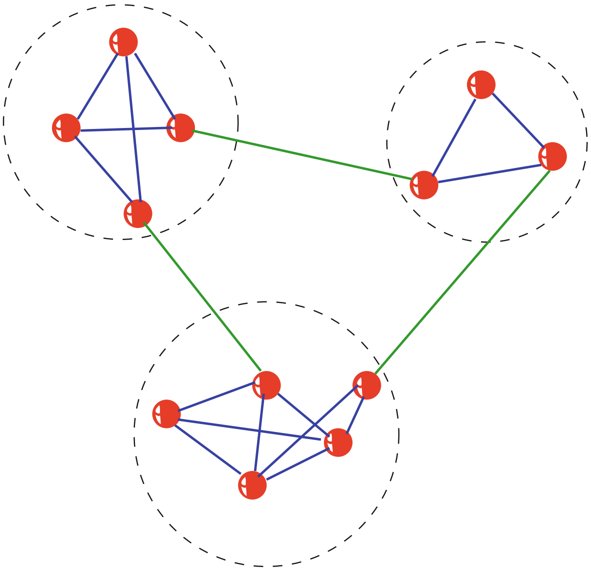
Definition of the Subject

Graph vertices are often organized into groups that seem to live fairly independently of the rest of the graph, with which they share but a few edges, whereas the relationships between group members are stronger, as shown by the large number of mutual connections. Such groups of vertices, or communities, can be considered as independent compartments of a graph. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. The task is very hard, though, both conceptually, due to the ambiguity in the definition of community and in the discrimination of different partitions and practically, because algorithms must find "good" partitions among an exponentially large number of them. Other complications are represented by the possible occurrence of hierarchies, i. e. communities which are nested inside larger communities, and by the existence of overlaps between communities, due to the presence of nodes belonging to more groups. All these aspects are dealt with in some detail and many methods are described, from traditional approaches used in computer science and sociology to recent techniques developed mostly within statistical physics.

Introduction

The origin of graph theory dates back to Euler's solution [1] of the puzzle of Königsberg's bridges in 1736. Since then a lot has been learned about graphs and their mathematical properties [2]. In the 20th century they have also become extremely useful as representation of a wide variety of systems in different areas. Biological, social, technological, and information networks can be studied as graphs, and graph analysis has become crucial to understand the features of these systems. For instance, social network analysis started in the 1930s and has become one of the most important topics in sociology [3,4]. In recent times, the computer revolution has provided scholars with a huge amount of data and computational resources to process and analyze these data. The size of real networks one can potentially handle has also grown considerably, reaching millions or even billions of vertices. The need to deal with such a large number of units has produced a deep change in the way graphs are approached [5,6,7,8,9].

Real networks are not random graphs. The random graph, introduced by P. Erdös and A. Rényi [10], is the paradigm of a disordered graph: in it, the probability of having an edge between a pair of vertices is equal for all possible pairs. In a random graph, the distribution of edges among the vertices is highly homogeneous. For instance, the distribution of the number of neighbors of a vertex, or *degree*, is binomial, so most vertices have equal or similar degree. In many real networks, instead, there are big inhomogeneities, revealing a high level of order and organization. The degree distribution is broad, with a tail that often follows a power law: therefore, many vertices with low degree coexist with some vertices with large degree. Furthermore, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of nodes, and low concentrations between these groups. This feature of real networks is called *community structure* and is the topic of this chapter. In Fig. 1 a schematic example of a graph with community structure is shown.



Community Structure in Graphs, Figure 1
A simple graph with three communities, highlighted by the *dashed circles*

Communities are groups of vertices which probably share common properties and/or play similar roles within the graph. So, communities may correspond to groups of pages of the World Wide Web dealing with related topics [11], to functional modules such as cycles and pathways in metabolic networks [12,13], to groups of related individuals in social networks [14,15], to compartments in food webs [16,17], and so on.

Community detection is important for other reasons, too. Identifying modules and their boundaries allows for a classification of vertices, according to their topological position in the modules. So, vertices with a central position in their clusters, i. e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group; vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities. Such classification seems to be meaningful in social [18,19,20] and metabolic networks [12]. Finally, one can study the graph where vertices are the communities and edges are set between modules if there are connections between some of their vertices in the original graph and/or if the modules overlap. In this way one attains a coarse-grained description of the original graph, which unveils the relationships between modules. Recent studies indicate that networks of communities have a different degree distribution with respect to the full graphs [13]; however, the origin of their structures can be explained by the same mechanism [21].

The aim of community detection in graphs is to identify the modules only based on the topology. The problem has a long tradition and it has appeared in various forms in several disciplines. For instance, in parallel computing it is crucial to know what is the best way to allocate tasks to processors so as to minimize the communications between them and enable a rapid performance of the calculation. This can be accomplished by splitting the computer cluster into groups with roughly the same number of processors, such that the number of physical connections between processors of different groups is minimal. The mathematical formalization of this problem is called *graph partitioning*. The first algorithms for graph partitioning were proposed in the early 1970s. Clustering analysis is also an important aspect in the study of social networks. The most popular techniques are hierarchical clustering and k-means clustering, where vertices are joined into groups according to their mutual similarity.

In a seminal paper, Girvan and Newman proposed a new algorithm, aiming at the identification of edges lying between communities and their successive removal, a procedure that after a few iterations leads to the isolation of modules [14]. The intercommunity edges are detected according to the values of a centrality measure, the edge betweenness, that expresses the importance of the role of the edges in processes where signals are transmitted across the graph following paths of minimal length. The paper triggered a big activity in the field, and many new methods have been proposed in the last years. In particular, physicists entered the game, bringing in their tools and techniques: spin models, optimization, percolation, random walks, synchronization, etc., became ingredients of new original algorithms. Earlier reviews of the topic can be found in [22,23].

Section "Elements of Community Detection" is about the basic elements of community detection, starting from the definition of community. The classical problem of graph partitioning and the methods for clustering analysis in sociology are presented in Sect. "Computer Science: Graph Partitioning" and "Social Science: Hierarchical and k-Means Clustering", respectively. Section "New Methods" is devoted to a description of the new methods. In Sect. "Testing Methods" the problem of testing algorithms is discussed. Section "The Mesoscopic Description of a Graph" introduces the description of graphs at the level of communities. Finally, Sect. "Future Directions" highlights the perspectives of the field and sorts out promising research directions for the future.

This chapter makes use of some basic concepts of graph theory, that can be found in any introductory textbook, like [2]. Some of them are briefly explained in the text.

Elements of Community Detection

The problem of community detection is, at first sight, intuitively clear. However, when one needs to formalize it in detail things are not so well defined. In the intuitive concept some ambiguities are hidden and there are often many equally legitimate ways of resolving them. Hence the term "Community Detection" actually indicates several rather different problems.

First of all, there is no unique way of translating into a precise prescription the intuitive idea of community. Many possibilities exist, as discussed below. Some of these possible definitions allow for vertices to belong to more than one community. It is then possible to look for overlapping or nonoverlapping communities. Another ambiguity has to do with the concept of community structure. It may be intended as a single partition of the graph or as a hierarchy of partitions, at different levels of coarse-graining. There is then a problem of comparison. Which one is the best partition (or the best hierarchy)? If one could, in principle, analyze all possible partitions of a graph, one would need a sensible way of measuring their "quality" to single out the partitions with the strongest community structure. It may even occur that one graph has no community structure and one should be able to realize it. Finding a good method for comparing partitions is not a trivial task and different choices are possible. Last but not least, the number of possible partitions grows faster than exponentially with the graph size, so that, in practice, it is not possible to analyze them all. Therefore one has to devise smart methods to find 'good' partitions in a reasonable time. Again, a very hard problem.

Before introducing the basic concepts and discussing the relevant questions it is important to stress that the identification of topological clusters is possible only if the graphs are *sparse*, i. e. if the number of edges m is of the order of the number of nodes n of the graph. If $m \gg n$, the distribution of edges among the nodes is too homogeneous for communities to make sense.

Definition of Community

The first and foremost problem is how to define precisely what a community is. The intuitive notion presented in the Introduction is related to the comparison of the number of edges joining vertices within a module ("intracommunity edges") with the number of edges joining vertices of different modules ("intercommunity edges"). A module is characterized by a larger density of links "inside" than "outside". This notion can be however formalized in many ways. Social network analysts have devised many definitions of subgroups with various degrees of internal cohesion among vertices [3,4]. Many other definitions have been introduced by computer scientists and physicists. In general, the definitions can be classified in three main categories.

- *Local definitions*. Here the attention is focused on the vertices of the subgraph under investigation and on its immediate neighborhood, disregarding the rest of the graph. These prescriptions come mostly from social network analysis and can be further subdivided in *self-referring*, when one considers the subgraph alone, and *comparative*, when the mutual cohesion of the vertices of the subgraph is compared with their cohesion with the external neighbors. Self-referring definitions identify classes of subgraphs like *cliques*, *n-cliques*, *k-plexes*, etc. They are *maximal subgraphs*, which cannot be enlarged with the addition of new vertices and edges without losing the property which defines them. The concept of clique is very important and often recurring when one studies graphs. A clique is a maximal subgraph where each vertex is adjacent to all the others. In the literature it is common to call cliques also non-maximal subgraphs. Triangles are the simplest cliques, and are frequent in real networks. Larger cliques are rare, so they are not good models of communities. Besides, finding cliques is computationally very demanding: the Bron–Kerbosch method [24] runs in a time growing exponentially with the size of the graph. The definition of clique is very strict. A softer constraint is represented by the concept of n-clique, which is a maximal subgraph such that the distance of each pair of its vertices is not larger than n. A k-plex is a maximal subgraph such that each vertex is adjacent to all the others except at most k of them. In contrast, a *k-core* is a maximal subgraph where each vertex is adjacent to at least k vertices within the subgraph. Comparative definitions include that of *LS set*, or *strong community*, and that of *weak community*. An LS set is a subgraph where each node has more neighbors inside than outside the subgraph. Instead, in a weak community, the total degree of the nodes inside the community exceeds the external total degree, i. e. the number of links lying between the community and the rest of the graph. LS sets are also weak communities, but the inverse is not true, in general. The notion of weak community was introduced by Radicchi et al. [25].
- *Global definitions*. Communities are structural units of the graph, so it is reasonable to think that their distinctive features can be recognized if one analyzes a subgraph with respect to the graph as a whole. Global definitions usually start from a *null model*, i. e. a graph which matches the original in some of its topological features, but which does not display community structure. After that, the linking properties of subgraphs of the initial graph are compared with those of the corresponding subgraphs in the null model. The simplest way to design a null model is to introduce randomness in the distribution of edges among the vertices. A random graph à la Erdös–Rényi, for instance, has no community structure, as any two vertices have the same probability to be adjacent, so there is no preferential linking involving special groups of vertices. The most popular null model is that proposed by Newman and Girvan and consists of a randomized version of the original graph, where edges are rewired at random, under the constraint that each vertex keeps its degree [26]. This null model is the basic concept behind the definition of *modularity*, a function which evaluates the goodness of partitions of a graph into modules (see Sect. "Evaluating Partitions: Quality Functions"). Here a subset of vertices is a community if the number of edges inside the subset exceeds the expected number of internal edges that the subset would have in the null model. A more general definition, where one counts small connected subgraphs (*motifs*), and not necessarily edges, can be found in [27]. A general class of null models, including that of modularity, has been designed by Reichardt and Bornholdt [28].
- *Definitions based on vertex similarity*. In this last category, communities are groups of vertices which are similar to each other. A quantitative criterion is chosen to evaluate the similarity between each pair of vertices, connected or not. The criterion may be local or global: for instance one can estimate the distance between a pair of vertices. Similarities can be also extracted from eigenvector components of special matrices, which are usually close in value for vertices belonging to the same community. Similarity measures are at the basis of the method of hierarchical clustering, to be discussed in Sect. "Social Science: Hierarchical and k-Means Clustering". The main problem in this case is the need to introduce an additional criterion to select meaningful partitions.

It is worth remarking that, in spite of the wide variety of definitions, in many detection algorithms communities are not defined at all, but are a byproduct of the procedure. This is the case of the divisive algorithms described in Sect. "Divisive Algorithms" and of the dynamic algorithms of Sect. "Dynamic Algorithms".

Evaluating Partitions: Quality Functions

Strictly speaking, a partition of a graph in communities is a split of the graph in clusters, with each vertex assigned to only one cluster. The latter condition may be relaxed, as shown in Sect. "Overlapping Communities". Whatever the definition of community is, there is usually a large number of possible partitions. It is then necessary to establish which partitions exhibit a real community structure. For that, one needs a *quality function*, i. e. a quantitative criterion to evaluate how good a partition is. The most popular quality function is the modularity of Newman and Girvan [26]. It can be written in several ways, as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) , \tag{1}$$

where the sum runs over all pairs of vertices, A is the adjacency matrix, k_i the degree of vertex i and m the total number of edges of the graph. The element A_{ij} of the adjacency matrix is 1 if vertices i and j are connected, otherwise it is 0. The δ-function yields one if vertices i and j are in the same community, zero otherwise. Because of that, the only contributions to the sum come from vertex pairs belonging to the same cluster: by grouping them together the sum over the vertex pairs can be rewritten as a sum over the modules

$$Q = \sum_{s=1}^{n_m} \left[\frac{l_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right] . \tag{2}$$

Here, n_m is the number of modules, l_s the total number of edges joining vertices of module s and d_s the sum of the degrees of the vertices of s. In Eq. (2), the first term of each summand is the fraction of edges of the graph inside the module, whereas the second term represents the expected fraction of edges that would be there if the graph were a random graph with the same degree for each vertex. In such a case, a vertex could be attached to any other vertex of the graph, and the probability of a connection between two vertices is proportional to the product of their degrees. So, for a vertex pair, the comparison between real and expected edges is expressed by the corresponding summand of Eq. (1).

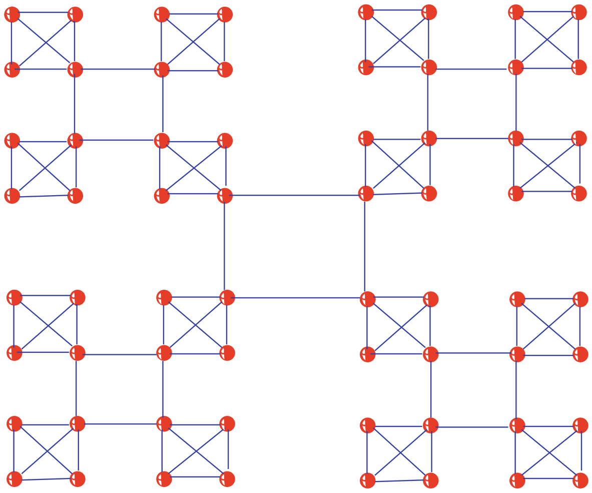
Equation (2) embeds an implicit definition of community: a subgraph is a module if the number of edges inside it is larger than the expected number in modularity's null model. If this is the case, the vertices of the subgraph are more tightly connected than expected. Basically, if each summand in Eq. (2) is non-negative, the corresponding subgraph is a module. Besides, the larger the difference between real and expected edges, the more "modular" the subgraph. So, large positive values of Q are expected to indicate good partitions. The modularity of the whole graph, taken as a single community, is zero, as the two terms of the only summand in this case are equal and opposite. Modularity is always smaller than one, and can be negative as well. For instance, the partition in which each vertex is a community is always negative. This is a nice feature of the measure, implying that, if there are no partitions with positive modularity, the graph has no community structure. On the contrary, the existence of partitions with large negative modularity values may hint to the existence of subgroups with very few internal edges and many edges lying between them (*multipartite structure*).

Modularity has been employed as a quality function in many algorithms, like some of the divisive algorithms of Sect. "Divisive Algorithms". In addition, modularity optimization is itself a popular method for community detection (see Sect. "Modularity Optimization"). Modularity also allows to assess the stability of partitions [29] and to transform a graph into a smaller one by preserving its community structure [30].

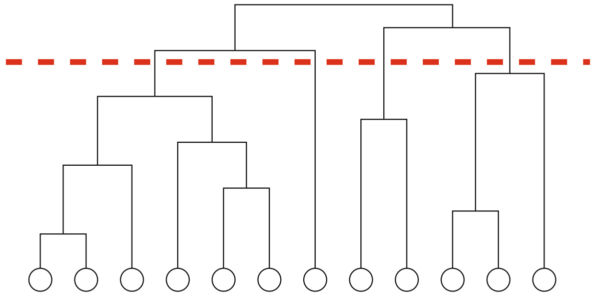
However, there are some caveats on the use of the measure. The most important concerns the value of modularity for a partition. For which values one can say that there is a clear community structure in a graph? The question is tricky: if two graphs have the same type of modular structure, but different sizes, modularity will be larger for the larger graph. So, modularity values cannot be compared for different graphs. Moreover, one would expect that partitions of random graphs will have modularity values close to zero, as no community structure is expected there. Instead, it has been shown that partitions of random graphs may attain fairly large modularity values, as the probability that the distribution of edges on the vertices is locally inhomogeneous in specific realizations is not negligible [31]. Finally, a recent analysis has proved that modularity increases if subgraphs smaller than a characteristic size are merged [32]. This fact represents a serious bias when one looks for communities via modularity optimization and is discussed in more detail in Sect. "Modularity Optimization".

Hierarchies

Graph vertices can have various levels of organization. Modules can display an internal community structure, i. e.



Community Structure in Graphs, Figure 2
Schematic example of a hierarchical graph. Sixteen modules with four vertices each are clearly organized in groups of four



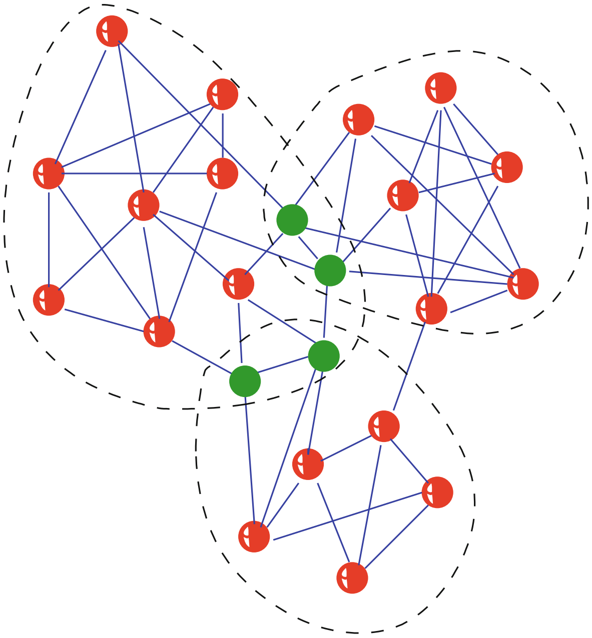
Community Structure in Graphs, Figure 3
A dendrogram, or hierarchical tree. *Horizontal cuts* correspond to partitions of the graph in communities. Reprinted figure with permission from [26]

they can contain smaller modules, which can in turn include other modules, and so on. In this case one says that the graph is hierarchical (see Fig. 2). For a clear classification of the vertices and their roles inside a graph, it is important to find all modules of the graph as well as their hierarchy.

A natural way to represent the hierarchical structure of a graph is to draw a *dendrogram*, like the one illustrated in Fig. 3. Here, partitions of a graph with twelve vertices are shown. At the bottom, each vertex is its own module. By moving upwards, groups of vertices are successively aggregated. Merges of communities are represented by horizontal lines. The uppermost level represents the whole graph as a single community. Cutting the diagram horizontally at some height, as shown in the figure (dashed line), displays one level of organization of the graph vertices. The diagram is hierarchical by construction: each community belonging to a level is fully included in a community at a higher level. Dendrograms are regularly used in sociology and biology. The technique of hierarchical clustering, described in Sect. "Social Science: Hierarchical and *k*-Means Clustering", lends itself naturally to this kind of representation.

Overlapping Communities

As stated in Sect. "Evaluating Partitions: Quality Functions", in a partition each vertex is generally attributed only to one module. However, vertices lying at the boundary between modules may be difficult to assign to one module or another, based on their connections with the other vertices. In this case, it makes sense to consider such intermediate vertices as belonging to more groups, which are then called *overlapping communities* (Fig. 4). Many real networks are characterized by a modular structure with sizeable overlaps between different clusters. In social networks, people usually belong to more communities, according to their personal life and interests: for instance a person may have tight relationships both with the people of its working environment and with other individuals involved in common free time activities.



Community Structure in Graphs, Figure 4
Overlapping communities. In the partition highlighted by the *dashed* contours, some vertices are shared between more groups

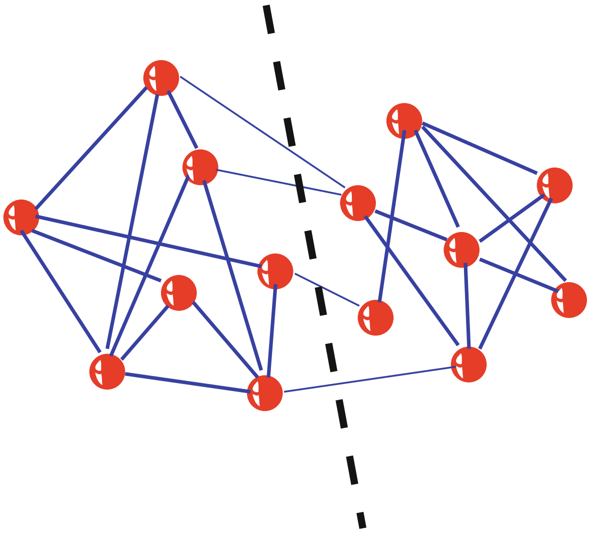
Accounting for overlaps is also a way to better exploit the information that one can derive from topology. Ideally, one could estimate the degree of participation of a vertex in different communities, which corresponds to the likelihood that the vertex belongs to the various groups. Community detection algorithms, instead, often disagree in the classification of peripheral vertices of modules, because they are forced to put them in a single cluster, which may be the wrong one.

The problem of community detection is so hard that very few algorithms consider the possibility of having overlapping communities. An interesting method has been recently proposed by G. Palla et al. [13] and is described in Sect. "Clique Percolation". For standard algorithms, the problem of identifying overlapping vertices could be addressed by checking for the stability of partitions against slight variations in the structure of the graph, as described in [33].

Computer Science: Graph Partitioning

The problem of graph partitioning consists in dividing the vertices in g groups of predefined size, such that the number of edges lying between the groups is minimal. The number of edges running between modules is called *cut size*. Figure 5 presents the solution of the problem for a graph with fourteen vertices, for $g = 2$ and clusters of equal size.

The specification of the number of modules of the partition is necessary. If one simply imposed a partition with the minimal cut size, and left the number of modules free, the solution would be trivial, corresponding to all vertices ending up in the same module, as this would yield a vanishing cut size.



Community Structure in Graphs, Figure 5
Graph partitioning. The *cut* shows the partition in two groups of equal size

Graph partitioning is a fundamental issue in parallel computing, circuit partitioning and layout, and in the design of many serial algorithms, including techniques to solve partial differential equations and sparse linear systems of equations. Most variants of the graph partitioning problem are NP-hard, i. e. it is unlikely that the solution can be computed in a time growing as a power of the graph size. There are however several algorithms that can do a good job, even if their solutions are not necessarily optimal [34]. Most algorithms perform a bisection of the graph, which is already a complex task. Partitions into more than two modules are usually attained by iterative bisectioning.

The *Kernighan–Lin algorithm* [35] is one of the earliest methods proposed and is still frequently used, often in combination with other techniques. The authors were motivated by the problem of partitioning electronic circuits onto boards: the nodes contained in different boards need to be linked to each other with the least number of connections. The procedure is an optimization of a benefit function Q, which represents the difference between the number of edges inside the modules and the number of edges lying between them. The starting point is an initial partition of the graph in two clusters of the predefined size: such initial partition can be random or suggested by some information on the graph structure. Then, subsets consisting of equal numbers of vertices are swapped between the two groups, so that Q has the maximal increase. To reduce the risk to be trapped in local maxima of Q, the procedure includes some swaps that decrease the function Q. After a series of swaps with positive and negative gains, the partition with the largest value of Q is selected and used as starting point of a new series of iterations. The Kernighan–Lin algorithm is quite fast, scaling as $O(n^2)$ in worst-case time, n being the number of vertices. The partitions found by the procedure are strongly dependent on the initial configuration and other algorithms can do better. However, the method is used to improve on the partitions found through other techniques, by using them as starting configurations for the algorithm.

Another popular technique is the *spectral bisection method*, which is based on the properties of the Laplacian matrix. The Laplacian matrix (or simply Laplacian) of a graph is obtained from the adjacency matrix A by placing on the diagonal the degrees of the vertices and by changing the signs of the other elements. The Laplacian has all non-negative eigenvalues and at least one zero eigenvalue, as the sum of the elements of each row and column of the matrix is zero. If a graph is divided into g connected components, the Laplacian would have g degenerate eigenvectors with eigenvalue zero and can be written in block-diagonal form, i. e. the vertices can be ordered in such a way that the Laplacian displays g square blocks along the diagonal, with entries different from zero, whereas all other elements vanish. Each block is the Laplacian of the corresponding subgraph, so it has the trivial eigenvector with components $(1, 1, 1, \ldots, 1, 1)$. Therefore, there are g degenerate eigenvectors with equal non-vanishing components in correspondence of the vertices of a block, whereas all other components are zero. In this way, from the components of the eigenvectors one can identify the connected components of the graph.

If the graph is connected, but consists of g subgraphs which are weakly linked to each other, the spectrum will have one zero eigenvalue and $g - 1$ eigenvalues which are close to zero. If the groups are two, the second lowest eigenvalue will be close to zero and the corresponding eigenvector, also called *Fiedler vector*, can be used to identify the two clusters as shown below.

Every partition of a graph with n vertices in two groups can be represented by an index vector $\mathbf{s}$, whose component $\mathbf{s}_i$ is $+1$ if vertex i is in one group and -1 if it is in the other group. The cut size R of the partition of the graph in the two groups can be written as

$$R = \frac{1}{4}\mathbf{s}^{\mathrm{T}}\mathbf{L}\mathbf{s} \, , \tag{3}$$

where $\mathbf{L}$ is the Laplacian matrix and $\mathbf{s}^{\mathrm{T}}$ the transpose of vector $\mathbf{s}$. Vector $\mathbf{s}$ can be written as $\mathbf{s} = \sum_i a_i \mathbf{v}_i$, where $\mathbf{v}_i$, $i = 1, \ldots, n$ are the eigenvectors of the Laplacian. If $\mathbf{s}$ is properly normalized, then

$$R = \sum_i a_i^2 \lambda_i \, , \tag{4}$$

where λ_i is the Laplacian eigenvalue corresponding to eigenvector $\mathbf{v}_i$. It is worth remarking that the sum contains at most $n - 1$ terms, as the Laplacian has at least one zero eigenvalue. Minimizing R equals to the minimization of the sum on the right-hand side of Eq. (4). This task is still very hard. However, if the second lowest eigenvector λ_2 is close enough to zero, a good approximation of the minimum can be attained by choosing $\mathbf{s}$ parallel to the Fiedler vector $\mathbf{v}_2$: this would reduce the sum to λ_2, which is a small number. But the index vector cannot be perfectly parallel to $\mathbf{v}_2$ by construction, because all its components are equal in modulus, whereas the components of $\mathbf{v}_2$ are not. The best one can do is to match the signs of the components. So, one can set $\mathbf{s}_i = +1(-1)$ if $\mathbf{v}_2^i > 0 \; (< 0)$. It may happen that the sizes of the two corresponding groups do not match the predefined sizes one wishes to have. In this case, if one aims at a split in n_1 and $n_2 = n - n_1$ vertices, the best strategy is to order the components of the Fiedler vector from the lowest to the largest values and to put in one group the vertices corresponding to the first n_1 components from the top or the bottom, and the remaining vertices in the second group. If there is a discrepancy between n_1 and the number of positive or negative components of $\mathbf{v}_2$, this procedure yields two partitions: the better solution is the one that gives the smallest cut size.

The spectral bisection method is quite fast. The first eigenvectors of the Laplacian can be computed by using the Lanczos method [36], that scales as $m/(\lambda_3 - \lambda_2)$, where m is the number of edges of the graph. If the eigenvalues λ_2 and λ_3 are well separated, the running time of the algorithm is much shorter than the time required to calculate the complete set of eigenvectors, which scales as $O(n^3)$. The method gives in general good partitions, that can be further improved by applying the Kernighan–Lin algorithm.

Other methods for graph partitioning include level-structure partitioning, the geometric algorithm, multilevel algorithms, etc. A good description of these algorithms can be found in [34].

Graph partitioning algorithms are not good for community detection, because it is necessary to provide as input both the number of groups and their size, about which in principle one knows nothing. Instead, one would like an algorithm capable to produce this information in its output. Besides, using iterative bisectioning to split the graph in more pieces is not a reliable procedure.

Social Science: Hierarchical and *k*-Means Clustering

In social network analysis, one partitions actors/vertices in clusters such that actors in the same cluster are more similar between themselves than actors of different clusters. The two most used techniques to perform clustering analysis in sociology are *hierarchical clustering* and *k-means clustering*.

The starting point of hierarchical clustering is the definition of a similarity measure between vertices. After a measure is chosen, one computes the similarity for each pair of vertices, no matter if they are connected or not. At the end of this process, one is left with a new $n \times n$ matrix X, the similarity matrix. Initially, there are n groups, each containing one of the vertices. At each step, the two most similar groups are merged; the procedure continues until all vertices are in the same group.

There are different ways to define the similarity between groups out of the matrix X. In *single linkage clustering*, the similarity between two groups is the minimum element x_{ij}, with i in one group and j in the other. On the contrary, the maximum element x_{ij} for vertices of different groups is used in the procedure of *complete linkage clustering*. In *average linkage clustering* one has to compute the average of the x_{ij}.

The procedure can be better illustrated by means of dendrograms, like the one in Fig. 3. One should note that hierarchical clustering does not deliver a single partition, but a set of partitions.

There are many possible ways to define a similarity measure for the vertices based on the topology of the network. A possibility is to define a distance between vertices, like

$$x_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2} \,. \tag{5}$$

This is a dissimilarity measure, based on the concept of structural equivalence. Two vertices are structurally equivalent if they have the same neighbors, even if they are not adjacent themselves. If i and j are structurally equivalent, $x_{ij} = 0$. Vertices with large degree and different neighbors are considered very "far" from each other. Another measure related to structural equivalence is the Pearson correlation between columns or rows of the adjacency matrix,

$$x_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i\sigma_j} \,, \tag{6}$$

where the averages $\mu_i = (\sum_j A_{ij})/n$ and the variances $\sigma_i = \sum_j (A_{ij} - \mu_i)^2/n$.

An alternative measure is the number of edge- (or vertex-) independent paths between two vertices. Independent paths do not share any edge (vertex), and their number is related to the maximum flow that can be conveyed between the two vertices under the constraint that each edge can carry only one unit of flow (max-flow/min-cut theorem). Similarly, one could consider all paths running between two vertices. In this case, there is the problem that the total number of paths is infinite, but this can be avoided if one performs a weighted sum of the number of paths, where paths of length l are weighted by the factor α^l, with $\alpha < 1$. So, the weights of long paths are exponentially suppressed and the sum converges.

Hierarchical clustering has the advantage that it does not require a preliminary knowledge on the number and size of the clusters. However, it does not provide a way to discriminate between the many partitions obtained by the procedure, and to choose that or those that better represent the community structure of the graph. Moreover, the results of the method depend on the specific similarity measure adopted. Finally, it does not correctly classify all vertices of a community, and in many cases some vertices are missed even if they have a central role in their clusters [22].

Another popular clustering technique in sociology is k-means clustering [37]. Here, the number of clusters is preassigned, say k. The vertices of the graph are embedded in a metric space, so that each vertex is a point and a distance measure is defined between pairs of points in the space. The distance is a measure of dissimilarity between vertices. The aim of the algorithm is to identify k points in this space, or *centroids*, so that each vertex is associated to one centroid and the sum of the distances of all vertices from their respective centroids is minimal. To achieve this, one starts from an initial distribution of centroids such that they are as far as possible from each other. In the first iteration, each vertex is assigned to the nearest centroid. Next, the centers of mass of the k clusters are estimated and become a new set of centroids, which allows for a new classification of the vertices, and so on. After a sufficient number of iterations, the positions of the centroids are stable, and the clusters do not change any more. The solution found is not necessarily optimal, as it strongly depends on the initial choice of the centroids. The result can be improved by performing more runs starting from different initial conditions.

The limitation of k-means clustering is the same as that of the graph partitioning algorithms: the number of clusters must be specified at the beginning, the method is not able to derive it. In addition, the embedding in a metric space can be natural for some graphs, but rather artificial for others.

New Methods

From the previous two sections it is clear that traditional approaches to derive graph partitions have serious limits. The most important problem is the need to provide the algorithms with information that one would like to derive from the algorithms themselves, like the number of clusters and their size. Even when these inputs are not necessary, like in hierarchical clustering, there is the question of estimating the goodness of the partitions, so that one can pick the best one. For these reasons, there has been a major effort in the last years to devise algorithms capable of extracting a complete information about the community structure of graphs. These methods can be grouped into different categories.

Divisive Algorithms

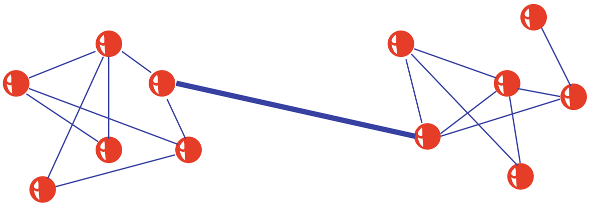
A simple way to identify communities in a graph is to detect the edges that connect vertices of different communities and remove them, so that the clusters get disconnected from each other. This is the philosophy of divisive algorithms. The crucial point is to find a property of intercommunity edges that could allow for their identification. Any divisive method delivers many partitions, which are by construction hierarchical, so that they can be represented with dendrograms.

Algorithm of Girvan and Newman. The most popular algorithm is that proposed by Girvan and Newman [14]. The method is also historically important, because it marked the beginning of a new era in the field of community detection. Here edges are selected according to the values of measures of *edge centrality*, estimating the importance of edges according to some property or process running on the graph. The steps of the algorithm are:

1. Computation of the centrality for all edges;
2. Removal of edge with largest centrality;
3. Recalculation of centralities on the running graph;
4. Iteration of the cycle from step 2.

Girvan and Newman focused on the concept of *betweenness*, which is a variable expressing the frequency of the participation of edges to a process. They considered three alternative definitions: edge betweenness, current-flow betweenness and random walk betweenness.

Edge betweenness is the number of shortest paths between all vertex pairs that run along the edge. It is an extension to edges of the concept of site betweenness, introduced by Freeman in 1977 [20]. It is intuitive that intercommunity edges have a large value of the edge betweenness, because many shortest paths connecting vertices of different communities will pass through them (Fig. 6). The betweenness of all edges of the graph can be calculated in a time that scales as $O(mn)$, with techniques based on breadth-first-search [26,38].



Community Structure in Graphs, Figure 6
Edge betweenness is highest for edges connecting communities. In the figure, the *thick* edge in the middle has a much higher betweenness than all other edges, because all shortest paths connecting vertices of the two communities run through it

Current-flow betweenness is defined by considering the graph a resistor network, with edges having unit resistance. If a voltage difference is applied between any two vertices, each edge carries some amount of current, that can be calculated by solving Kirchoff's equations. The procedure is repeated for all possible vertex pairs: the current-flow betweenness of an edge is the average value of the current carried by the edge. Calculation of current-flow betweenness requires the inversion of an $n \times n$ matrix (once), followed by obtaining and averaging the current for all pairs of nodes. Each of these two tasks takes a time $O(n^3)$ for a sparse matrix.

The random-walk betweenness of an edge says how frequently a random walker running on the graph goes across the edge. We remind that a random walker moving from a vertex follows each edge with equal probability. A pair of vertices is chosen at random, s and t. The walker starts at s and keeps moving until it hits t, where it stops. One computes the probability that each edge was crossed by the walker, and averages over all possible choices for the vertices s and t. The complete calculation requires a time $O(n^3)$ on a sparse graph. It is possible to show that this measure is equivalent to current-flow betweenness [39].

Calculating edge betweenness is much faster than current-flow or random walk betweenness ($O(n^2)$ versus $O(n^3)$ on sparse graphs). In addition, in practical applications the Girvan–Newman algorithm with edge betweenness gives better results than adopting the other centrality measures. Numerical studies show that the recalculation step 3 of Girvan–Newman algorithm is essential to detect meaningful communities. This introduces an additional factor m in the running time of the algorithm: consequently, the edge betweenness version scales as $O(m^2 n)$, or $O(n^3)$ on a sparse graph. Because of that, the algorithm is quite slow, and applicable to graphs with up to $n \sim 10\,000$ vertices, with current computational resources. In the original version of Girvan–Newman's algorithm [14], the authors had to deal with the whole hierarchy of partitions, as they had no procedure to say which partition is the best. In a successive refinement [26], they selected the partition with the largest value of modularity (see Sect. "Evaluating Partitions: Quality Functions"), a criterion that has been frequently used ever since. There have been countless applications of the Girvan–Newman method: the algorithm is now integrated in well known libraries of network analysis programs.

Algorithm of Tyler et al. Tyler, Wilkinson and Huberman proposed a modification of the Girvan–Newman algorithm, to improve the speed of the calculation [40].

The modification consists in calculating the contribution to edge betweenness only from a limited number of vertex pairs, chosen at random, deriving a sort of Monte Carlo estimate. The procedure induces statistical errors in the values of the edge betweenness. As a consequence, the partitions are in general different for different choices of the sampling pairs of vertices. However, the authors showed that, by repeating the calculation many times, the method gives good results, with a substantial gain of computer time. In practical examples, only vertices lying at the boundary between communities may not be clearly classified, and be assigned sometimes to a group, sometimes to another. The method has been applied to a network of people corresponding through email [40] and to networks of gene co-occurrences [41].

Algorithm of Fortunato et al. An alternative measure of centrality for edges is information centrality. It is based on the concept of efficiency [42], which estimates how easily information travels on a graph according to the length of shortest paths between vertices. The information centrality of an edge is the variation of the efficiency of the graph if the edge is removed. In the algorithm by Fortunato, Latora and Marchiori [43], edges are removed according to decreasing values of information centrality. The method is analogous to that of Girvan and Newman, but slower, as it scales as $O(n^4)$ on a sparse graph. On the other hand, it gives a better classification of vertices when communities are fuzzy, i. e. with a high degree of interconnectedness.

Algorithm of Radicchi et al. Because of the high density of edges within communities, it is easy to find loops in them, i. e. closed non-intersecting paths. On the contrary, edges lying between communities will hardly be part of short loops. Based on this intuitive idea, Radicchi et al. proposed a new measure, the edge clustering coefficient, such that low values of the measure are likely to correspond to intercommunity edges [25]. The edge clustering coefficient generalizes to edges the notion of clustering coefficient introduced by Watts and Strogatz for vertices [44]. The latter is the number of triangles including a vertex divided by the number of possible triangles that can be formed. The edge clustering coefficient is the number of loops of length g including the edge divided by the number of possible cycles. Usually, loops of length $g = 3$ or 4 are considered. At each iteration, the edge with smallest clustering coefficient is removed, the measure is recalculated again, and so on. The procedure stops when all clusters obtained are LS-sets or "weak" communities (see Sect. "Definition of Community"). Since the edge clustering coefficient is a local measure, involving at most an extended neighborhood of the edge, it can be calculated very quickly. The running time of the algorithm to completion is $O(m^4/n^2)$, or $O(n^2)$ on a sparse graph, so it is much shorter than the running time of the Girvan–Newman method. On the other hand, the method may give poor results when the graph has few loops, as it happens in several non-social networks. In this case, in fact, the edge clustering coefficient is small and fairly similar for all edges, and the algorithm may fail to identify the bridges between communities.

Modularity Optimization

If Newman–Girvan modularity Q (Sect. "Evaluating Partitions: Quality Functions") is a good indicator of the quality of partitions, the partition corresponding to its maximum value on a given graph should be the best, or at least a very good one. This is the main motivation for modularity maximization, perhaps the most popular class of methods to detect communities in graphs. An exhaustive optimization of Q is impossible, due to the huge number of ways in which it is possible to partition a graph, even when the latter is small. Besides, the true maximum is out of reach, as it has been recently proved that modularity optimization is an NP-hard problem [45], so it is probably impossible to find the solution in a time growing polynomially with the size of the graph. However, there are currently several algorithms able to find fairly good approximations of the modularity maximum in a reasonable time.

Greedy techniques. The first algorithm devised to maximize modularity was a greedy method of Newman [46]. It is an agglomerative method, where groups of vertices are successively joined to form larger communities such that modularity increases after the merging. One starts from n clusters, each containing a single vertex. Edges are not initially present, they are added one by one during the procedure. However, modularity is always calculated from the full topology of the graph, since one wants to find its partitions. Adding a first edge to the set of disconnected vertices reduces the number of groups from n to $n-1$, so it delivers a new partition of the graph. The edge is chosen such that this partition gives the maximum increase of modularity with respect to the previous configuration. All other edges are added based on the same principle. If the insertion of an edge does not change the partition, i. e. the clusters are the same, modularity stays the same. The number of partitions found during the procedure is n, each with a different number of clusters, from n to 1. The largest value of modularity in this subset of partitions is the approximation of the modularity maximum given by the algorithm. The update of the modularity value at each iteration step can be performed in a time $O(n+m)$, so the algorithm runs to completion in a time $O((m+n)n)$, or $O(n^2)$ on a sparse graph, which is fast. In a later paper by Clauset et al. [47], it was shown that the calculation of modularity during the procedure can be performed much more quickly by use of max-heaps, special data structures created using a binary tree. By doing that, the algorithm scales as $O(md \log n)$, where d is the depth of the dendrogram describing the successive partitions found during the execution of the algorithm, which grows as $\log n$ for graphs with a strong hierarchical structure. For those graphs, the running time of the method is then $O(n \log^2 n)$, which allows to analyze the community structure of very large graphs, up to 10^7 vertices. The greedy algorithm is currently the only algorithm that can be used to estimate the modularity maximum on such large graphs. On the other hand, the approximation it finds is not that good, as compared with other techniques. The accuracy of the algorithm can be considerably improved if one accounts for the size of the groups to be merged [48], or if the hierarchical agglomeration is started from some good intermediate configuration, rather than from the individual vertices [49].

Simulated annealing. Simulated annealing [50] is a probabilistic procedure for global optimization used in different fields and problems. It consists in performing an exploration of the space of possible states, looking for the global optimum of a function F, say its maximum. Transitions from one state to another occur with probability 1 if F increases after the change, otherwise with a probability $\exp(\beta \Delta F)$, where ΔF is the decrease of the function and β is an index of stochastic noise, a sort of inverse temperature, which increases after each iteration. The noise reduces the risk that the system gets trapped in local optima. At some stage, the system converges to a stable state, which can be an arbitrarily good approximation of the maximum of F, depending on how many states were explored and how slowly β is varied. Simulated annealing was first employed for modularity optimization by R. Guimerá et al. [31]. Its standard implementation combines two types of "moves": local moves, where a single vertex is shifted from one cluster to another, taken at random; global moves, consisting of merges and splits of communities. In practical applications, one typically combines n^2 local moves with n global ones in one iteration. The method can potentially come very close to the true modularity maximum, but it is slow. Therefore, it can be used for small graphs, with up to about 10^4 vertices. Applications include studies of potential energy landscapes [51] and of metabolic networks [12].

Extremal optimization. Extremal optimization is a heuristic search procedure proposed by Boettcher and Percus [52], in order to achieve an accuracy comparable with simulated annealing, but with a substantial gain in computer time. It is based on the optimization of local variables, expressing the contribution of each unit of the system to the global function at study. This technique was used for modularity optimization by Duch and Arenas [53]. Modularity can be indeed written as a sum over the vertices: the local modularity of a vertex is the value of the corresponding term in this sum. A fitness measure for each vertex is obtained by dividing the local modularity of the vertex by its degree. One starts from a random partition of the graph in two groups. At each iteration, the vertex with the lowest fitness is shifted to the other cluster. The move changes the partition, so the local fitnesses need to be recalculated. The process continues until the global modularity Q cannot be improved any more by the procedure. At this stage, each cluster is considered as a graph on its own and the procedure is repeated, as long as Q increases for the partitions found. The algorithm finds an excellent approximation of the modularity maximum in a time $O(n^2 \log n)$, so it represents a good tradeoff between accuracy and speed.

Spectral optimization. Modularity can be optimized using the eigenvalues and eigenvectors of a special matrix, the modularity matrix B, whose elements are

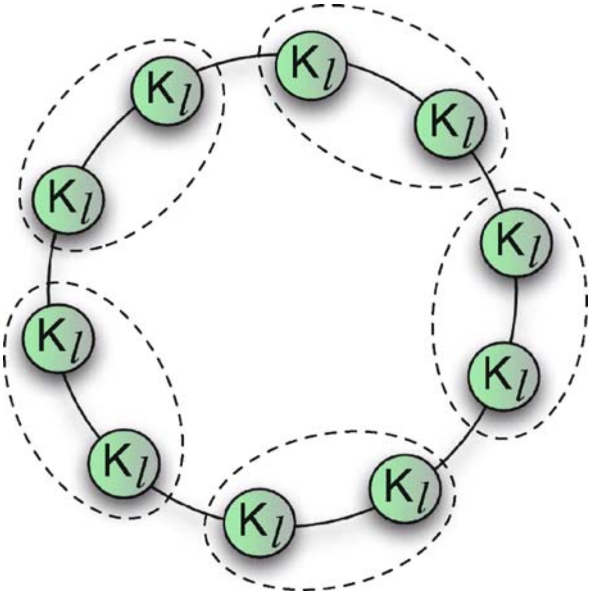
$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \,, \tag{7}$$

where the notation is the same used in Eq. (1). The method [54,55] is analogous to spectral bisection, described in Sect. "Computer Science: Graph Partitioning". The difference is that here the Laplacian matrix is replaced by the modularity matrix. Between Q and B there is the same relation as between R and L in Eq. (3), so modularity can be written as a weighted sum of the eigenvalues of B, just like Eq. (4). Here one has to look for the eigenvector of B with largest eigenvalue, $\mathbf{u}_1$, and group the vertices according to the signs of the components of $\mathbf{u}_1$, just like in Sect. "Computer Science: Graph Partitioning". The Kernighan–Lin algorithm can then be used to improve the result. The procedure is repeated for each of the clusters separately, and the number of communities increases as long as modularity does. The advantage over spectral bisection is that it is not necessary to specify the size of the two groups, because it is determined by taking the partition with largest modularity. The drawback is similar as for spectral bisection, i. e. the algorithm gives the best results for bisections, whereas it is less accurate when the number of communities is larger than two. The situation could be improved by using the other eigenvectors with positive eigenvalues of the modularity matrix. In addition, the eigenvectors with the most negative eigenvalues are important to detect a possible multipartite structure of the graph, as they give the most relevant contribution to the modularity minimum. The algorithm typically runs in a time $O(n^2 \log n)$ for a sparse graph, when one computes only the first eigenvector, so it is faster than extremal optimization, and slightly more accurate, especially for large graphs.

Finally, some general remarks on modularity optimization and its reliability. A large value for the modularity maximum does not necessarily mean that a graph has a community structure. Random graphs can also have partitions with large modularity values, even though clusters are not explicitly built in [31,56]. Therefore, the modularity maximum of a graph reveals its community structure only if it is appreciably larger than the modularity maximum of random graphs of the same size [57].

In addition, one assumes that the modularity maximum delivers the "best" partition of the network in communities. However, this is not always true [32]. In the definition of modularity (Eq. (2)) the graph is compared with a random version of it, that keeps the degrees of its vertices. If groups of vertices in the graphs are more tightly connected than they would be in the randomized graph, modularity optimization would consider them as parts of the same module. But if the groups have less than $\sqrt{m}$ internal edges, the expected number of edges running between them in modularity's null model is less than one, and a single interconnecting edge would cause the merging of the two groups in the optimal partition. This holds for every density of edges inside the groups, even in the limit case in which all vertices of each group are connected to each other, i. e. if the groups are cliques. In Fig. 7 a graph is made out of n_c identical cliques, with l vertices each, connected by single edges. It is intuitive to think that the modules of the best partition are the single cliques: instead, if n_c is larger than about l^2, modularity would be higher for the partition in which pairs of consecutive cliques are parts of the same module (indicated by the dashed lines in the figure). The problem holds for a wide class of possible null models [58]. Attempts have been made to solve it within the modularity framework [59,60,61].

Modifications of the measure have also been suggested. Massen and Doye proposed a slight variation of modularity's null model [51]: it is still a graph with the same degree sequence as the original, and with edges rewired at random among the vertices, but one imposes the additional constraint that there can be neither multiple edges between a pair of vertices nor edges joining a vertex with itself (self-edges). Muff, Rao and Caflisch remarked that modularity's null model implicitly assumes that each vertex could be attached to any other, whether in real cases a cluster is usually connected to few other clusters [62]. Therefore, they proposed a local version of modularity, in which the expected number of edges within a module is not calculated with respect to the full graph, but considering just a portion of it, namely the subgraph including the module and its neighboring modules.

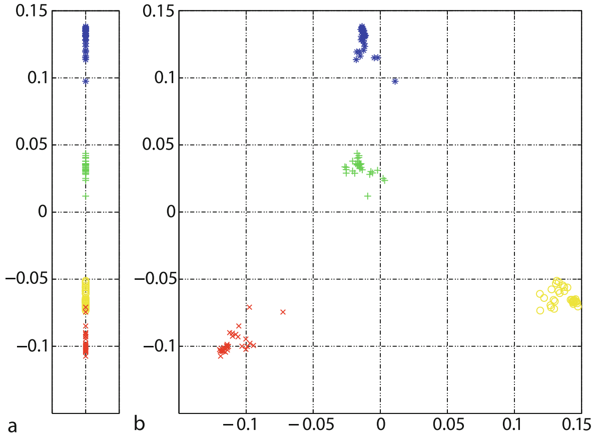


Community Structure in Graphs, Figure 7
Resolution limit of modularity optimization. The natural community structure of the graph, represented by the individual cliques (*circles*), is not recognized by optimizing modularity, if the cliques are smaller than a scale depending on the size of the graph. Reprinted figure with permission from [32]

Spectral Algorithms

As discussed above, spectral properties of graph matrices are frequently used to find partitions. Traditional methods are in general unable to predict the number and size of the clusters, which instead must be fed into the procedure. Recent algorithms, reviewed below, are more powerful.

Algorithm of Donetti and Muñoz. An elegant method based on the eigenvectors of the Laplacian matrix has been devised by Donetti and Muñoz [63]. The idea is simple: the values of the eigenvector components are close for vertices in the same community, so one can use them as coordinates to represent vertices as points in a metric space. So, if one uses M eigenvectors, one can embed the vertices in an M-dimensional space. Communities appear as groups of points well separated from each other, as illustrated in Fig. 8. The separation is the more visible, the larger the number of dimensions/eigenvectors M. The space points



Community Structure in Graphs, Figure 8
Spectral algorithm by Donetti and Muñoz. Vertex *i* is represented by the values of the *i*th components of Laplacian eigenvectors. In this example, the graph has an adhoc division in four communities, indicated by different symbols. The communities are better separated in two dimensions (b) than in one (a). Reprinted figure with permission from [63]

are grouped in communities by hierarchical clustering (see Sect. "Social Science: Hierarchical and *k*-Means Clustering"). The final partition is the one with largest modularity. For the similarity measure between vertices, Donetti and Muñoz used both the Euclidean distance and the angle distance. The angle distance between two points is the angle between the vectors going from the origin of the M-dimensional space to either point. Applications show that the best results are obtained with complete-linkage clustering. The algorithm runs to completion in a time $O(n^3)$, which is not fast. Moreover, the number M of eigenvectors that are needed to have a clean separation of the clusters is not known a priori.

Algorithm of Capocci et al. Similarly to Donetti and Muñoz, Capocci et al. used eigenvector components to identify communities [64]. In this case the eigenvectors are those of the *normal matrix*, that is derived from the adjacency matrix by dividing each row by the sum of its elements. The eigenvectors can be quickly calculated by performing a constrained optimization of a suitable cost function. A similarity matrix is built by calculating the correlation between eigenvector components: the similarity between vertices i and j is the Pearson correlation coefficient between their corresponding eigenvector components, where the averages are taken over the set of eigenvectors used. The method can be extended to directed graphs. It is useful to estimate vertex similarities, however it does not provide a well-defined partition of the graph.

Algorithm of Wu and Huberman. A fast algorithm by Wu and Huberman identifies communities based on the properties of resistor networks [65]. It is essentially a method for bisectioning graph, similar to spectral bisection, although partitions in an arbitrary number of communities can be obtained by iterative applications. The graph is transformed into a resistor network where each edge has unit resistance. A unit potential difference is set between two randomly chosen vertices. The idea is that, if there is a clear division in two communities of the graph, there will be a visible gap between voltage values for vertices at the borders between the clusters. The voltages are calculated by solving Kirchoff's equations: an exact resolution would be too time consuming, but it is possible to find a reasonably good approximation in a linear time for a sparse graph with a clear community structure, so the more time consuming part of the algorithm is the sorting of the voltage values, which takes time $O(n \log n)$. Any possible vertex pair can be chosen to set the initial potential difference, so the procedure should be repeated for all possible vertex pairs. The authors showed that this is not necessary, and that a limited number of sampling pairs is sufficient to get good results, so the algorithm scales as $O(n \log n)$ and is very fast. An interesting feature of the method is that it can quickly find the natural community of any vertex, without determining the complete partition of the graph. For that, one uses the vertex as source voltage and places the sink at an arbitrary vertex. The same feature is present in an older algorithm by Flake et al. [11], where one uses max-flow instead of current flow.

Previous works have shown that also the eigenvectors of the *transfer matrix* $\mathbf{T}$ can be used to extract useful information on community structure [66,67]. The element T_{ij} of the transfer matrix is $1/k_j$ if i and j are neighbors, where k_j is the degree of j, otherwise it is zero. The transfer matrix rules the process of diffusion on graphs.

Dynamic Algorithms

This section describes methods employing processes running on the graph, focusing on spin-spin interactions, random walk and synchronization.

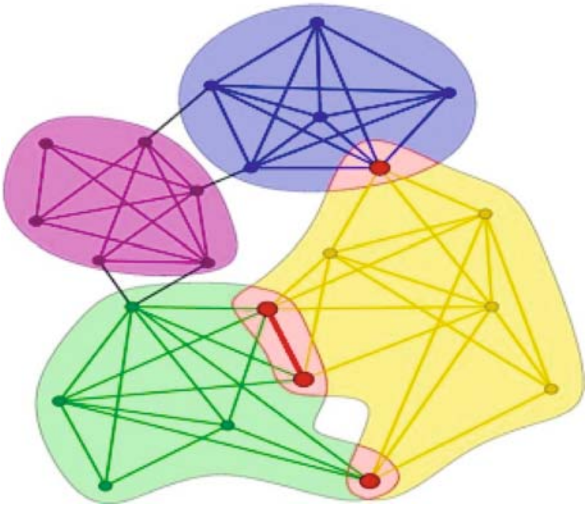
Q-state Potts model. The Potts model is among the most popular models in statistical mechanics [68]. It describes a system of spins that can be in q different states. The interaction is ferromagnetic, i. e. it favors spin alignment, so at zero temperature all spins are in the same state. If antiferromagnetic interactions are also present, the ground state of the system may not be the one where all spins are aligned, but a state where different spin values coexist, in homogeneous clusters. If Potts spin variables are assigned to the vertices of a graph with community structure, and the interactions are between neighboring spins, it is likely that the topological clusters could be recovered from like-valued spin clusters of the system, as there are many more interactions inside communities than outside. Based on this idea, inspired by an earlier paper by Blatt, Wiseman and Domany [69], Reichardt and Bornholdt proposed a method to detect communities that maps the graph onto a q-Potts model with nearest-neighbors interactions [70]. The Hamiltonian of the model, i. e. its energy, is the sum of two competing terms, one favoring spin alignment, one antialignment. The relative weight of these two terms is expressed by a parameter γ, which is usually set to the value of the density of edges of the graph. The goal is to find the ground state of the system, i. e. to minimize the energy. This can be done with simulated annealing [50], starting from a configuration where spins are randomly assigned to the vertices and the number of states q is very high. The procedure is quite fast and the results do not depend on q. The method also allows to identify vertices shared between communities, from the comparison of partitions corresponding to global and local energy minima. More recently, Reichardt and Bornholdt derived a general framework [28], in which detecting community structure is equivalent to finding the ground state of a q-Potts model spin glass [71]. Their previous method and modularity optimization are recovered as special cases. Overlapping communities can be discovered by comparing partitions with the same (minimal) energy, and hierarchical structure can be investigated by tuning a parameter acting on the density of edges of a reference graph without community structure.

Random walk. Using random walks to find communities comes from the idea that a random walker spends a long time inside a community due to the high density of edges and consequent number of paths that could be followed. Zhou used random walks to define a distance between pairs of vertices [72]: the distance d_{ij} between i and j is the average number of edges that a random walker has to cross to reach j starting from i. Close vertices are likely to belong to the same community. Zhou defines the "global attractor" of a vertex i to be the closest vertex to i (smallest d_{ij}), whereas the "local attractor" of i is its closest neighbor. Two types of communities are defined, according to local or global attractors: a vertex i has to be put in the same community of its attractor and of all other vertices for which i is an attractor. Communities must be minimal subgraphs, i. e. they cannot include smaller subgraphs which are communities according to the chosen criterion. Applications to real and artificial networks show that the method can find meaningful partitions. In a successive paper [73], Zhou introduced a measure of dissimilarity between vertices based on the distance defined above. The measure resembles the definition of distance based on structural equivalence of Eq. (5), where the elements of the adjacency matrix are replaced by the corresponding distances. Graph partitions are obtained with a divisive procedure that, starting from the graph as a single community, performs successive splits based on the criterion that vertices in the same cluster must be less dissimilar than a running threshold, which is decreased during the process. The hierarchy of partitions derived by the method is representative of actual community structures for several real and artificial graphs. In another work [74], Zhou and Lipowsky defined distances with biased random walkers, where the bias is due to the fact that walkers move preferentially towards vertices sharing a large number of neighbors with the starting vertex. A different distance measure between vertices based on random walks was introduced by Latapy and Pons [75]. The distance is calculated from the probabilities that the random walker moves from a vertex to another in a fixed number of steps. Vertices are then grouped into communities through hierarchical clustering. The method is quite fast, running to completion in a time $O(n^2 \log n)$ on a sparse graph.

Synchronization. Synchronization is another promising dynamic process to reveal communities in graphs. If oscillators are placed at the vertices, with initial random phases, and have nearest-neighbor interactions, oscillators in the same community synchronize first, whereas a full synchronization requires a longer time. So, if one follows the time evolution of the process, states with synchronized clusters of vertices can be quite stable and long-lived, so they can be easily recognized. This was first shown by Arenas, Díaz–Guilera and Pérez–Vicente [76]. They used Kuramoto oscillators [77], which are coupled two-dimensional vectors endowed with a proper frequency of oscillations. If the interaction coupling exceeds a threshold, the dynamics leads to synchronization. Arenas et al. showed that the time evolution of the system reveals some intermediate time scales, corresponding to topological scales of the graph, i. e. to different levels of organization of the vertices. Hierarchical community structure can be revealed in this way. Based on the same principle, Boccaletti et al. designed a community detection method based on synchronization [79]. The synchronization dynamics is a variation of Kuramoto's model, the opinion changing rate (OCR) model [80]. The evolution equations of the model are solved for decreasing values of a parameter that tunes the strength of the interaction coupling between neighboring vertices. In this way, different partitions are recovered: the partition with the largest value of modularity is chosen. The algorithm scales in a time $O(mn)$, or $O(n^2)$ on sparse graphs, and gives good results on practical examples. However, synchronization-based algorithms may not be reliable when communities are very different in size.

Clique Percolation

In most of the approaches examined so far, communities have been characterized and discovered, directly or indirectly, by some global property of the graph, like betweenness, modularity, etc., or by some process that involves the graph as a whole, like random walks, synchronization, etc. But communities can be also interpreted as a form of local organization of the graph, so they could be defined from some property of the groups of vertices themselves, regardless of the rest of the graph. Moreover, very few of the algorithms presented so far are able to deal with the problem of overlapping communities (Sect. "Overlapping Communities"). A method that accounts both for the locality of the community definition and for the possibility of having overlapping communities is the Clique Percolation Method (CPM) by Palla et al. [13]. It is based on the concept that the internal edges of community are likely to form cliques due to their high density. On the other hand, it is unlikely that intercommunity edges form cliques: this idea was already used in the divisive method of Radicchi et al. (see Sect. "Divisive Algorithms"). Palla et al. define a k-clique as a complete graph with k vertices. Notice that this definition is different from the definition of n-clique (see Sect. "Definition of Community") used in social science. If it were possible for a clique to move on a graph, in some way, it would probably get trapped inside its original community, as it could not cross the bottleneck formed by the intercommunity edges. Palla et al. introduced a number of concepts to implement this idea. Two k-cliques are *adjacent* if they share $k-1$ vertices. The union of adjacent k-cliques is called *k-clique chain*. Two k-cliques are connected if they are part of a k-clique chain. Finally, a *k-clique community* is the largest connected subgraph obtained by the union of a k-clique and of all k-cliques which are connected to it. Examples of k-clique communities are shown in Fig. 9. One could say that a k-clique community is identified by making a k-clique "roll" over adjacent k-cliques, where rolling means rotating a k-clique about the $k-1$ vertices it shares with any adjacent k-clique. By construction, k-clique communities can share vertices, so they can be overlapping. There may be vertices belonging to non-adjacent k-cliques, which could be reached by different paths and end up in different clusters. In order to find k-clique communities, one searches first for maximal cliques, a task that is known to require a running time that grows exponentially with the size of the graph. However, the authors found that, for the real networks they analyzed, the procedure is quite fast, allowing to analyze graphs with up to 10^5 vertices in a reasonably short time. The actual scalability of the algorithm depends on many factors, and cannot be expressed in closed form. The algorithm has been extended to the analysis of weighted [81] and directed [82] graphs. It was recently used to study the evolution of community structure in social networks [83]. A special software, called *CFinder*, based on the CPM, has been designed by Palla and coworkers and is freely available. The CPM has the same limit as the algorithm of Radicchi et al.: It assumes that the graph has a large number of cliques, so it may fail to give meaningful partitions for graphs with just a few cliques, like technological networks.



Community Structure in Graphs, Figure 9
Clique Percolation Method. The example shows communities spanned by adjacent 3-cliques (*triangles*). Overlapping vertices are shown by the *bigger* dots. Reprinted figure with permission from [13]

Other Techniques

This section describes some algorithms that do not fit in the previous categories, although some overlap is possible.

Markov Cluster Algorithm (MCL). This method, invented by van Dongen [84], simulates a peculiar process of flow diffusion in a graph. One starts from the *stochastic matrix* of the graph, which is obtained from the adjacency matrix by dividing each element A_{ij} by the degree of i. The element S_{ij} of the stochastic matrix gives the probability that a random walker, sitting at vertex i, moves to j. The sum of the elements of each column of S is one. Each iteration of the algorithm consists of two steps. In the first step, called expansion, the stochastic matrix of the graph is raised to an integer power p (usually $p = 2$). The entry M_{ij} of the resulting matrix gives the probability that a random walker, starting from vertex i, reaches j in p steps (diffusion flow). The second step, which has no physical counterpart, consists in raising each single entry of the matrix M to some power α, where α is now real-valued. This operation, called inflation, enhances the weights between pairs of vertices with large values of the diffusion flow, which are likely to be in the same community. Next, the elements of each row must be divided by their sum, such that the sum of the elements of the row equals one and a new stochastic matrix is recovered. After some iterations, the process delivers a stable matrix, with some remarkable properties. Its elements are either zero or one, so it is a sort of adjacency matrix. Most importantly, the graph described by the matrix is disconnected, and its connected components are the communities of the original graph. The method is really simple to implement, which is the main reason of its success: as of now, the MCL is one of the most used clustering algorithms in bioinformatics. Due to the matrix multiplication of the expansion step, the algorithm should scale as $O(n^3)$, even if the graph is sparse, as the running matrix becomes quickly dense after a few steps of the algorithm. However, while computing the matrix multiplication, MCL keeps only a maximum number k of non-zero elements per column, where k is usually much smaller than n. So, the actual worst-case running time of the algorithm is $O(nk^2)$ on a sparse graph. A problem of the method is the fact that the final partition is sensitive to the parameter α used in the inflation step. Therefore several partitions can be obtained, and it is not clear which are the most meaningful or representative.

Maximum likelihood. Newman and Leicht have recently proposed an algorithm based on traditional tools and techniques of statistical inference [85]. The method consists in deducing the group structure of the graph by checking which possible partition better "fits" the graph topology. The goodness of the fit is measured by the likelihood that the observed graph structure was generated by the particular set of relationships between vertices that define a partition. The latter is described by two sets of model parameters, expressing the size of the clusters and the connection preferences among the vertices, i. e. the probabilities that vertices of one cluster are linked to any vertex. The partition corresponding to the maximum likelihood is obtained by iterating a set of coupled equations for the variables, starting from a suitable set of initial conditions. Convergence is fast, so the algorithm could be applied to fairly large graphs, with up to about 10^6 vertices. A nice feature of the method is that it discovers more general types of vertex classes than communities. For instance, multipartite structure could be uncovered, or mixed patterns where multipartite subgraphs coexist with communities, etc. In this respect, it is more powerful than most methods of community detection, which are bound to focus only on proper communities, i. e. subgraphs with more internal than external edges. In addition, since partitions are defined by assigning probability values to the vertices, expressing the extent of their membership in a group, it is possible that some vertices are not clearly assigned to a group, but to more groups, so the method is able to deal with overlapping communities. The main drawback of the algorithm is the fact that one needs to specify the number of groups at the beginning of the calculation, a number that is often unknown for real networks. It is possible to derive this information self-consistently by maximizing the probability that the data are reproduced by partitions with a given number of clusters. But this procedure involves some degree of approximation, and the results are often not good.

L-shell method. This is an agglomerative method designed by Bagrow and Bollt [86]. The algorithm finds the community of any vertex, although the authors also presented a more general procedure to identify the full community structure of the graph. Communities are defined locally, based on a simple criterion involving the number of edges inside and outside a group of vertices. One starts from a vertex-origin and keeps adding vertices lying on successive shells, where a shell is defined as a set of vertices at a fixed geodesic distance from the origin. The first shell includes the nearest neighbors of the origin, the second the next-to-nearest neighbors, and so on. At each iteration, one calculates the number of edges connecting vertices of the new layer to vertices inside and outside the running cluster. If the ratio of these two numbers ("emerging degree") exceeds some predefined threshold, the vertices of the new shell are added to the cluster, otherwise the process stops. Because of the local nature of the process, the algorithm is very fast and can identify communities very quickly. By repeating the process starting from every vertex, one could derive a *membership matrix M*: the element M_{ij} is one if vertex j belongs to the community of vertex i, otherwise it is zero. The membership matrix can be rewritten by suitably permutating rows and columns based on their mutual distances. The distance between two rows (or columns) is defined as the number of entries whose elements differ. If the graph has a clear community structure, the membership matrix takes a block-diagonal form, where the blocks identify the communities. Unfortunately, the rearrangement of the matrix requires a time $O(n^3)$, so it is quite slow. In a different algorithm, local communities are discovered through greedy maximization of a local modularity measure [87].

Algorithm of Eckmann and Moses. This is another method where communities are defined based on a local criterion [88]. The idea is to use the clustering coefficient [44] of a vertex as a quantity to distinguish tightly connected groups of vertices. Many edges mean many loops inside a community, so the vertices of a community are likely to have a large clustering coefficient. The latter can be related to the average distance between pairs of neighbors of the vertex. The possible values of the distance are 1 (if neighbors are connected) or 2 (if they are not), so the average distance lies between 1 and 2. The more triangles there are in the subgraph, the shorter the average distance. Since each vertex has always distance 1 from its neighbors, the fact that the average distance between its neighbors is different from 1 reminds what happens when one measures segments on a curved surface. Endowed with a metric, represented by the geodesic distance between vertices/points, and a curvature, the graph can be embedded in a geometric space. Communities appear as portions of the graph with a large curvature. The algorithm was applied to the graph representation of the World Wide Web, where vertices are Web pages and edges are the hyperlinks that take users from a page to the other. The authors found that communities correspond to Web pages dealing with the same topic.

Algorithm of Sales–Pardo et al. This is an algorithm designed to detect hierarchical community structure (see Sect. "Hierarchies"), a realistic feature of many natural, social and technological networks, that most algorithms usually neglect. The authors [89] introduce first a similarity measure between pairs of vertices based on Newman–Girvan modularity: basically the similarity between two vertices is the frequency with which they coexist in the same community in partitions corresponding to local optima of modularity. The latter are configurations for which modularity is stable, i. e. it cannot increase if one shifts one vertex from one cluster to another or by merging or splitting clusters. Next, the similarity matrix is put in block-diagonal form, by minimizing a cost function expressing the average distance of connected vertices from the diagonal. The blocks correspond to the communities and the recovered partition represents the largest scale organization level. To determine levels at lower scales, one iterates the procedure for each subgraph identified at the previous level, which is considered as an independent graph. The method yields then a hierarchy by construction, as communities at each level are nested within communities at higher levels. The algorithm is not fast, as both the search of local optima for modularity and the rearrangement of the similarity matrix are performed with simulated annealing, but delivers good results for computer generated networks, and meaningful partitions for some social, technological and biological networks.

Algorithm by Rosvall and Bergstrom. The modular structure can be considered as a reduced description of a graph to approximate the whole information contained in its adjacency matrix. Based on this idea, Rosvall and Bergstrom [90] envisioned a communication process in which a partition of a network in communities represents a synthesis Y of the full structure that a signaler sends to a receiver, who tries to infer the original graph topology X from it. The best partition corresponds to the signal Y that contains the most information about X. This can be quantitatively assessed by the maximization of the mutual information $I(X;Y)$ [91]. The method is better than modularity optimization, especially when communities are of different size. The optimization of the mutual information is performed by simulated annealing, so the method is rather slow and can be applied to graphs with up to about 10^4 vertices.

Testing Methods

When a community detection algorithm is designed, it is necessary to test its performance, and compare it with other methods. Ideally, one would like to have graphs with known community structure and check whether the algorithm is able to find it, or how closely can come to it. In any case, one needs to compare partitions found by the method with "real" partitions. How can different partitions of the same graph be compared? Danon et al. [92] used a measure borrowed from information theory, the *normalized mutual information*. One builds a *confusion matrix* N, whose element N_{ij} is the number of vertices of the real community i that are also in the detected community j. Since the partitions to be compared may have different numbers of clusters, N is usually not a square matrix. The similarity of two partitions A and B is given by the following expression

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log(N_{ij}N/N_{i.}N_{.j})}{\sum_{i=1}^{c_A} N_{i.}\log(N_{i.}/N) + \sum_{j=1}^{c_B} N_{.j}\log(N_{.j}/N)}, \tag{8}$$

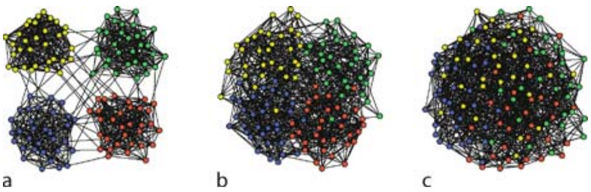
where $c_B(c_A)$ is the number of communities in partition $A(B)$, $N_{i.}$ is the sum of the elements of N on row i and $N_{.j}$ is the sum of the elements of N on column j. Another useful measure of similarity between partitions is the *Jaccard index*, which is regularly used in scientometric research.

Given two partitions A and B, the Jaccard index is defined as

$$I_J(A, B) = \frac{n_{11}}{n_{11} + n_{01} + n_{10}} , \tag{9}$$

where n_{11} is the number of pairs of vertices which are in the same community in both partitions and $n_{01}(n_{10})$ denotes the number of pairs of elements which are put in the same community in $A(B)$ and in different communities in $B(A)$. A nice presentation of criteria to compare partitions can be found in [93].

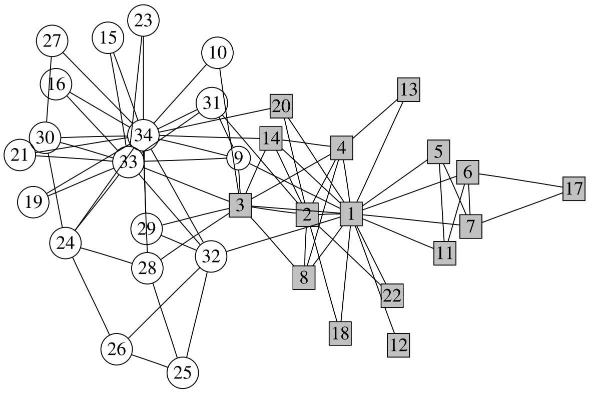
In the literature on community detection, algorithms have been generally tested on two types of graphs: computer generated graphs and real networks. The most famous computer generated benchmark is a class of graphs designed by Girvan and Newman [14]. Each graph consists of 128 vertices, arranged in four groups with 32 vertices each: 1–32, 33–64, 65–96 and 97–128. The average degree of each vertex is set to 16. The density of edges inside the groups is tuned by a parameter z_{in}, expressing the average number of edges shared by each vertex of a group with the other members (internal degree). Naturally, when z_{in} is close to 16, there is a clear community structure (see Fig. 10a), as most edges will join vertices of the same community, whereas when $z_{\text{in}} \leq 8$ there are more edges connecting vertices of different communities and the graph looks fuzzy (see Fig. 10c). In this way, one can realize different degrees of mixing between the groups. In this case the test consists in calculating the similarity between the partitions determined by the method at study and the natural partition of the graph in the four equal-sized groups. The similarity can be calculated by using the measure of Eq. (8), but in the literature one used a different quantity, i. e. the fraction of correctly classified vertices. A vertex is correctly classified if it is in the same cluster with at least 16 of its "natural" partners. If the model partition has clusters given by the merging of two or more natural groups, all vertices of the cluster are considered incorrectly classified. The number of correctly classified vertices is then divided by the total size of the graph, to yield a number between 0 and 1. One usually builds many realizations of the graph for a particular value of z_{in} and computes the average fraction of correctly classified vertices, which is a measure of the sensitivity of the method. The procedure is then iterated for different values of z_{in}. Many different algorithms have been compared with each other according to the diagram where the fraction of correctly classified vertices is plotted against z_{in}. Most algorithms usually do a good job for large z_{in} and start to fail when z_{in} approaches 8. The recipe to label vertices as correctly or incorrectly classified is somewhat arbitrary, though, and measures like those of Eq. (8) and (9) are probably more objective. There is also a subtle problem concerning the reliability of the test. Because of the randomness involved in the process of distributing edges among the vertices, it may well be that, in specific realizations of the graph, some vertices share more edges with members of another group than of their own. In this case, it is inappropriate to consider the initial partition in four groups as the real partition of the graph.



Community Structure in Graphs, Figure 10
Benchmark of Girvan and Newman. The three pictures correspond to $z_{\text{in}} = 15$ (a), $z_{\text{in}} = 11$ (b) and $z_{\text{in}} = 8$ (c). In c the four groups are basically invisible. Reprinted figure with permission from [12]

Tests on real networks usually focus on a limited number of examples, for which one has precise information about the vertices and their properties.

The most popular real network with a known community structure is the social network of Zachary's karate club (see Fig. 11). This is a social network representing the personal relationships between members of a karate club at an American university. During two years, the sociologist Wayne Zachary observed the ties between members, both inside and outside the club [94]. At some point, a conflict arose between the club's administrator (vertex 1) and one of the teachers (vertex 33), which led to the split of the club



Community Structure in Graphs, Figure 11
Zachary's karate club network, an example of graph with known community structure. Reprinted figure with permission from [26]

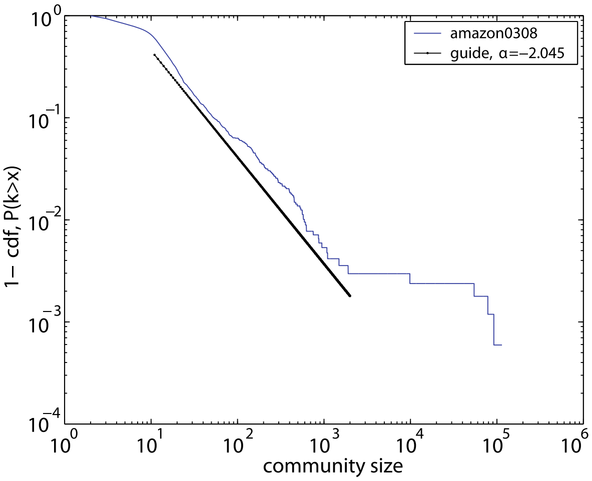
in two smaller clubs, with some members staying with the administrator and the others following the instructor. Vertices of the two groups are highlighted by squares and circles in Fig. 11. The question is whether the actual social split could be predicted from the network topology. Several algorithms are actually able to identify the two classes, apart from a few intermediate vertices, which may be misclassified (e. g. vertices 3, 10). Other methods are less successful: for instance, the maximum of Newman–Girvan modularity corresponds to a split of the network in four groups [53,63]. It is fundamental however to stress that the comparison of community structures detected by the various methods with the split of Zachary's karate club is based on a very strong assumption: that the split actually reproduced the separation of the social network in two communities. There is no real argument, beyond common wisdom, supporting this assumption.

Two other networks have frequently been used to test community detection algorithms: the network of American college football teams derived by Girvan and Newman [14] and the social network of bottlenose dolphins constructed by Lusseau [95]. Also for these networks the caveat applies: Nothing guarantees that "reasonable" communities, defined on the basis of non-topological information, must coincide with those detected by methods based only on topology.

The Mesoscopic Description of a Graph

Community detection algorithms have been applied to a huge variety of real systems, including social, biological and technological networks. The partitions found for each system are usually similar, as the algorithms, in spite of their specific implementations, are all inspired by close intuitive notions of community. What are the general properties of these partitions? The analysis of partitions and their properties delivers a *mesoscopic description* of the graph, where the communities, and not the vertices, are the elementary units of the topology. The term mesoscopic is used because the relevant scale here lies between the scale of the vertices and that of the full graph. A simple question is whether the communities of a graph are usually about of the same size or whether the community sizes have some special distribution. It turns out that the distribution of community sizes is skewed, with a tail that obeys a power law with exponents in the range between 1 and 3 [13,22,23,47]. So, there seems to be no characteristic size for a community: small communities usually coexist with large ones.

As an example, Fig. 12 shows the cumulative distribution of community sizes for a recommendation network of the online vendor Amazon.com. Vertices are products and there is a connection between item A and B if B was frequently purchased by buyers of A. We remind that the cumulative distribution is the integral of the probability distribution: if the cumulative distribution is a power law with exponent α, the probability distribution is also a power law with exponent $\alpha + 1$.



Community Structure in Graphs, Figure 12
Cumulative distribution of community sizes for the Amazon purchasing network. The partition is derived by greedy modularity optimization. Reprinted figure with permission from [47]

If communities are overlapping, one could derive a network, where the communities are the vertices and pairs of vertices are connected if their corresponding communities overlap [13]. Such networks seem to have some special properties. For instance, the degree distribution is a particular function, with an initial exponential decay followed by a slower power law decay. A recent analysis has shown that such distribution can be reproduced by assuming that the graph grows according to a simple preferential attachment mechanism, where communities with large degree have an enhanced chance to interact/overlap with new communities [21].

Finally, by knowing the community structure of a graph, it is possible to classify vertices according to their roles within their community, which may allow to infer individual properties of the vertices. A nice classification has been proposed by Guimerá and Amaral [12,96]. The role of a vertex depends on the values of two indices, the z-score and the participation ratio, that determine the position of the vertex within its own module and with respect to the other modules. The z-score compares the internal degree of the vertex in its module with the average internal degree of the vertices in the module. The participation ratio says how the edges of the vertex are distributed among the modules. Based on these two indices, Guimerá and Amaral distinguish seven roles for a vertex. These roles seem to be correlated to functions of vertices: in metabolic networks, for instance, vertices sharing many edges with vertices of other modules ("connectors") are often metabolites which are more conserved across species than other metabolites, i. e. they have an evolutionary advantage [12].

Future Directions

The problem of community detection is truly interdisciplinary. It involves scientists of different disciplines both in the design of algorithms and in their applications. The past years have witnessed huge progresses and novelties in this topic. Many methods have been developed, based on various principles. Their scalability has improved by at least one power in the graph size in just a couple of years. Currently partitions in graphs with up to millions of vertices can be found. From this point of view, the limit is close, and future improvements in this sense are unlikely. Algorithms running in linear time are very quick, but their results are often not very good.

The major breakthrough introduced by the new methods is the possibility of extracting graph partitions with no preliminary knowledge or inputs about the community structure of the graph. Most new algorithms do not need to know how many communities there are, a major drawback of computer science approaches: they derive this information from the graph topology itself. Similarly, algorithms of new generation are able to select one or a few meaningful partitions, whereas social science approaches usually produce a whole hierarchy of partitions, which they are unable to discriminate. Especially in the last two years, the quality of the output produced by some algorithms has considerably improved. Realistic aspects of community structure, like overlapping and hierarchical communities, are now often taken into account.

The main question is: is there at present a good method to detect communities in graphs? The answer depends on what is meant by "good". Several algorithms give satisfactory results when they are tested as described in Sect. "Testing Methods": in this respect, they can be considered good. However, if examined in more detail, some methods disclose serious limits and biases. For instance, the most popular method used nowadays, modularity optimization, is likely to give problems in the analysis of large graphs. Most algorithms are likely to fail in some limit, still one can derive useful indications from them: from the comparison of partitions derived by different methods one could extract the cores of real communities. The ideal method is one that delivers meaningful partitions and handles overlapping communities and hierarchy, possibly in a short time. No such method exists yet.

Finding a good method for community detection is a crucial endeavor in biology, sociology and computer science. In particular, biologists often rely on the application of clustering techniques to classify their data. Due to the bioinformatics revolution, gene regulatory networks, protein–protein interaction networks, metabolic networks, etc., are now much better known that they used to be in the past and finally susceptible to solid quantitative investigations. Uncovering their modular structure is an open challenge and a necessary step to discover properties of elementary biological constituents and to understand how biological systems work.

Bibliography

Primary Literature

1. Euler L (1736) Solutio problematis ad geometriam situs pertinentis. Commentarii Academiae Petropolitanae 8:128–140
2. Bollobás B (1998) Modern Graph Theory. Springer, New York
3. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge
4. Scott JP (2000) Social Network Analysis. Sage Publications Ltd, London
5. Barabási AL, Albert R (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
6. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: from biological nets to the Internet and WWW. Oxford University Press, Oxford
7. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
8. Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the Internet: A statistical physics approach. Cambridge University Press, Cambridge
9. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex Networks: Structure and Dynamics. Phys Rep 424:175–308
10. Erdös P, Rényi A (1959) On Random Graphs. Publicationes Mathematicae Debrecen 6:290–297
11. Flake GW, Lawrence S, Lee Giles C, Coetzee FM (2002) Self-Organization and Identification of Web Communities. IEEE Comput 35(3):66–71
12. Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433:895–900
13. Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818
14. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Nat Acad Sci USA 99(12):7821–7826
15. Lusseau D, Newman MEJ (2004) Identifying the role that animals play in their social networks. Proc R Soc Lond B 271:S477–S481

16. Pimm SL (1979) The structure of food webs. Theor Popul Biol 16:144–158
17. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments exposed in food-web structure. Nature 426:282–285
18. Granovetter M (1973) The Strength of Weak Ties. Am J Sociol 78:1360–1380
19. Burt RS (1976) Positions in Networks. Soc Force 55(1):93–122
20. Freeman LC (1977) A Set of Measures of Centrality Based on Betweenness. Sociometry 40(1):35–41
21. Pollner P, Palla G, Vicsek T (2006) Preferential attachment of communities: The same principle, but a higher level. Europhys Lett 73(3):478–484
22. Newman MEJ (2004) Detecting community structure in networks. Eur Phys J B 38:321–330
23. Danon L, Duch J, Arenas A, Díaz-Guilera A (2007) Community structure identification. In: Caldarelli G, Vespignani A (eds) Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. World Scientific, Singapore, pp 93–114
24. Bron C, Kerbosch J (1973) Finding all cliques on an undirected graph. Commun ACM 16:575–577
25. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Nat Acad Sci USA 101(9):2658–2663
26. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
27. Arenas A, Fernández A, Fortunato S, Gómez S (2007) Motif-based communities in complex networks. arXiv:0710.0059
28. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74:016110
29. Massen CP, Doye JPK (2006) Thermodynamics of community structure. arXiv:cond-mat/0610077
30. Arenas A, Duch J, Fernándes A, Gómez S (2007) Size reduction of complex networks preserving modularity. New J Phys 9(6):176–180
31. Guimerà R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70:025101(R)
32. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Nat Acad Sci USA 104(1):36–41
33. Gfeller D, Chappelier J-C, De Los Rios P (2005) Finding instabilities in the community structure of complex networks. Phys Rev E 72:056135
34. Pothen A (1997) Graph partitioning algorithms with applications to scientific computing. In: Keyes DE, Sameh A, Venkatakrishnan V (eds) Parallel Numerical Algorithms. Kluwer Academic Press, Boston, pp 323–368
35. Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49:291–307
36. Golub GH, Van Loan CF (1989) Matrix computations. John Hopkins University Press, Baltimore
37. JB MacQueen (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley. University of California Press, pp 281–297
38. Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177
39. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27:39–54
40. Tyler JR, Wilkinson DM, Huberman BA (2003) Email as spectroscopy: automated discovery of community structure within organizations. In: Huysman M, Wenger E, Wulf V (eds) Proceeding of the First International Conference on Communities and Technologies. Kluwer Academic Press, Amsterdam
41. Wilkinson DM, Huberman BA (2004) A method for finding communities of related genes. Proc Nat Acad Sci USA 101(1): 5241–5248
42. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87:198701
43. Fortunato S, Latora V, Marchiori M (2004) A method to find community structures based on information centrality. Phys Rev E 70:056104
44. Watts D, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
45. Brandes U, Delling D, Gaertler M, Görke R, Hoefer M, Nikoloski Z, Wagner D (2007) On finding graph clusterings with maximum modularity. In: Proceedings of the 33rd International Workshop on Graph-Theoretical Concepts in Computer Science (WG'07). Springer, Berlin
46. Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69:066133
47. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111
48. Danon L, Díaz-Guilera A, Arenas A (2006) The effect of size heterogeneity on community identification in complex networks. J Stat Mech Theory Exp 11:P11010
49. Pujol JM, Béjar J, Delgado J (2006) Clustering algorithm for determining community structure in large networks. Phys Rev E 74:016107
50. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
51. Massen CP, Doye JPK (2005) Identifying communities within energy landscapes. Phys Rev E 71:046101
52. Boettcher S, Percus AG (2001) Optimization with extremal dynamics. Phys Rev Lett 86:5211–5214
53. Duch J, Arenas A (2005) Community detection in complex networks using extremal optimization. Phys Rev E 72:027104
54. Newman MEJ (2006) Modularity and community structure in networks. Proc Nat Acad Sci USA 103 (23):8577–8582
55. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104
56. Reichardt J, Bornholdt S (2007) Partitioning and modularity of graphs with arbitrary degree distribution. Phys Rev E 76:015102(R)
57. Reichardt J, Bornholdt S (2006) When are networks truly modular? Phys D 224:20–26
58. Kumpula JM, Saramäki J, Kaski K, Kertész J (2007) Limited resolution in complex network community detection with Potts model approach. Eur Phys J B 56:41–45
59. Arenas A, Fernándes A, Gómez S (2007) Multiple resolution of the modular structure of complex networks. arXiv:physics/0703218
60. Ruan J, Zhang W (2007) Identifying network communities with high resolution. arXiv:0704.3759
61. Kumpula JM, Saramäki J, Kaski K, Kertész J (2007) Limited resolution and multiresolution methods in complex network community detection. In: Kertész J, Bornholdt S, Mantegna RN (eds) Noise and Stochastics in Complex Systems and Finance. Proc SPIE 6601:660116
62. Muff S, Rao F, Caflisch A (2005) Local modularity measure for network clusterizations. Phys Rev E 72:056107

63. Donetti L, Muñoz MA (2004) Detecting network communities: a new systematic and efficient algorithm. J Stat Mech Theory Exp P10012
64. Capocci A, Servedio VDP, Caldarelli G, Colaiori F (2004) Detecting communities in large networks. Phys A 352(2–4):669–676
65. Wu F, Huberman BA (2004) Finding communities in linear time: a physics approach. Eur Phys J B 38:331–338
66. Eriksen KA, Simonsen I, Maslov S, Sneppen K (2003) Modularity and extreme edges of the Internet. Phys Rev Lett 90(14):148701
67. Simonsen I, Eriksen KA, Maslov S, Sneppen K (2004) Diffusion on complex networks: a way to probe their large-scale topological structure. Physica A 336:163–173
68. Wu FY (1982) The Potts model. Rev Mod Phys 54:235–268
69. Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. Phys Rev Lett 76(18):3251–3254
70. Reichardt J, Bornholdt S (2004) Detecting fuzzy community structure in complex networks. Phys Rev Lett 93(21):218701
71. Mezard M, Parisi G, Virasoro M (1987) Spin glass theory and beyond. World Scientific Publishing Company, Singapore
72. Zhou H (2003) Network landscape from a Brownian particle's perspective. Phys Rev E 67:041908
73. Zhou H (2003) Distance, dissimilarity index, and network community structure. Phys Rev E 67:061901
74. Zhou H, Lipowsky R (2004) Network Brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. Lect Notes Comput Sci 3038:1062–1069
75. Latapy M, Pons P 92005) Computing communities in large networks using random walks. Lect Notes Comput Sci 3733: 284–293
76. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) Synchronization reveals topological scales in complex networks. Phys Rev Lett 96:114102
77. Kuramoto Y (1984) Chemical Oscillations, Waves and Turbulence. Springer, Berlin
78. Arenas A, Díaz-Guilera A (2007) Synchronization and modularity in complex networks. Eur Phys J ST 143:19–25
79. Boccaletti S, Ivanchenko M, Latora V, Pluchino A, Rapisarda A (2007) Detecting complex network modularity by dynamical clustering. Phys Rev E 76:045102(R)
80. Pluchino A, Latora V, Rapisarda A (2005) Changing opinions in a changing world: a new perspective in sociophysics. Int J Mod Phys C 16(4):505–522
81. Farkas I, Ábel D, Palla G, Vicsek T (2007) Weighted network modules. New J Phys 9:180
82. Palla G, Farkas IJ, Pollner P, Derényi I, Vicsek T (2007) Directed network modules. New J Phys 9:186
83. Palla G, Barabási A-L, Vicsek T (2007) Quantifying social groups evolution. Nature 446:664–667
84. van Dongen S (2000) Graph Clustering by Flow Simulation. Ph D thesis, University of Utrecht, The Netherlands
85. Newman MEJ, Leicht E (2007) Mixture models and exploratory analysis in networks. Proc Nat Acad Sci USA 104(23):9564–9569
86. Bagrow JP, Bollt EM (2005) Local method for detecting communities. Phys Rev E 72:046108
87. Clauset A (2005) Finding local community structure in networks. Phys Rev E 72:026132
88. Eckmann J-P, Moses E (2002) Curvature of co-links uncovers hidden thematic layers in the World Wide Web. Proc Nat Acad Sci USA 99(9):5825–5829
89. Sales-Pardo M, Guimerá R, Amaral LAN (2007) Extracting the hierarchical organization of complex systems. arXiv:0705.1679
90. Rosvall M, Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. Proc Nat Acad Sci USA 104(18):7327–7331
91. Shannon CE, Weaver V (1949) The Mathematical Theory of Communication. University of Illinois Press, Champaign
92. Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech Theory Exp P09008
93. Gustafsson M, Hörnquist M, Lombardi A (2006) Comparison and validation of community structures in complex networks. Physica A 367:559–576
94. Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthr Res 33:452–473
95. Lusseau D (2003) The emergent properties of a dolphin social network. Proc R Soc Lond B 270(2):S186–188
96. Guimerá R, Amaral LAN (2005) Cartography of complex networks: modules and universal roles. J Stat Mech Theory Exp P02001

Books and Reviews

Bollobás B (2001) Random Graphs. Cambridge University Press, Cambridge

Chung FRK (1997) Spectral Graph Theory. CBMS Regional Conference Series in Mathematics 92. American Mathematical Society, Providence

Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford

Elsner U (1997) Graph Partitioning: a Survey. Technical Report 97-27, Technische Universität Chemnitz, Chemnitz

Comparison of Discrete and Continuous Wavelet Transforms

PALLE E. T. JORGENSEN[1], MYUNG-SIN SONG[2]

[1] Department of Mathematics, The University of Iowa, Iowa City, USA

[2] Department of Mathematics and Statistics, Southern Illinois University, Edwardsville, USA

Article Outline

This *glossary* consists of a list of terms used inside the paper in mathematics, in probability, in engineering, and, on occasion, in physics. To clarify the seemingly confusing use of up to four different names for the same idea or concept, we have further added informal explanations spelling out the reasons behind the differences in current terminology from neighboring fields.

DISCLAIMER: This glossary has the structure of four areas. A number of terms are listed line by line, and each line is followed by explanation. Some "terms" have up to four separate (yet commonly accepted) names.

Glossary

MATHEMATICS: function (measurable), PROBABILITY: random variable, ENGINEERING: signal, PHYSICS: state Mathematically, functions may map between any two sets, say, from X to Y; but if X is a probability space (typically called Ω), it comes with a σ-algebra $\mathcal{B}$ of measurable sets, and probability measure P. Elements E in $\mathcal{B}$ are called events, and $\mathrm{P}(E)$ the probability of E. Corresponding measurable functions with values in a vector space are called random variables, a terminology which suggests a stochastic viewpoint. The function values of a random variable may represent the outcomes of an experiment, for example "throwing of a die".

Yet, function theory is widely used in engineering where functions are typically thought of as signal. In this case, X may be the real line for time, or $\mathbb{R}^d$. Engineers visualize functions as signals. A particular signal may have a stochastic component, and this feature simply introduces an extra stochastic variable into the "signal", for example noise.

Turning to physics, in our present application, the physical functions will be typically be in some L^2-space, and L^2-functions with unit norm represent quantum mechanical "states".

MATHEMATICS: sequence (incl. vector-valued), PROBABILITY: random walk, ENGINEERING: time-series, PHYSICS: measurement Mathematically, a sequence is a function defined on the integers $\mathbb{Z}$ or on subsets of $\mathbb{Z}$, for example the natural numbers $\mathbb{N}$. Hence, if time is discrete, this to the engineer represents a time series, such as a speech signal, or any measurement which depends on time. But we will also allow functions on lattices such as $\mathbb{Z}^d$.

In the case $d = 2$, we may be considering the grayscale numbers which represent exposure in a digital camera. In this case, the function (grayscale) is defined on a subset of $\mathbb{Z}^2$, and is then simply a matrix.

A random walk on $\mathbb{Z}^d$ is an assignment of a sequential and random motion as a function of time. The randomness presupposes assigned probabilities. But we will use the term "random walk" also in connection with random walks on combinatorial trees.

MATHEMATICS: nested subspaces, PROBABILITY: refinement, ENGINEERING: multiresolution, PHYSICS: scales of visual resolutions While finite or infinite families of nested subspaces are ubiquitous in mathematics, and have been popular in Hilbert space theory for generations (at least since the 1930s), this idea was revived in a different guise in 1986 by Stéphane Mallat, then an engineering graduate student. In its adaptation to wavelets, the idea is now referred to as the multiresolution method.

What made the idea especially popular in the wavelet community was that it offered a skeleton on which various discrete algorithms in applied mathematics could be attached and turned into wavelet constructions in harmonic analysis. In fact what we now call multiresolutions have come to signify a crucial link between the world of discrete wavelet algorithms, which are popular in computational mathematics and in engineering (signal/image processing, data mining, etc.) on the one side, and on the other side continuous wavelet bases in function spaces, especially in $L^2(\mathbb{R}^d)$. Further, the multiresolution idea closely mimics how fractals are analyzed with the use of finite function systems.

But in mathematics, or more precisely in operator theory, the underlying idea dates back to work of John von Neumann, Norbert Wiener, and Herman Wold, where nested and closed subspaces in Hilbert space were used extensively in an axiomatic approach to stationary processes, especially for time series. Wold proved that any (stationary) time series can be decomposed into two different parts: The first (deterministic) part can be exactly described by a linear combination of its own past, while the second part is the opposite extreme; it is *unitary*, in the language of von Neumann.

Von Neumann's version of the same theorem is a pillar in operator theory. It states that every isometry in a Hilbert space $\mathcal{H}$ is the unique sum of a shift isometry and a unitary operator, i. e., the initial Hilbert space $\mathcal{H}$ splits canonically as an orthogonal sum of two subspaces $\mathcal{H}_s$ and $\mathcal{H}_u$ in $\mathcal{H}$, one which carries the shift operator, and the other $\mathcal{H}_u$ the unitary part. The shift isometry is defined from a nested scale of closed spaces V_n, such that the intersection of these spaces is $\mathcal{H}_u$. Specifically,

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_n \subset V_{n+1} \subset \cdots$$

$$\bigwedge_n V_n = \mathcal{H}_u \,, \quad \text{and} \quad \bigvee_n V_n = \mathcal{H} \,.$$

However, Stéphane Mallat was motivated instead by the notion of scales of resolutions in the sense of optics. This, in turn, is based on a certain "artificial-intelligence" approach to vision and optics, developed earlier by David Marr at MIT, an approach which imitates the mechanism of vision in the human eye.

The connection from these developments in the 1980s back to von Neumann is this: Each of the closed subspaces V_n corresponds to a level of resolution in such a way that a larger subspace represents a finer resolution. Resolutions are relative, not absolute! In this view, the relative complement of the smaller (or coarser) subspace in larger space then represents the visual detail which is added in passing from a blurred image to a finer one, i. e., to a finer visual resolution.

This view became an instant hit in the wavelet community, as it offered a repository for the fundamental father and the mother functions, also called the scaling function φ, and the wavelet function ψ. Via a system of translation and scaling operators, these functions then generate nested subspaces, and we recover the scaling identities which initialize the appropriate algorithms. What results is now called the family of pyramid algorithms in wavelet analysis. The approach itself is called the multiresolution approach (MRA) to wavelets. And in the meantime various generalizations (GMRAs) have emerged.

In all of this, there was a second "accident" at play: As it turned out, pyramid algorithms in wavelet analysis now lend themselves via multiresolutions, or nested scales of closed subspaces, to an analysis based on frequency bands. Here we refer to bands of frequencies as they have already been used for a long time in signal processing.

One reason for the success in varied disciplines of the same geometric idea is perhaps that it is closely modeled on how we historically have represented numbers in the positional number system. Analogies to the Euclidean algorithm seem especially compelling.

MATHEMATICS: **operator,** PROBABILITY: **process,** ENGINEERING: **black box,** PHYSICS: **observable (if selfadjoint)** In linear algebra students are familiar with the distinctions between (linear) transformations T (here called "operators") and matrices. For a fixed operator $T\colon V \to W$, there is a variety of matrices, one for each choice of basis in V and in W. In many engineering applications, the transformations are not restricted to be linear, but instead represent some experiment ("black box", in Norbert Wiener's terminology), one with an input and an output, usually functions of time. The input could be an external voltage function, the black box an electric circuit, and the output the resulting voltage in the circuit. (The output is a solution to a differential equation.)

This context is somewhat different from that of quantum mechanical (QM) operators $T\colon V \to V$ where V is a Hilbert space. In QM, selfadjoint operators represent observables such as position Q and momentum P, or time and energy.

MATHEMATICS: **Fourier dual pair,** PROBABILITY: **generating function,** ENGINEERING: **time/frequency,** PHYSICS: ***P*/*Q*** The following dual pairs position Q/momentum P, and time/energy may be computed with the use of Fourier series or Fourier transforms; and in this sense they are examples of Fourier dual pairs. If for example time is discrete, then frequency may be represented by numbers in the interval $[\,0, 2\pi)$; or in $[\,0, 1)$ if we enter the number 2π into the Fourier exponential. Functions of the frequency are then periodic, so the two endpoints are identified. In the case of the interval $[\,0, 1)$, 0 on the left is identified with 1 on the right. So a low frequency band is an interval centered at 0, while a high frequency band is an interval centered at 1/2. Let a function W on $[\,0, 1)$ represent a probability assignment. Such functions W are thought of as "filters" in signal processing. We say that W is low-pass if it is 1 at 0, or if it is near 1 for frequencies near 0.

Low-pass filters pass signals with low frequencies, and block the others.

If instead some filter W is 1 at 1/2, or takes values near 1 for frequencies near 1/2, then we say that W is high-pass; it passes signals with high frequency.

MATHEMATICS: **convolution,** PROBABILITY: **—,** ENGINEERING: **filter,** PHYSICS: **smearing** Pointwise multiplication of functions of frequencies corresponds in the Fourier dual time-domain to the operation of convolution (or of Cauchy product if the time-scale is discrete.) The process of modifying a signal with a fixed convolution is called a linear filter in signal processing. The corresponding Fourier dual frequency function is then referred to as "frequency response" or the "frequency response function".

More generally, in the continuous case, since convolution tends to improve smoothness of functions, physicists call it "smearing."

MATHEMATICS: **decomposition (e. g., Fourier coefficients in a Fourier expansion),** PROBABILITY: **–,** ENGINEERING: **analysis,** PHYSICS: **frequency components** Calculating the Fourier coefficients is "analysis," and adding up the pure frequencies (i. e., summing the Fourier series) is called synthesis. But this view carries over more generally to engineering where there are more operations involved on the two sides, e. g., breaking up a signal into its frequency bands, transforming further, and then adding up the "banded" functions in the end. If the signal out is the same as the signal in, we say that the analysis/synthesis yields perfect reconstruction.

MATHEMATICS: **integrate (e. g., inverse Fourier transform),** PROBABILITY: **reconstruct,** ENGINEERING: **synthesis,** PHYSICS: **superposition** Here the terms related to "synthesis" refer to the second half of the kind of signal-processing design outlined in the previous paragraph.

MATHEMATICS: **subspace,** PROBABILITY: **—,** ENGINEERING: **resolution,** PHYSICS: **(signals in a) frequency band** For a space of functions (signals), the selection of certain frequencies serves as a way of selecting special signals. When the process of scaling is introduced into optics of a digital camera, we note that a nested family of subspaces corresponds to a grading of visual resolutions.

MATHEMATICS: **Cuntz relations,** PROBABILITY: **—,** ENGINEERING: **perfect reconstruction from subbands,** PHYSICS: **subband decomposition**

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1} \,, \quad \text{and} \quad S_i^* S_j = \delta_{i,j}\mathbf{1} \,.$$

MATHEMATICS: **inner product,** PROBABILITY: **correlation,** ENGINEERING: **transition probability,** PHYSICS: **probability of transition from one state to another** In many applications, a vector space with inner product captures perfectly the geometric and probabilistic features of the situation. This can be axiomatized in the language of Hilbert space; and the inner product is the most crucial ingredient in the familiar axiom system for Hilbert space.

MATHEMATICS: $\boldsymbol{f_{\text{out}} = T f_{\text{in}}}$, PROBABILITY: **—,** ENGINEERING: **input/output,** PHYSICS: **transformation of states** Systems theory language for operators $T\colon V \to W$. Then vectors in V are input, and in the range of T output.

MATHEMATICS: **fractal,** PROBABILITY: **—,** ENGINEERING: **—,** PHYSICS: **—** Intuitively, think of a fractal as reflecting similarity of scales such as is seen in fern-like images that look "roughly" the same at small and at large scales. Fractals are produced from an infinite iteration of a finite set of maps, and this algorithm is perfectly suited to the kind of subdivision which is a cornerstone of the discrete wavelet algorithm. Self-similarity could refer alternately to space, and to time. And further versatility is added, in that flexibility is allowed into the definition of "similar".

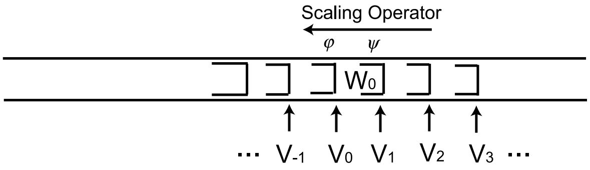
MATHEMATICS: **—,** PROBABILITY: **—,** ENGINEERING: **data mining,** PHYSICS: **—** The problem of how to handle and make use of large volumes of data is a corollary of the digital revolution. As a result, the subject of data mining itself changes rapidly. Digitized information (data) is now easy to capture automatically and to store electronically. In science, commerce, and industry, data represent collected observations and information: In business, there are data on markets, competitors, and customers. In manufacturing, there are data for optimizing production opportunities, and for improving processes. A tremendous potential for data mining exists in medicine, genetics, and energy. But raw data are not always directly usable, as is evident by inspection. A key to advances is our ability to *extract information and knowledge* from the data (hence "data mining"), and to understand the phenomena governing data sources. Data mining is now taught in a variety of forms in engineering departments, as well as in statistics and computer science departments.

One of the structures often hidden in data sets is some degree of *scale*. The goal is to detect and identify one or more natural global and local scales in the data. Once this is done, it is often possible to detect associated similarities of scale, much like the familiar scale-similarity from multidimensional wavelets, and from fractals. Indeed, various adaptations of wavelet-like algorithms have been shown to be useful. These algorithms themselves are useful in *detecting* scale-similarities, and are applicable to other types of pattern recognition. Hence, in this context, generalized multiresolutions offer another tool for discovering structures in large data sets, such as those stored in the resources of the Internet. Because of the sheer volume of data involved, a strictly manual analysis is out of the question. Instead, sophisticated query processors based on statistical and mathematical techniques are used in generating insights and extracting conclusions from data sets.

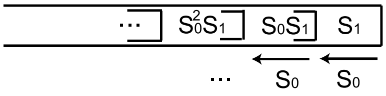
Multiresolutions Haar's work in 1909–1910 had implicitly the key idea which got wavelet mathematics started on a roll 75 years later with Yves Meyer, Ingrid Daubechies, Stéphane Mallat, and others—namely the idea of a multiresolution. In that respect Haar was ahead of his time. See Figs. 1 and 2 for details.

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \,, \quad V_0 + W_0 = V_1$$

The word "multiresolution" suggests a connection to optics from physics. So that should have been a hint to



Comparison of Discrete and Continuous Wavelet Transforms, Figure 1
Multiresolution. $L^2(\mathbb{R}^d)$-version (continuous); $\varphi \in V_0$, $\psi \in W_0$



Comparison of Discrete and Continuous Wavelet Transforms, Figure 2
Multiresolution. $l^2(\mathbb{Z})$-version (discrete); $\varphi \in V_0$, $\psi \in W_0$

mathematicians to take a closer look at trends in signal and image processing! Moreover, even staying within mathematics, it turns out that as a general notion this same idea of a "multiresolution" has long roots in mathematics, even in such modern and pure areas as operator theory and Hilbert-space geometry. Looking even closer at these interconnections, we can now recognize scales of subspaces (so-called multiresolutions) in classical algorithmic construction of orthogonal bases in inner-product spaces, now taught in lots of mathematics courses under the name of the Gram–Schmidt algorithm. Indeed, a closer look at good old Gram–Schmidt reveals that it is a matrix algorithm, Hence new mathematical tools involving non-commutativity!

If the signal to be analyzed is an image, then why not select a fixed but suitable *resolution* (or a subspace of signals corresponding to a selected resolution), and then do the computations there? The selection of a fixed "resolution" is dictated by practical concerns. That idea was key in turning computation of wavelet coefficients into iterated matrix algorithms. As the matrix operations get large, the computation is carried out in a variety of paths arising from big matrix products. The dichotomy, continuous vs. discrete, is quite familiar to engineers. The industrial engineers typically work with huge volumes of numbers.

Numbers! – So why wavelets? Well, what matters to the industrial engineer is not really the wavelets, but the fact that special wavelet functions serve as an efficient way to encode large data sets – I mean encode for computations. And the wavelet algorithms are computational. They work on numbers. Encoding numbers into pictures, images, or graphs of functions comes later, perhaps at the very end of the computation. But without the graphics, I doubt that we would understand any of this half as well as we do now. The same can be said for the many issues that relate to the crucial mathematical concept of self-similarity, as we know it from fractals, and more generally from recursive algorithms.

Definition of the Subject

In this paper we outline several points of view on the interplay between discrete and continuous wavelet transforms; stressing both pure and applied aspects of both. We outline some new links between the two transform technologies based on the theory of representations of generators and relations. By this we mean a finite system of generators which are represented by operators in Hilbert space. We further outline how these representations yield sub-band filter banks for signal and image processing algorithms.

The word "wavelet transform" (WT) means different things to different people: Pure and applied mathematicians typically give different answers the questions "What is the WT?" And engineers in turn have their own preferred quite different approach to WTs. Still there are two main trends in how WTs are used, the *continuous* WT on one side, and the *discrete* WT on the other. Here we offer a user friendly outline of both, but with a slant toward geometric methods from the theory of operators in Hilbert space.

Our paper is organized as follows: For the benefit of diverse reader groups, we begin with Glossary (Sect. "Glossary"). This is a substantial part of our account, and it reflects the multiplicity of how the subject is used.

The concept of multiresolutions or multiresolution analysis (MRA) serves as a link between the discrete and continuous theory.

In Sect. "List of Names and Discoveries", we summarize how different mathematicians and scientists have contributed to and shaped the subject over the years.

The next two sections then offer a technical overview of both discrete and the continuous WTs. This includes basic tools from Fourier analysis and from operators in Hilbert space. In Sect. "Tools from Mathematics" and Sect. "A Transfer Operator", we outline the connections between the separate parts of mathematics and their applications to WTs.

Introduction

While applied problems such as time series, signals and processing of digital images come from engineering and from the sciences, they have in the past two decades taken a life of their own as an exciting new area of applied mathematics. While searches in Google on these keywords typically yield sites numbered in the millions, the diversity of applications is wide, and it seems reasonable here to narrow our focus to some of the approaches that are both more mathematical and more recent. For references, see for example [1,6,23,31]. In addition, our own interests (e. g., [20,21,27,28]) have colored the presentation below. Each of the two areas, the discrete side, and the continuous theory is huge as measured by recent journal publications. A leading theme in our article is the independent interest in a multitude of interconnections between the discrete algorithm and their uses in the more mathematical analysis of function spaces (continuous wavelet transforms). The mathematics involved in the study and the applications of this interaction we feel is of benefit to both mathematicians and to engineers. See also [20]. An early paper [9] by Daubechies and Lagarias was especially influential in connecting the two worlds, discrete and continuous.

The Discrete vs. Continuous Wavelet Algorithms

The Discrete Wavelet Transform

If one stays with function spaces, it is then popular to pick the d-dimensional Lebesgue measure on $\mathbb{R}^d$, $d = 1, 2, \ldots$, and pass to the Hilbert space $L^2(\mathbb{R}^d)$ of all square integrable functions on $\mathbb{R}^d$, referring to d-dimensional Lebesgue measure. A wavelet basis refers to a family of basis functions for $L^2(\mathbb{R}^d)$ generated from a finite set of normalized functions ψ_i, the index i chosen from a fixed and finite index set I, and from two operations, one called scaling, and the other translation. The scaling is typically specified by a d by d matrix over the integers $\mathbb{Z}$ such that all the eigenvalues in modulus are bigger than one, lie outside the closed unit disk in the complex plane. The d-lattice is denoted $\mathbb{Z}^d$, and the translations will be by vectors selected from $\mathbb{Z}^d$. We say that we have a wavelet basis if the triple indexed family $\psi_{i,j,k}(x) := |\det A|^{j/2}\psi(A^j x + k)$ forms an orthonormal basis (ONB) for $L^2(\mathbb{R}^d)$ as i varies in I, $j \in \mathbb{Z}$, and $k \in \mathbb{R}^d$. The word "orthonormal" for a family F of vectors in a Hilbert space $\mathcal{H}$ refers to the norm and the inner product in $\mathcal{H}$: The vectors in an orthonormal family F are assumed to have norm one, and to be mutually orthogonal. If the family is also total (i. e., the vectors in F span a subspace which is dense in $\mathcal{H}$), we say that F is an orthonormal basis (ONB.)

While there are other popular wavelet bases, for example frame bases, and dual bases (see e. g., [2,18] and the papers cited there), the ONBs are the most agreeable at least from the mathematical point of view.

That there are bases of this kind is not at all clear, and the subject of wavelets in this continuous context has gained much from its connections to the discrete world of signal- and image-processing.

Here we shall outline some of these connections with an emphasis on the mathematical context. So we will be stressing the theory of Hilbert space, and bounded linear operators acting in Hilbert space $\mathcal{H}$, both individual operators, and families of operators which form algebras.

As was noticed recently the operators which specify particular subband algorithms from the discrete world of signal-processing turn out to satisfy relations that were found (or rediscovered independently) in the theory of operator algebras, and which go under the name of Cuntz algebras, denoted $\mathcal{O}_N$ if n is the number of bands. For additional details, see e. g., [21].

In symbols the C^*-algebra has generators $(S_i)_{i=0}^{N-1}$, and the relations are

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1} \tag{1}$$

(where $\mathbf{1}$ is the identity element in $\mathcal{O}_N$) and

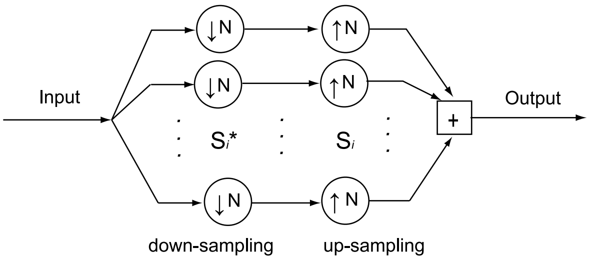
$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}\,, \quad \text{and} \quad S_i^* S_j = \delta_{i,j}\mathbf{1}\,. \tag{2}$$

In a representation on a Hilbert space, say $\mathcal{H}$, the symbols S_i turn into bounded operators, also denoted S_i, and the identity element $\mathbf{1}$ turns into the identity operator I in $\mathcal{H}$, i. e., the operator $I\colon h \to h$, for $h \in \mathcal{H}$. In operator language, the two formulas (1) and (2) state that each S_i is an isometry in $\mathcal{H}$, and that the respective ranges $S_i\mathcal{H}$ are mutually orthogonal, i. e., $S_i\mathcal{H} \perp S_j\mathcal{H}$ for $i \neq j$. Introducing the projections $P_i = S_i S_i^*$, we get $P_i P_j = \delta_{i,j} P_i$, and

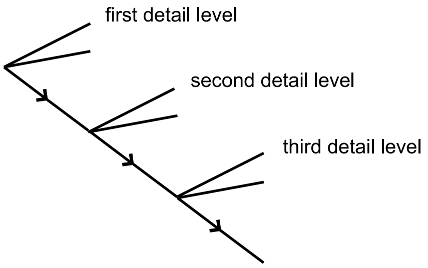
$$\sum_{i=0}^{N-1} P_i = I\,.$$

In the engineering literature this takes the form of programming diagrams: Fig. 3. If the process of Fig. 3 is repeated, we arrive at the discrete wavelet transform (Fig. 4) or stated in the form of images ($n = 5$) (Fig. 5).

Selecting a resolution subspace V_0 = closure span $\{\varphi(\cdot - k) | k \in \mathbb{Z}\}$, we arrive at a wavelet subdivision $\{\psi_{j,k} | j \geq 0, k \in \mathbb{Z}\}$, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, and the continuous expansion $f = \sum_{j,k} \langle \psi_{j,k} | f \rangle \psi_{j,k}$ or the discrete analogue derived from the isometries, $i = 1, 2, \ldots, N-1$, $S_0^k S_i$ for $k = 0, 1, 2, \ldots$; called the discrete wavelet transform.



Comparison of Discrete and Continuous Wavelet Transforms, Figure 3
Perfect reconstruction in a subband filtering as used in signal- and image-processing

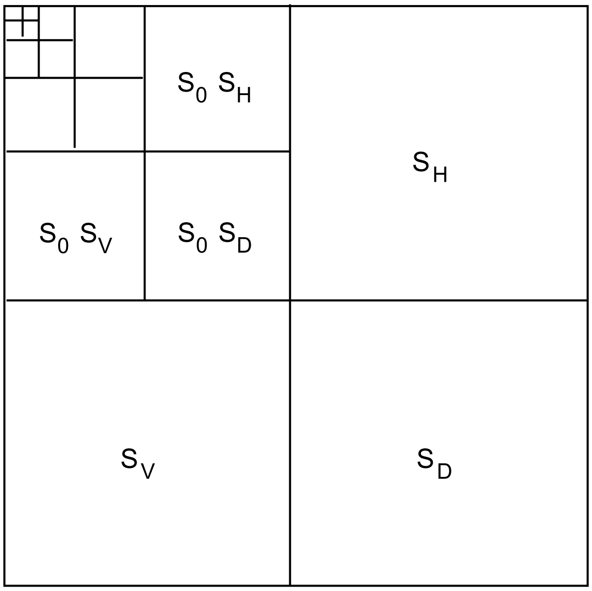


Comparison of Discrete and Continuous Wavelet Transforms, Figure 4
Binary decision tree for a sequential selection of filters.

Notational Convention In algorithms, the letter N is popular, and often used for counting more than one thing.

In the present contest of the discrete wavelet algorithm (DWA) or DWT, we count two things, "the number of times a picture is decomposed via subdivision". We have used n for this. The other related but different number N is the number of subbands, $N = 2$ for the dyadic DWT, and $N = 4$ for the image DWT. The image-processing WT in our present context is the tensor product of the 1-D dyadic WT, so $2 \times 2 = 4$. Caution: Not all DWAs arise as tensor products of $N = 2$ models. The wavelets coming from tensor products are called separable. When a particular image-processing scheme is used for generating continuous wavelets it is not transparent if we are looking at a separable or inseparable wavelet!

To clarify the distinction, it is helpful to look at the representations of the Cuntz relations by operators in Hilbert space. We are dealing with representations of the two distinct algebras $\mathcal{O}_2$, and $\mathcal{O}_4$; two frequency subbands vs. four subbands. Note that the Cuntz $\mathcal{O}_2$, and $\mathcal{O}_4$ are given axiomatic, or purely symbolically. It is only when subband filters are chosen that we get representations. This also means that the choice of N is made initially; and the same N is used in different runs of the programs. In contrast, the number of times a picture is decomposed varies from one experiment to the next!



Comparison of Discrete and Continuous Wavelet Transforms, Figure 5
The subdivided squares represent the use of the pyramid subdivision algorithm to image processing, as it is used on pixel squares. At each subdivision step the *top left-hand square* represents averages of nearby pixel numbers, averages taken with respect to the chosen low-pass filter; while the three directions, horizontal, vertical, and diagonal represent detail differences, with the three represented by separate bands and filters. So in this model, there are four bands, and they may be realized by a tensor product construction applied to dyadic filters in the separate x- and the y-directions in the plane. For the discrete WT used in image-processing, we use iteration of four isometries S_0, S_H, S_V, and S_D with mutually orthogonal ranges, and satisfying the following sum-rule $S_0S_0^* + S_HS_H^* + S_VS_V^* + S_DS_D^* = I$, with I denoting the identity operator in an appropriate l^2-space

Summary: $N = 2$ for the dyadic DWT: The operators in the representation are S_0, S_1. One average operator, and one detail operator. The detail operator S_1 "counts" local detail variations.

Image-processing. Then $N = 4$ is fixed as we run different images in the DWT: The operators are now: S_0, S_H, S_V, S_D. One average operator, and three detail operator for local detail variations in the three directions in the plane.

The Continuous Wavelet Transform

Consider functions f on the real line $\mathbb{R}$. We select the Hilbert space of functions to be $L^2(\mathbb{R})$ To start a continuous WT, we must select a function $\psi \in L^2(\mathbb{R})$ and $r, s \in \mathbb{R}$

Comparison of Discrete and Continuous Wavelet Transforms, Figure 6
$n = 2$ Jorgensen. The selection of filters (Fig 4) is represented by the use of one of the operators S_i in Fig 3. A planar version of this principle is illustrated in Fig 6. For a more detailed discussion, see e. g., [3]

such that the following family of functions

$$\psi_{r,s}(x) = r^{-1/2}\psi\left(\frac{x-s}{r}\right)$$

creates an over-complete basis for $L^2(\mathbb{R})$. An over-complete family of vectors in a Hilbert space is often called a coherent decomposition. This terminology comes from quantum optics. What is needed for a continuous WT in the simplest case is the following representation valid for all $f \in L^2(\mathbb{R})$:

$$f(x) = C_\psi^{-1} \iint_{\mathbb{R}^2} \langle\psi_{r,s}|f\rangle\psi_{r,s}(x)\frac{\mathrm{d}r\mathrm{d}s}{r^2}$$

where $C_\psi := \int_\mathbb{R} |\hat{\psi}(\omega)|^2 \frac{\mathrm{d}\omega}{\omega}$ and where $\langle\psi_{r,s}|f\rangle = \int_\mathbb{R} \overline{\psi_{r,s}(y)} f(y)\mathrm{d}y$. The refinements and implications of this are spelled out in tables in Sect. "Connections to Group Theory".

Some Background on Hilbert Space

Wavelet theory is the art of finding a special kind of basis in Hilbert space. Let $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$ and denote the inner product $\langle\cdot\mid\cdot\rangle$. For us, it is assumed linear in the second variable. If $\mathcal{H} = L^2(\mathbb{R})$, then

$$\langle f \mid g\rangle := \int_\mathbb{R} \overline{f(x)}\, g(x)\,\mathrm{d}x\,.$$

If $\mathcal{H} = \ell^2(\mathbb{Z})$, then

$$\langle \xi \mid \eta\rangle := \sum_{n\in\mathbb{Z}} \bar{\xi}_n \eta_n\,.$$

Let $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$. If $\mathcal{H} = L^2(\mathbb{T})$, then

$$\langle f \mid g\rangle := \frac{1}{2\pi}\int_{-\pi}^{\pi} \overline{f(\theta)}\, g(\theta)\,\mathrm{d}\theta\,.$$

Functions $f \in L^2(\mathbb{T})$ have Fourier series: Setting $e_n(\theta) = \mathrm{e}^{in\theta}$,

$$\hat{f}(n) := \langle e_n \mid f\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-in\theta} f(\theta)\,\mathrm{d}\theta\,,$$

and

$$\|f\|^2_{L^2(\mathbb{T})} = \sum_{n\in\mathbb{Z}} \left|\hat{f}(n)\right|^2\,.$$

Similarly if $f \in L^2(\mathbb{R})$, then

$$\hat{f}(t) := \int_\mathbb{R} \mathrm{e}^{-ixt} f(x)\,\mathrm{d}x\,,$$

and

$$\|f\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi}\int_\mathbb{R} \left|\hat{f}(t)\right|^2 \mathrm{d}t\,.$$

Let J be an index set. We shall only need to consider the case when J is countable. Let $\{\psi_\alpha\}_{\alpha\in J}$ be a family of nonzero vectors in a Hilbert space $\mathcal{H}$. We say it is an *orthonormal basis* (ONB) if

$$\langle \psi_\alpha \mid \psi_\beta\rangle = \delta_{\alpha,\beta} \qquad \text{(Kronecker delta)} \tag{3}$$

and if

$$\sum_{\alpha\in J} |\langle \psi_\alpha \mid f\rangle|^2 = \|f\|^2 \quad \text{holds for all} f \in \mathcal{H}\,. \tag{4}$$

If only Eq. (4) is assumed, but not Eq. (3), we say that $\{\psi_\alpha\}_{\alpha\in J}$ is a (normalized) *tight frame*. We say that it is a *frame* with *frame constants* $0 < A \leq B < \infty$ if

$$A\,\|f\|^2 \leq \sum_{\alpha\in J} |\langle \psi_\alpha \mid f\rangle|^2 \leq B\,\|f\|^2$$

holds for all $f \in \mathcal{H}$.

Introducing the rank-one operators $Q_\alpha := |\psi_\alpha\rangle\langle\psi_\alpha|$ of Dirac's terminology, see [3], we see that $\{\psi_\alpha\}_{\alpha\in J}$ is an ONB if and only if the Q_α's are projections, and

$$\sum_{\alpha\in J} Q_\alpha = I \qquad (= \text{the identity operator in } \mathcal{H})\,. \tag{5}$$

It is a (normalized) tight frame if and only if Eq. (5) holds but with no further restriction on the rank-one operators Q_α. It is a frame with frame constants A and B if the operator

$$S := \sum_{\alpha\in J} Q_\alpha$$

satisfies

$$AI \leq S \leq BI$$

in the order of Hermitian operators. (We say that operators $H_i = H_i^*$, $i = 1, 2$, satisfy $H_1 \leq H_2$ if $\langle f \mid H_1 f\rangle \leq \langle f \mid H_2 f\rangle$ holds for all $f \in \mathcal{H}$). If h, k are vectors in a Hilbert space $\mathcal{H}$, then the operator $A = |h\rangle\langle k|$ is defined by the identity $\langle u \mid Av\rangle = \langle u \mid h\rangle\langle k \mid v\rangle$ for all $u, v \in \mathcal{H}$.

Wavelets in $L^2(\mathbb{R})$ are generated by simple operations on one or more functions ψ in $L^2(\mathbb{R})$, the operations come in pairs, say scaling and translation, or phase-modulation and translations. If $N \in \{2, 3, \ldots\}$ we set

$$\psi_{j,k}(x) := N^{j/2}\psi\left(N^j x - k\right) \quad \text{for } j, k \in \mathbb{Z}\,.$$

Increasing the Dimension In wavelet theory, [7] there is a tradition for reserving φ for the father function and ψ for the mother function. A 1-level wavelet transform of an $N \times M$ image can be represented as

$$\mathbf{f} \mapsto \begin{pmatrix} \mathbf{a}^1 & | & \mathbf{h}^1 \\ - & & - \\ \mathbf{v}^1 & | & \mathbf{d}^1 \end{pmatrix} \tag{6}$$

where the subimages $\mathbf{h}^1, \mathbf{d}^1, \mathbf{a}^1$ and $\mathbf{v}^1$ each have the dimension of $N/2$ by $M/2$.

$$\begin{aligned}
\mathbf{a}^1 &= V_m^1 \otimes V_n^1 : \varphi^A(x, y) = \varphi(x)\varphi(y) \\
&= \sum_i \sum_j h_i h_j \varphi(2x - i)\varphi(2y - j) \\
\mathbf{h}^1 &= V_m^1 \otimes W_n^1 : \psi^H(x, y) = \psi(x)\varphi(y) \\
&= \sum_i \sum_j g_i h_j \varphi(2x - i)\varphi(2y - j) \\
\mathbf{v}^1 &= W_m^1 \otimes V_n^1 : \psi^V(x, y) = \varphi(x)\psi(y) \\
&= \sum_i \sum_j h_i g_j \varphi(2x - i)\varphi(2y - j) \\
\mathbf{d}^1 &= W_m^1 \otimes W_n^1 : \psi^D(x, y) = \psi(x)\psi(y) \\
&= \sum_i \sum_j g_i g_j \varphi(2x - i)\varphi(2y - j)
\end{aligned} \tag{7}$$

where φ is the father function and ψ is the mother function in sense of wavelet, V space denotes the average space and the W spaces are the difference space from multiresolution analysis (MRA) [7].

In the formulas, we have the following two indexed number systems $\mathbf{a} := (h_i)$ and $\mathbf{d} := (g_i)$, $\mathbf{a}$ is for averages, and $\mathbf{d}$ is for local differences. They are really the input for the DWT. But they also are the key link between the two transforms, the discrete and continuous. The link is made up of the following scaling identities:

$$\begin{aligned}
\varphi(x) &= 2\sum_{i\in\mathbb{Z}} h_i \varphi(2x - i)\,; \\
\psi(x) &= 2\sum_{i\in\mathbb{Z}} g_i \varphi(2x - i)\,;
\end{aligned}$$

and (low-pass normalization) $\sum_{i\in\mathbb{Z}} h_i = 1$. The scalars (h_i) may be real or complex; they may be finite or infinite in number. If there are four of them, it is called the "four tap", etc. The finite case is best for computations since it corresponds to compactly supported functions. This means that the two functions φ and ψ will vanish outside some finite interval on a real line.

The two number systems are further subjected to orthogonality relations, of which

$$\sum_{i\in\mathbb{Z}} \bar{h}_i h_{i+2k} = \frac{1}{2}\delta_{0,k} \tag{8}$$

is the best known.

The systems h and g are both low-pass and high-pass filter coefficients. In Eq. (7), $\mathbf{a}^1$ denotes the first averaged image, which consists of average intensity values of the original image. Note that only φ function, V space and h coefficients are used here. Similarly, $\mathbf{h}^1$ denotes the first detail image of horizontal components, which consists of intensity difference along the vertical axis of the original image. Note that φ function is used on y and ψ function on x, W space for x values and V space for y values; and both h and g coefficients are used accordingly. The data $\mathbf{v}^1$ denote the first detail image of vertical components, which consists of intensity difference along the horizontal axis of the original image. Note that φ function is used on x and ψ function on y, W space for y values and V space for x values; and both h and g coefficients are used accordingly. Finally, $\mathbf{d}^1$ denotes the first detail image of diagonal components, which consists of intensity difference along the diagonal axis of the original image. The original image is reconstructed from the decomposed image by taking the sum of the averaged image and the detail images and scaling by a scaling factor. It could be noted that only ψ function, W space and g coefficients are used here. See [28,33].

This decomposition not only limits to one step but it can be done again and again on the averaged detail depending on the size of the image. Once it stops at certain level, quantization (see [26,32]) is done on the image. This quantization step may be lossy or lossless. Then the lossless entropy encoding is done on the decomposed and quantized image.

The relevance of the system of identities Eq. (8) may be summarized as follows. Set

$$m_0(z) := \frac{1}{2}\sum_{k\in\mathbb{Z}} h_k z^k \quad \text{for all } \; z \in \mathbb{T}\;;$$

$$g_k := (-1)^k \bar{h}_{1-k} \quad \text{for all } \; k \in \mathbb{Z}\;;$$

$$m_1(z) := \frac{1}{2}\sum_{k\in\mathbb{Z}} g_k z^k; \quad \text{and}$$

$$(S_j f)(z) = \sqrt{2} m_j(z) f(z^2)\,,$$
$$\text{for } \; j = 0, 1\,, \quad f \in L^2(\mathbb{T})\,, \quad z \in \mathbb{T}\,.$$

Then the following conditions are equivalent:

(a) The system of Eq. (8) is satisfied.
(b) The operators S_0 and S_1 satisfy the Cuntz relations.
(c) We have perfect reconstruction in the subband system of Fig. 4.

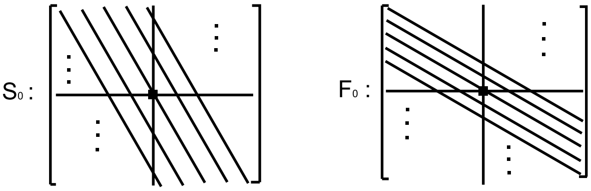
Note that the two operators S_0 and S_1 have equivalent matrix representations. Recalling that by Parseval's formula, we have $L^2(\mathbb{T}) \simeq l^2(\mathbb{Z})$. So representing S_0 instead as an $\infty \times \infty$ matrix acting on column vectors $x = (x_j)_{j\in\mathbb{Z}}$ we get

$$(S_0 x)_i = \sqrt{2}\sum_{j\in\mathbb{Z}} h_{i-2j} x_j$$

and for the adjoint operator $F_0 := S_0^*$, we get the matrix representation

$$(F_0 x)_i = \frac{1}{\sqrt{2}}\sum_{j\in\mathbb{Z}} \bar{h}_{j-2i} x_j$$

with the overbar signifying complex conjugation. This is computational significance to the two matrix representations, both the matrix for S_0, and for $F_0 := S_0^*$, is slanted.



Comparison of Discrete and Continuous Wavelet Transforms, Figure 7
Matrix representation of filters operations.

However, the slanting of one is the mirror-image of the other, i. e., Fig. 7.

Significance of Slanting The slanted matrix representations refers to the corresponding operators in L^2. In general operators in Hilbert function spaces have many matrix representations, one for each orthonormal basis (ONB), but here we are concerned with the ONB consisting of the Fourier frequencies z^j, $j \in \mathbb{Z}$. So in our matrix representations for the S operators and their adjoints we will be acting on column vectors, each infinite column representing a vector in the sequence space l^2. A vector in l^2 is said to be of finite size if it has only a finite set of non-zero entries.

It is the matrix F_0 that is effective for iterated matrix computation. Reason: When a column vector x of a fixed size, say 2s is multiplied, or acted on by F_0, the result is a vector y of half the size, i. e., of size s. So $y = F_0 x$. If we use F_0 and F_1 together on x, then we get two vectors, each of size s, the other one $z = F_1 x$, and we can form the combined column vector of y and z; stacking y on top of z. In our application, y represents averages, while z represents local differences: Hence the wavelet algorithm.

$$\begin{bmatrix} \vdots \\ y_{-1} \\ y_0 \\ y_1 \\ \vdots \\ - \\ \vdots \\ z_{-1} \\ z_0 \\ z_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} F_0 \\ - \\ F_1 \end{bmatrix} \begin{bmatrix} \vdots \\ x_{-2} \\ x_{-1} \\ x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix}$$

$$y = F_0 x$$
$$z = F_1 x$$

Comparison of Discrete and Continuous Wavelet Transforms, Table 1
Splitting of a signal into filtered components: average and a scale of detail. Several instances: Continuous vs. discrete Operator representation

$N = 2$	Overcomplete Basis	Dual Bases
Continuous resolution	$C_\psi^{-1} \iint_{\mathbb{R}^2} \frac{drds}{r^2} \lvert\psi_{r,s}\rangle\langle\psi_{r,s}\rvert = \mathbf{1}$	$C_{\psi,\tilde\psi}^{-1} \iint_{\mathbb{R}^2} \frac{drds}{r^2} \lvert\psi_{r,s}\rangle\langle\tilde\psi_{r,s}\rvert = \mathbf{1}$
Discrete resolution	$\sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \lvert\psi_{j,k}\rangle\langle\psi_{j,k}\rvert = \mathbf{1}$, $\psi_{j,k}$ Corresponding to $r = 2^{-j}, s = k2^{-j}$	$\sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \lvert\psi_{j,k}\rangle\,\langle\tilde\psi_{j,k}\rvert = \mathbf{1}$

$N \geq 2$	Isometries in ℓ^2	Dual Operator System in ℓ^2
Sequence spaces	$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}$, where $S_0, \dots, S_{N-1}$ are adjoints to the quadrature mirror filter operators F_i, i. e., $S_i = F_i^*$	$\sum_{i=0}^{N-1} S_i \tilde S_i^* = \mathbf{1}$, for a dual operator system $S_0, \dots, S_{N-1}$, $\tilde S_0, \dots, \tilde S_{N-1}$

Comparison of Discrete and Continuous Wavelet Transforms, Table 2
Application of the operator representation to specific signals, contiuous and discrete

$C_\psi^{-1} \iint_{\mathbb{R}^2} \frac{drds}{r^2} \lvert\langle \psi_{r,s} \mid f \rangle\rvert^2 = \lVert f\rVert_{L^2}^2 \quad \forall f \in L^2(\mathbb{R})$	$C_{\psi,\tilde\psi}^{-1} \iint_{\mathbb{R}^2} \frac{drds}{r^2} \langle f \mid \psi_{r,s} \rangle \langle \tilde\psi_{r,s} \mid g \rangle = \langle f \mid g \rangle \quad \forall f, g \in L^2(\mathbb{R})$
$\sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \lvert\langle \psi_{j,k} \mid f \rangle\rvert^2 = \lVert f\rVert_{L^2}^2 \quad \forall f \in L^2(\mathbb{R})$	$\sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle f \mid \psi_{j,k} \rangle \langle \tilde\psi_{j,k} \mid g \rangle = \langle f \mid g \rangle \quad \forall f, g \in L^2(\mathbb{R})$
$\sum_{i=0}^{N-1} \lVert S_i^* c\rVert^2 = \lVert c\rVert^2 \quad \forall c \in \ell^2$	$\sum_{i=0}^{N-1} \langle S_i^* c \mid \tilde S_i^* d \rangle = \langle c \mid d \rangle \quad \forall c, d \in \ell^2$

Connections to Group Theory

The first line in the two tables below is the continuous wavelet transform. It comes from what in physics is called *coherent vector decompositions*. Both transforms applies to vectors in Hilbert space $\mathcal{H}$, and $\mathcal{H}$ may vary from case to case. Common to all transforms is vector input and output. If the input agrees with output we say that the combined process yields the identity operator image. $\mathbf{1}\colon \mathcal{H} \to \mathcal{H}$ or written $\mathbf{1}_{\mathcal{H}}$. So for example if $(S_i)_{i=0}^{N-1}$ is a finite operator system, and input/output operator example may take the form

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}_{\mathcal{H}} \,.$$

The Summary of and variations on the resolution of the identity operator $\mathbf{1}$ in L^2 or in ℓ^2, for ψ and $\tilde\psi$ where $\psi_{r,s}(x) = r^{-\frac{1}{2}} \psi\left(\frac{x-s}{r}\right)$,

$$C_\psi = \int_{\mathbb{R}} \frac{d\omega}{|\omega|} |\hat\psi(\omega)|^2 < \infty \,,$$

similarly for $\tilde\psi$ and $C_{\psi,\tilde\psi} = \int_{\mathbb{R}} \frac{d\omega}{|\omega|} \overline{\hat\psi(\omega)}\, \hat{\tilde\psi}(\omega)$ is given in Table 1.

Then the assertions in Table 1 amount to the equations in Table 2.

A function ψ satisfying the resolution identity is called a *coherent vector* in mathematical physics. The representation theory for the $(ax + b)$-group, i. e., the matrix group $G = \left\{ \left(\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right) \mid a \in \mathbb{R}_+,\ b \in \mathbb{R} \right\}$, serves as its underpinning. Then the tables above illustrate how the $\{\psi_{j,k}\}$ wavelet system arises from a discretization of the following unitary representation of G:

$$\left(U_{\left(\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right)} f \right)(x) = a^{-\frac{1}{2}} f\left(\frac{x-b}{a}\right)$$

acting on $L^2(\mathbb{R})$. This unitary representation also explains the discretization step in passing from the first line to the second in the tables above. The functions $\{\, \psi_{j,k} \mid j, k \in \mathbb{Z} \,\}$ which make up a wavelet system result from the choice of a suitable coherent vector $\psi \in L^2(\mathbb{R})$, and then setting

$$\psi_{j,k}(x) = \left(U_{\left(\begin{smallmatrix} 2^{-j} & k\cdot 2^{-j} \\ 0 & 1 \end{smallmatrix}\right)} \psi \right)(x) = 2^{\frac{j}{2}} \psi\left(2^j x - k\right) \,.$$

Even though this representation lies at the historical origin of the subject of wavelets, the $(ax + b)$-group seems to be now largely forgotten in the next generation of the wavelet community. But Chaps. 1–3 of [7] still serve as a beautiful presentation of this (now much ignored) side of the subject. It also serves as a link to mathematical physics and to classical analysis.

List of Names and Discoveries

Many of the main discoveries summarized below are now lore.

1807 *Jean Baptiste Joseph Fourier mathematics, physics (heat conduction)*
Expressing functions as sums of sine and cosine waves of frequencies in arithmetic progression (now called Fourier series).

1909 *Alfred Haar mathematics*
Discovered, while a student of David Hilbert, an orthonormal basis consisting of step functions, applicable both to functions on an interval, and functions on the whole real line. While it was not realized at the time, Haar's construction was a precursor of what is now known as the Mallat subdivision, and multiresolution method, as well as the subdivision wavelet algorithms.

1946 *Denes Gabor (Nobel Prize): physics (optics, holography)*
Discovered basis expansions for what might now be called time-frequency wavelets, as opposed to time-scale wavelets.

1948 *Claude Elwood Shannon mathematics, engineering (information theory)*
A rigorous formula used by the phone company for sampling speech signals. Quantizing information, entropy, founder of what is now called the mathematical theory of communication.

1976 *Claude Garland, Daniel Esteban (both) signal processing*
Discovered subband coding of digital transmission of speech signals over the telephone.

1981 *Jean Morlet petroleum engineer*
Suggested the term "ondelettes." J.M. decomposed reflected seismic signals into sums of "wavelets (Fr.: ondelettes) of constant shape," i. e., a decomposition of signals into wavelet shapes, selected from a library of such shapes (now called wavelet series). Received somewhat late recognition for his work. Due to contributions by A. Grossman and Y. Meyer, Morlet's discoveries have now come to play a central role in the theory.

1985 *Yves Meyer mathematics, applications*
Mentor for A. Cohen, S. Mallat, and other of the wavelet pioneers, Y.M. discovered infinitely often differentiable wavelets.

1989 *Albert Cohen mathematics (orthogonality relations), numerical analysis*
Discovered the use of wavelet filters in the analysis of wavelets — the so-called Cohen condition for orthogonality.

1986 *Stéphane Mallat mathematics, signal and image processing*
Discovered what is now known as the subdivision, and multiresolution method, as well as the subdivision wavelet algorithms. This allowed the effective use of operators in the Hilbert space $L^2(\mathbb{R})$, and of the parallel computational use of recursive matrix algorithms.

1987 *Ingrid Daubechies mathematics, physics, and communications*
Discovered differentiable wavelets, with the number of derivatives roughly half the length of the support interval. Further found polynomial algorithmic for their construction (with coauthor Jeff Lagarias; joint spectral radius formulas).

1991 *Wayne Lawton mathematics (the wavelet transfer operator)*
Discovered the use of a transfer operator in the analysis of wavelets: orthogonality and smoothness.

1992 *The FBI using wavelet algorithms in digitizing and compressing fingerprints*
C. Brislawn and his group at Los Alamos created the theory and the codes which allowed the compression of the enormous FBI fingerprint file, creating A/D, a new database of fingerprints.

2000 *The International Standards Organization*
A wavelet-based picture compression standard, called JPEG 2000, for digital encoding of images.

1994 *David Donoho statistics, mathematics*
Pioneered the use of wavelet bases and tools from statistics to "denoise" images and signals.

History

While wavelets as they have appeared in the mathematics literature (e. g., [7]) for a long time, starting with Haar in 1909, involve function spaces, the connections to a host of discrete problems from engineering is more subtle. Moreover the deeper connections between the discrete algorithms and the function spaces of mathematical analysis are of a more recent vintage, see e. g., [31] and [21].

Here we begin with the function spaces. This part of wavelet theory refers to continuous wavelet transforms (details below). It dominated the wavelet literature in the 1980s, and is beautifully treated in the first four chapters in [7] and in [8]. The word "continuous" refers to the continuum of the real line $\mathbb{R}$. Here we consider spaces of functions in one or more real dimensions, i. e., functions on the line $\mathbb{R}$ (signals), the plane $\mathbb{R}^2$ (images), or in higher dimensions $\mathbb{R}^d$, functions of d real variables.

Tools from Mathematics

In our presentation, we will rely on tools from at least three separate areas of mathematics, and we will outline how they interact to form a coherent theory, and how they come together to form a link between what is now called the discrete and the continuous wavelet transform. It is the discrete case that is popular with engineers ([1,23,29,30]), while the continuous case has come to play a central role in the part of mathematics referred to as harmonic analysis, [8]. The three areas are, operator algebras, dynamical systems, and basis constructions:

a. Operator algebras. The theory of operator algebras in turn breaks up in two parts: One the study of "the algebras themselves" as they emerge from the axioms of von Neumann (von Neumann algebras), and Gelfand, Kadison and Segal (C^*-algebras.) The other has a more applied slant: It involves "the representations" of the algebras. By this we refer to the following: The algebras will typically be specified by generators and by relations, and by a certain norm-completion, in any case by a system of axioms. This holds both for the norm-closed algebras, the so called C^*-algebras, and for the weakly closed algebras, the von Neumann algebras. In fact there is a close connection between the two parts of the theory: For example, representations of C^*-algebras generate von Neumann algebras.
   To talk about representations of a fixed algebra say A we must specify a Hilbert space, and a homomorphism ρ from A into the algebra $\mathcal{B}(H)$ of all bounded operators on $\mathcal{H}$. We require that ρ sends the identity element in A into the identity operator acting on $\mathcal{H}$, and that $\rho(a^*) = (\rho(a))^*$ where the last star now refers to the adjoint operator.
   It was realized in the last ten years (see for example [3,21,22] that a family of representations that wavelets which are basis constructions in harmonic analysis, in signal/image analysis, and in computational mathematics may be built up from representations of an especially important family of simple C^*-algebras, the Cuntz algebras. The Cuntz algebras are denoted $\mathcal{O}_2, \mathcal{O}_3, \ldots$, including $\mathcal{O}_\infty$.
b. Dynamical systems. The connection between the Cuntz algebras $\mathcal{O}_N$ for $N = 2, 3, \ldots$ are relevant to the kind of dynamical systems which are built on branching-laws, the case of $\mathcal{O}_N$ representing N-fold branching. The reason for this is that if N is fixed, $\mathcal{O}_N$ includes in its definition an iterated subdivision, but within the context of Hilbert space. For more details, see e. g., [12,13,14,15,16,17,22].
c. Analysis of bases in function spaces. The connection to basis constructions using wavelets is this: The context for wavelets is a Hilbert space $\mathcal{H}$, where $\mathcal{H}$ may be $L^2(\mathbb{R}^d)$ where d is a dimension, $d = 1$ for the line (signals), $d = 2$ for the plane (images), etc. The more successful bases in Hilbert space are the orthonormal bases ONBs, but until the mid 1980s, there were no ONBs in $L^2(\mathbb{R}^d)$ which were entirely algorithmic and effective for computations. One reason for this is that the tools that had been used for 200 years since Fourier involved basis functions (Fourier wave functions) which were not localized. Moreover these existing Fourier tools were not friendly to algorithmic computations.

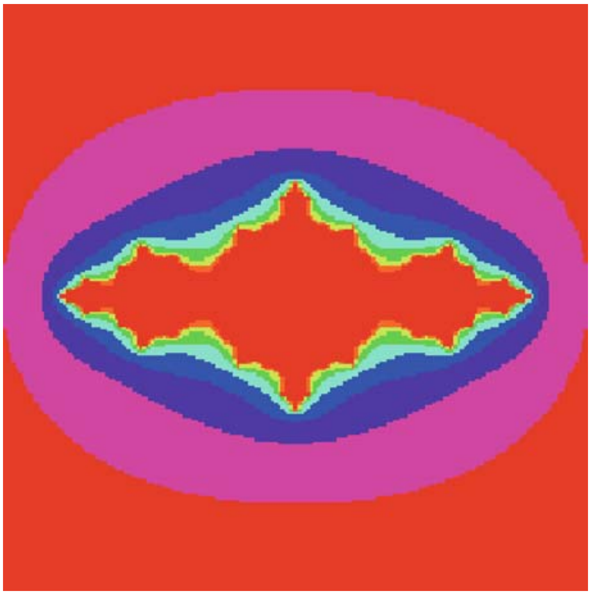
A Transfer Operator

A popular tool for deciding if a candidate for a wavelet basis is in fact an ONB uses a certain transfer operator. Variants of this operator is used in diverse areas of applied mathematics. It is an operator which involves a weighted average over a finite set of possibilities. Hence it is natural for understanding random walk algorithms. As remarked in for example [12,20,21,22], it was also studied in physics, for example by David Ruelle, who used to prove results on phase transition for infinite spin systems in quantum statistical mechanics. In fact the transfer operator has many incarnations (many of them known as Ruelle operators), and all of them based on N-fold branching laws.

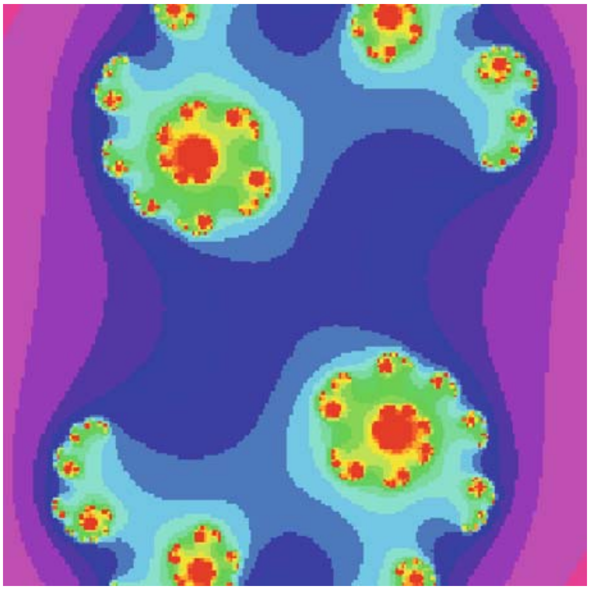
In our wavelet application, the Ruelle operator weights in input over the N branch possibilities, and the weighting is assigned by a chosen scalar function W. the and the W-Ruelle operator is denoted R_W. In the wavelet setting there is in addition a low-pass filter function m_0 which in its frequency response formulation is a function on the d-torus $\mathbf{T}^d = \mathbb{R}^d/\mathbb{Z}^d$.

Since the scaling matrix A has integer entries A passes to the quotient $\mathbb{R}^d/\mathbb{Z}^d$, and the induced transformation $r_A : \mathbb{T}^d \to \mathbb{T}^d$ is an N-fold cover, where $N = |\det A|$, i. e., for every x in $\mathbb{T}^d$ there are N distinct points y in $\mathbb{T}^d$ solving $r_A(y) = x$.

In the wavelet case, the weight function W is $W = |m_0|^2$. Then with this choice of W, the ONB problem for a candidate for a wavelet basis in the Hilbert space $L^2(\mathbb{R}^d)$



Comparison of Discrete and Continuous Wavelet Transforms, Figure 8
Julia set with $c = -1$ These images are generated by Mathematica by authors for different c values for $\varphi_c(z) = z^(2) + c$.



Comparison of Discrete and Continuous Wavelet Transforms, Figure 9
Julia set with $c = 0.45 - 0.1428i$ These images are generated by Mathematica by authors for different c values for $\varphi_c(z) = z^(2) + c$.

as it turns out may be decided by the dimension of a distinguished eigenspace for R_W, by the so called Perron–Frobenius problem.

This has worked well for years for the wavelets which have an especially simple algorithm, the wavelets that are initialized by a single function, called the scaling function. These are called the multiresolution analysis (MRA) wavelets, or for short the MRA-wavelets. But there are instances, for example if a problem must be localized in frequency domain, when the MRA-wavelets do not suffice, where it will by necessity include more than one scaling function. And we are then back to trying to decide if the output from the discrete algorithm, and the $\mathcal{O}_N$ representation is an ONB, or if it has some stability property which will serve the same purpose, in case where asking for an ONB is not feasible.

Future Directions

The idea of a scientific analysis by subdividing a fixed picture or object into its finer parts is not unique to wavelets. It works best for structures with an inherent self-similarity; this self-similarity can arise from numerical scaling of distances. But there are more subtle non-linear self-similarities. The Julia sets in the complex plane are a case in point [4,5,10,11,24,25]. The simplest Julia set come from a one parameter family of quadratic polynomials $\varphi_c(z) = z^2 + c$, where z is a complex variable and where c is a fixed parameter. The corresponding Julia sets J_c have a surprisingly rich structure. A simple way to understand them is the following: Consider the two branches of the inverse $\beta_\pm = z \mapsto \pm\sqrt{z-c}$. Then J_c is the unique minimal non-empty compact subset of $\mathbb{C}$, which is invariant under $\{\beta_\pm\}$. (There are alternative ways of presenting J_c but this one fits our purpose. The Julia set J of a holomorphic function, in this case $z \mapsto z^2 + c$, informally consists of those points whose long-time behavior under repeated iteration, or rather iteration of substitutions, can change drastically under arbitrarily small perturbations.) Here "long-time" refers to large n, where $\varphi^{(n+1)}(z) = \varphi(\varphi^{(n)}(z))$, $n = 0, 1, \ldots$, and $\varphi^{(0)}(z) = z$.

It would be interesting to adapt and modify the Haar wavelet, and the other wavelet algorithms to the Julia sets. The two papers [13,14] initiate such a development.

Literature

As evidenced by a simple Google check, the mathematical wavelet literature is gigantic in size, and the manifold applications spread over a vast number of engineering journals. While we cannot do justice to this volume st literature, we instead offer a collection of the classics [19] edited recently by C. Heil et al.

Acknowledgments

We thank Professors Dorin Dutkay and Judy Packer for helpful discussions. Work supported in part by the U.S. National Science Foundation.

Bibliography

1. Aubert G, Kornprobst P (2006) Mathematical problems in image processing. Springer, New York
2. Baggett L, Jorgensen P, Merrill K, Packer J (2005) A non-MRA C^r frame wavelet with rapid decay. Acta Appl Math 1–3:251–270
3. Bratelli O, Jorgensen P (2002) Wavelets through a looking glass: the world of the spectrum. Birkhäuser, Birkhäuser, Boston
4. Braverman M (2006) Parabolic julia sets are polynomial time computable. Nonlinearity 19(6):1383–1401
5. Braverman M, Yampolsky M (2006) Non-computable julia sets. J Amer Math Soc 19(3):551–578 (electronic)
6. Bredies K, Lorenz DA, Maass P (2006) An optimal control problem in medical image processing Springer, New York, pp 249–259
7. Daubechies I (1992) Ten lectures on wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics, vol 61, SIAM, Philadelphia
8. Daubechies I (1993) Wavelet transforms and orthonormal wavelet bases. Proc Sympos Appl Math, Amer Math Soc 47:1–33, Providence
9. Daubechies I, Lagarias JC (1992) Two-scale difference equations. II. Local regularity, infinite products of matrices and fractals. SIAM J Math Anal 23(4):1031–1079
10. Devaney RL, Look DM (2006) A criterion for sierpinski curve julia sets. Topology Proc 30(1):163–179. Spring Topology and Dynamical Systems Conference
11. Devaney RL, Rocha MM, Siegmund S (2007) Rational maps with generalized sierpinski gasket julia sets. Topol Appl 154(1):11–27
12. Dutkay DE (2004) The spectrum of the wavelet galerkin operator. Integral Equations Operator Theory 4:477–487
13. Dutkay DE, Jorgensen PET (2005) Wavelet constructions in non-linear dynamics. Electron Res Announc Amer Math Soc 11:21–33
14. Dutkay DE, Jorgensen PET (2006) Hilbert spaces built on a similarity and on dynamical renormalization. J Math Phys 47(5):20
15. Dutkay DE, Jorgensen PET (2006) Iterated function systems, ruelle operators, and invariant projective measures. Math Comp 75(256):1931–1970
16. Dutkay DE, Jorgensen PET (2006) Wavelets on fractals. Rev Mat Iberoam 22(1):131–180
17. Dutkay DE, Roysland K (2007) The algebra of harmonic functions for a matrix-valued transfer operator. arXiv:math/0611539
18. Dutkay DE, Roysland K (2007) Covariant representations for matrix-valued transfer operators. arXiv:math/0701453
19. Heil C, Walnut DF (eds) (2006) Fundamental papers in wavelet theory. Princeton University Press, Princeton, NJ
20. Jorgensen PET (2003) Matrix factorizations, algorithms, wavelets. Notices Amer Math Soc 50(8):880–894
21. Jorgensen PET (2006) Analysis and probability: wavelets, signals, fractals. grad texts math, vol 234. Springer, New York
22. Jorgensen PET (2006) Certain representations of the cuntz relations, and a question on wavelets decompositions. In: Operator theory, operator algebras, and applications. Contemp Math 414:165–188 Amer Math Soc, Providence
23. Liu F (2006) Diffusion filtering in image processing based on wavelet transform. Sci China Ser F 49(4):494–503
24. Milnor J (2004) Pasting together julia sets: a worked out example of mating. Exp Math 13(1):55–92
25. Petersen CL, Zakeri S (2004) On the julia set of a typical quadratic polynomial with a siegel disk. Ann Math (2) 159(1):1–52
26. Skodras A, Christopoulos C, Ebrahimi T (2001) JPEG 2000 still image compression standard. IEEE Signal Process Mag 18:36–58
27. Song MS (2006) Wavelet image compression. Ph.D. thesis, University of Iowa
28. Song MS (2006) Wavelet image compression. In: Operator theory, operator algebras, and applications. Contemp. Math., vol 414. Amer. Math. Soc., Providence, RI, pp 41–73
29. Strang G (1997) Wavelets from filter banks. Springer, Singapore, pp 59–110
30. Strang G (2000) Signal processing for everyone. Computional mathematics driven by industrial problems (Martina F 1999), pp 365–412. Lect Notes Math, vol 1739. Springer, Berlin
31. Strang G, Nguyen T (1996) Wavelets and filter banks. Wellesley-Cambridge Press, Wellesley
32. Usevitch BE (Sept. 2001) A tutorial on modern lossy wavelet image compression: foundations of jpeg 2000. IEEE Signal Process Mag 18:22–35
33. Walker JS (1999) A primer on wavelets and their scientific applications. Chapman & Hall, CRC

Complex Dynamics of Traffic Management, Introduction to

Boris S. Kerner
GR/ETI, HPC: G021, Daimler AG, Sindelfingen, Germany

Traffic management is an extremely complex dynamic process associated with the spatiotemporal behavior of many-particle non-linear systems. In this Encyclopedia of Complexity and Systems Science, there are several review articles devoted to various aspects of the complex dynamics of vehicular traffic management, air traffic management, and pedestrian traffic management.

The complexity of vehicular traffic management is associated with non-linear interactions between the following three main dynamic processes:

(i) travel *decision* behavior,
(ii) traffic *assignment* in a traffic network, and
(iii) traffic flow behavior, in particular, *traffic congestion* occurrence within the network.

Travel decision behavior determines travel demand. Traffic assignment in the network is associated with travel supply. However, traffic congestion occurring within the traffic network restricts free flow travel. This influences both travel decision behavior (e. g., because of traffic congestion, a person decides to stay at home or travel by train rather than by car) and traffic assignment in the network (e. g., because of traffic congestion on a route from an origin to a destination usually used, a person changes their route of travel). The articles presented in this Encyclopedia cover all these main scientific fields of traffic management.

The review article of Goulias (see ▶ Travel Behavior and Demand Analysis and Prediction) is devoted to a detailed consideration of the theoretical and modeling approaches to travel behavior analysis. As shown in Goulias's article, traveler values and attitudes are referred to motivational, cognitive, situational, and disposition factors determining human behavior. Goulias presents a sequential chronological order merging technological innovations and theoretical innovations to the modeling of travel behavior. There are a diverse variety of approaches to the modeling of travel behavior, which are based on mathematical methods like multi-agent microsimulation widely used in many other fields of systems science. Inputs to these models are the typical regional model data about social, economic, and demographic information of potential travelers and land use information to create schedules followed by people in their everyday life. The output are detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity. The in-depth understanding of transportation-related human behavior is essential to all kinds of traffic management. This is because travel behavior refers primarily to the modeling and analysis of travel demand.

Traffic assignment and control as well as methods for traffic prediction in various traffic networks are reviewed in the articles of Rakha and Tawfik (see ▶ Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment), Gartner and Stamatiadis (see ▶ Traffic Networks, Optimization and Control of Urban), Hegyi, Bellemans and De Schutter (see ▶ Freeway Traffic Management and Control), Rehborn and Klenov (see ▶ Traffic Prediction of Congested Patterns). In particular, Rakha and Tawfik review methods for modeling of dynamic traffic assignment in traffic networks. Gartner and Stamatiadis consider methods for traffic modeling and light signal control in city and urban traffic networks. Traffic control methods developed for traffic on freeway networks are discussed in the article of Hegyi, Bellemans, and De Schutter. Rehborn and Klenov discuss methods for traffic prediction required for dynamic traffic assignment in urban and freeway traffic networks.

In an urban network with short enough network links, dynamic traffic phenomena are determined mostly by traffic signals and other traffic regulations at link intersections. The dynamic traffic phenomena in the urban network can be nevertheless very complex, because traffic regulations at one of the link intersections can have a great influence on the probability of traffic congestion occurring on other links of the network. In their article, Gartner and Stamatiadis (▶ Traffic Networks, Optimization and Control of Urban) present a historical model development beginning from a consideration of traffic control methods at an isolated link intersection to a discussion of very complex approaches to dynamic traffic signal control and optimization in a complex city network.

Driver interactions, which are always essential for traffic dynamics on highways, lead to very complex spatiotemporal phenomena in vehicular traffic.

Hegyi, Bellemans and De Schutter (▶ Freeway Traffic Management and Control) present an introduction to control and managements methods currently used in field trials on freeways as well as to some recent control and managements approaches. These methods include on-ramp feedback metering, congestion pricing, speed limitation, collective and individual route guidance systems. Future traffic management methods will include also diverse driver assistance systems, in particular, with the use of vehicle ad-hoc networks and vehicle-to-infrastructure communication.

A dynamic traffic management in a traffic network requires a dynamic traffic assignment model. Rakha and Tawfik (▶ Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment) provide a detailed discussion of the problem of dynamic traffic assignment. A dynamic traffic assignment model should find the link inflows for the network. The model usually includes a traffic flow model (traffic modeler), which makes a prognosis of traffic in the network, and a traffic routing model (traffic router) associated with the problem of traffic optimization. The traffic router computes the sequence of roadways that minimize some utility objective function, i. e., minimize travel "costs" of the traffic network. Examples of the travel costs are travel time, fuel consumption, or HC and CO_2 emissions. In a real traffic network, traffic routing can be organized through the use for example of individual and collective driver guidance systems as well as congestion pricing. The traffic modeler and router are connected by a feedback loop. As a result, traffic congestion in the network predicted by the traffic modeler changes results of dynamic traffic assignment considerably. For this reason, the traffic modeler should model traffic congestion as close as possible to real traffic congestion found in empirical observations.

In the dynamic traffic management models, a reliable prognosis of traffic is also required. As shown in the article of Rehborn and Klenov (▶ Traffic Prediction of Congested Patterns), there are two main approaches to traffic prognosis: (a) "physics-of-traffic", i. e., traffic prognosis models and methods based on the understanding of the nature of empirical spatiotemporal features of traffic patterns; (b) methods of mathematical statistics or artificial intelligence, which are also called "data mining" methods, in which the understanding of traffic patterns is not necessarily needed.

Data mining models for traffic prognosis do not usually require the understanding of traffic data, which these methods use for traffic prediction. Using a huge number of historical traffic data, such a method learns predictable features of traffic data without the pretension of their understanding. The data mining methods used for traffic prognosis include statistical methods based on regression, a wavelet approach, or filtering models as well as neural networks. It should be noted that there are many articles in this Encyclopedia, in which these and other methods of artificial intelligence are reviewed.

The physics-of-traffic approach to traffic prognosis is based on the understanding of measurements of traffic variables made in space and time as well as on simulations of the traffic variables associated with empirical data. The key element for a successful prediction concept lies in the correct and reliable understanding of the empirical features of the traffic process and their reproducible characteristics. Furthermore, this understanding of real traffic should be incorporated in a traffic flow model, which should explain and predict the empirical traffic features. Therefore, in the physics-of-traffic approach a traffic flow prediction model is based on empirical features of phase transitions or/and resulting congested patterns; current and historical measured traffic data are used in the model, which reconstructs traffic phases and makes the tracking and prediction of the propagation of traffic congestion in a traffic network.

Thus we see that the understanding of empirical traffic congestion is the key for effective traffic management, control, organization, and all other traffic applications.

Empirical congested patterns exhibit a complex spatiotemporal behavior that was studied during the last 75 years by several generations of scientists. It was found that traffic congestion in the traffic network results from traffic breakdown in an initially free flow: vehicle speeds decrease abruptly to lower speeds in congested traffic. Traffic breakdown is observed mostly at highway bottlenecks. A bottleneck can be a result of road works, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, bad weather conditions, accidents, etc. In congested traffic a "stop-and-go" mode, i. e., a sequence of moving traffic jams is very often observed.

However, the puzzle of empirical spatiotemporal features of traffic congestion has been solved only recently. Consequently, an alternative traffic flow theory that is called three-phase traffic theory and the associated three-phase traffic flow models, which can predict and explain the empirical spatiotemporal features of traffic congestion, have had to be introduced. The development of various traffic flow theories and modeling approaches to traffic congestion is discussed in (▶ Traffic Congestion, Modeling Approaches to). These modeling approaches include microscopic, mesoscopic and macroscopic traffic flow models, in particular based on cellular automata and gas kinetic models widely used in many other fields of systems science. A probabilistic theory of traffic breakdown is considered in the review article of Kerner and Klenov (see ▶ Traffic Breakdown, Probabilistic Theory of). Spatiotemporal features of traffic congested patterns resulting from traffic breakdown are considered in the review article (▶ Traffic Congestion, Spatiotemporal Features of). Empirical observations of traffic congestion discussed in these articles show that phenomena of traffic breakdown and the subsequent evolution of resulting congested patterns are associated with diverse phase transitions and complex spatiotemporal self-organization effects in vehicular traffic.

On the one hand, the phase transitions and self-organization traffic phenomena, which can occur randomly over time and space within a traffic network, depend considerably on the spatiotemporal distribution of traffic demand in the network. On the other hand, the occurrence of traffic congestion changes travel decision behavior and traffic assignment considerably. In turn, travel decision behavior and traffic assignment determine the spatiotemporal distribution of traffic demand in this network. Thus the recent understanding of empirical features of the phase transitions and spatiotemporal effects in vehicular traffic (see ▶ Traffic Congestion, Modeling Approaches to and ▶ Traffic Congestion, Spatiotemporal Features of) should be taken into account in future models and theories of dynamic traffic assignment as well as in the subsequent development of traffic management and control strategies. This will be one of the most important and difficult challenges for transportation research in the next future.

Thus the complexity of traffic management is associated with diverse phase transitions in vehicular traffic resulting in various complex self-organizing spatiotemporal congested patterns propagating within a traffic network as well as with the necessity in the optimization of the traffic congested patterns. This optimization should ensure either the dissolution of traffic congestion or, if this is not possible to achieve, the minimization of the influence of traffic congestion on travel costs like travel time, fuel consumption, or HC and CO_2 emissions.

Modeling approaches to vehicular traffic management discussed in the mentioned above review articles of the Encyclopedia are also very important for studies of air traffic and pedestrian traffic.

Air traffic continues to grow at a steady pace in the world. For this reason, the discussion of air traffic dynamics as well as approaches to air traffic modeling made in the review article of Sridhar and Sheth (see ► Air Traffic Control, Complex Dynamics of) is an important contribution to the fields of traffic management.

Currently, we can observe a continuous growth of publications devoted to modeling approaches to pedestrian traffic and crowd dynamics. In particular, the interest to pedestrian traffic is associated with necessity in the development of effective evacuation strategies required for the relocation of people to safely escape hazardous disaster impacts. During the evacuation, destinations are chosen to escape the hazardous disaster impacts as quickly as possible. This is a peculiarity of the evacuation process in comparison with vehicle traffic in which every vehicle has usually a defined destination. This very interesting and important field of traffic management is considered in the review articles of Helbing and Johansson, Schadschneider et al., and Goudie.

Helbing and Johansson (see ► Pedestrian, Crowd and Evacuation Dynamics) and Schadschneider et al. (see ► Evacuation Dynamics: Empirical Results, Modeling and Applications) consider various aspects of the modeling of pedestrian interactions during evacuation as well as a comparison of the associated theoretical and empirical results. Based on an example of Australian fire zone research, Goudie (see ► Evacuation as a Communication and Social Phenomenon) considers evacuation as a communication and social phenomenon. Goudie's article describes the influence of communication and social aspects on possible evacuation scenarios.

Acknowledgments

I would like to thank Konstadinos Goulias, Sergey Klenov, Hesham Rakha, Hubert Rehborn, Andreas Schadschneider, and Bart De Schutter for useful suggestions.

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination

SUI HUANG, STUART A. KAUFFMAN
Institute for Biocomplexity and Informatics,
Department of Biological Sciences, University of Calgary,
Calgary, Canada

Article Outline

Glossary

Transcription factor A protein that binds to the regulatory region of a target gene (its promoter or enhancer regions) and thereby, controls its expression (*transcription* of the target gene into a mRNA which is ultimately *translated* into the *protein* encoded by the target gene). Transcription factors account for temporal and contextual specificity of the expression of genes; for instance, a developmentally regulated gene is expressed only during a particular phase in development and in particular tissues.

Gene regulatory network (GRN) Transcription factors regulate the expression of other transcription factor genes as well as other 'non-regulatory' genes which encode proteins, such as metabolic enzymes or structural proteins. A regulatory relationship between two genes thus is formalized as: "transcription factor A is the regulator of target gene B" or: A → B. The entirety of such regulatory interactions forms a network = the gene regulatory network (GRN). Synonyms: genetic network or gene network, transcriptional network.

Gene network architecture and topology The GRN can be represented as a "*directed graph*". The latter consists of a set of nodes (= vertices) representing the genes connected by arrows (directed *links* or *edges*) representing the regulatory interactions pointing from the regulator to the regulated target gene. [In contrast, in an *undirected* graph, the links are simple lines without arrowheads. The protein-interaction network can be represented as undirected graph]. The topology of a network is the structure of this graph and is an abstract notation, without physicality, of all the potential regulatory interactions between the genes. Topology usually is used to denote the simple interactions captured by the directed graph. For defining the network dynamics, however, additional aspects of interactions need to be specified, including: modalities or "sign" of an arrow (inhibitory vs. activating regulation), the 'transfer functions' (relationship between magnitude of input to that of the output = target gene) and the logical function (notably, in Boolean network, defining how multiple inputs are related to each other and are integrated to shaping the output). In this article when all this additional information is implied to be included, the term gene network *architecture* is used. Thus, the graph topology is a subset of network architecture.

Gene network dynamics The collective change of the gene expression levels of the genes in a network, essentially, the change over time of the network state *S*.

State space Phase space = the abstract space that contains all possible states *S* of a dynamical system. For (autonomous) gene regulatory networks, each state *S* is specified by the configuration of the expression levels of each of the *N* genes of the network; thus a system state S is one point in the *N*-dimensional state space. As the system changes its state over time, S moves along trajectories in the state space.

Transcriptome Gene expression pattern across the entire (or large portion of) the genome, measured at the level of mRNA levels. Used as synonym to "gene expression profile". The transcriptome can in a first approximation be considered a snapshot of the network state *S* of the GRN in gene network dynamics.

Cell type A distinct, whole-cell phenotype characteristic of a mature cell specialized to exert an organ-specific function. Example of cell types are: liver cell, red blood cell, skin fibroblast, heart muscle cell, fat cell, etc. Cell types are characterized by their distinct cell morphology and their gene expression pattern.

Cell fate A potential developmental outcome of a (stem or progenitor) cell. A cell fate of a stem cell can be the development into a particular mature cell type.

Multipotency The ability of a cell to generate multiple cell types; a hallmark of stem cells. Stem cells are said to be multipotent (see also under Stem cells).

Stem cell A multi-potent cell capable of "self-renewal" (division in which both daughter cells have the same degree of multi-potency as the mother cell) and can give rise to multiple cell types. There is a hierarchy of multipotency: a *toti-potent embryonic stem cell* can generate all possible cell types in the body, including extra-embryonic tissues, such as placenta. A *pluripotent embryonic stem cell* can generate tissues of three germ layers, i. e., it can produce all cell types of the foetus and the adult. A *multipotent (sensu stritiore) stem cell* of a tissue (e. g., blood) can give rise to all cell types of that tissue (e. g., a hematopoietic stem cell can produce all the blood cells). A multipotent *progenitor* cell can give rise to more than one cell types within a tissue (e. g. the granulocyte-monocyte progenitor cell).

Cell lineage Developmental trajectory of a multipotent cell towards one of multiple cell types, e. g., the "myeloid lineage" among blood cells, comprising the white blood cells granulocytes, monocytes, etc. Thus, a cell fate decision is a decision between multiple lineages accessible to a stem or progenitor cell.

Differentiation The process of cell fate decision in a stem or progenitor cell and the subsequent maturation into a mature cell type.

Definition of the Subject

Current studies of complex gene regulatory networks (GRN) in which thousands of genes regulate each others' expression have revealed interesting features of the network structure using graph theory methods. But how does the particular network architecture translate into biology? Since individual genes alter their expression level as a consequence of the network interactions, the genome-wide gene expression pattern (*transcriptome*), which manifests the dynamics of the entire network, changes as a whole in a highly constrained manner. The transcriptome in turn determines the cell phenotype. Hence, the constraints in the global dynamics of the GRN directly map into the most elementary "biological observable": the existence of distinct cell types in the metazoan body and their development from pluripotent stem cells.

In this article a historical overview of the various levels at which GRNs are studied, starting from network architecture analysis to the dynamics are first presented. An introduction is given to continuous and discrete value models of GRN commonly used to understand the dynamics of small genetic circuits or of large genome wide networks, respectively. This will allow us to explain how the intuitive metaphor of the "epigenetic landscape", a key idea that was proposed by Waddington in the 1940s to explain the generation of discrete cell fates, formally arises from gene network dynamics. This central idea appears in its modern form in formal and molecular terms as the concept that cell types represent *attractor states* of GRNs, first proposed by Kauffman in 1969. This raises two fundamental questions currently addressed by experimental biologists in the era of "systems biology": (1) Are cell types in the metazoan body indeed high-dimensional attractors of GRNs and (2) is the dynamics of GRNs in the "critical regime" – poised between order and chaos? A summary of recent experimental findings on these questions is given, and the broader implications of network concepts for cell fate commitment of stem cells are also briefly discussed. The idea of the epigenetic landscape is key to our understanding of how genes and gene regulatory networks give rise to the observable cell behavior, and thus to a formal and integrated view of molecular causation in biology.

Introduction

With the rise of "systems biology" over the past decade, molecular biology is moving away from the paradigm of linear genetic pathways which has long served as linear chains of causation (e. g., Gene A → Gene B→ Gene C→ phenotype) in explaining cell behaviors. It has begun to embrace the idea of molecular networks as an integrated information processing system of the cell [93,138]. The departure from the gene-centered, mechanistic 'arrow-arrow' schemes that embody 'proximal causation' [189] towards an integrative view will also entail a change in our paradigm of what an "explanation" means in biology: How do we map the collective behavior of thousands of interacting genes, obtained from molecular dissection, to the "***biological observable***"? The latter term, borrowed from the statistical physics idea of the "macroscopic observable", is most prosaically epitomized in whole-cell behavior in metazoan organisms: its capacity to adopt a large variety of easily recognizable, discretely distinct phenotypes, such as a liver cell vs. a red blood cell, or different functional states, such as the proliferative, quiescent or the apoptotic state.

All these morphologically and functionally distinct phenotypes are produced by the very same set of genes of the genome as a result of the joint action of the genes. This is achieved by the differential expression ("turning ON or OFF") of individual genes. Thus, in a first approximation, each cell phenotype is determined by a specific configuration of the status of expression of all individual genes across the genome. This genome-wide ***gene expression pattern or profile***, or ***transcriptome***, is the direct output of the GRN and maps almost uniquely into a cell phenotype (see overview in Fig. 1).

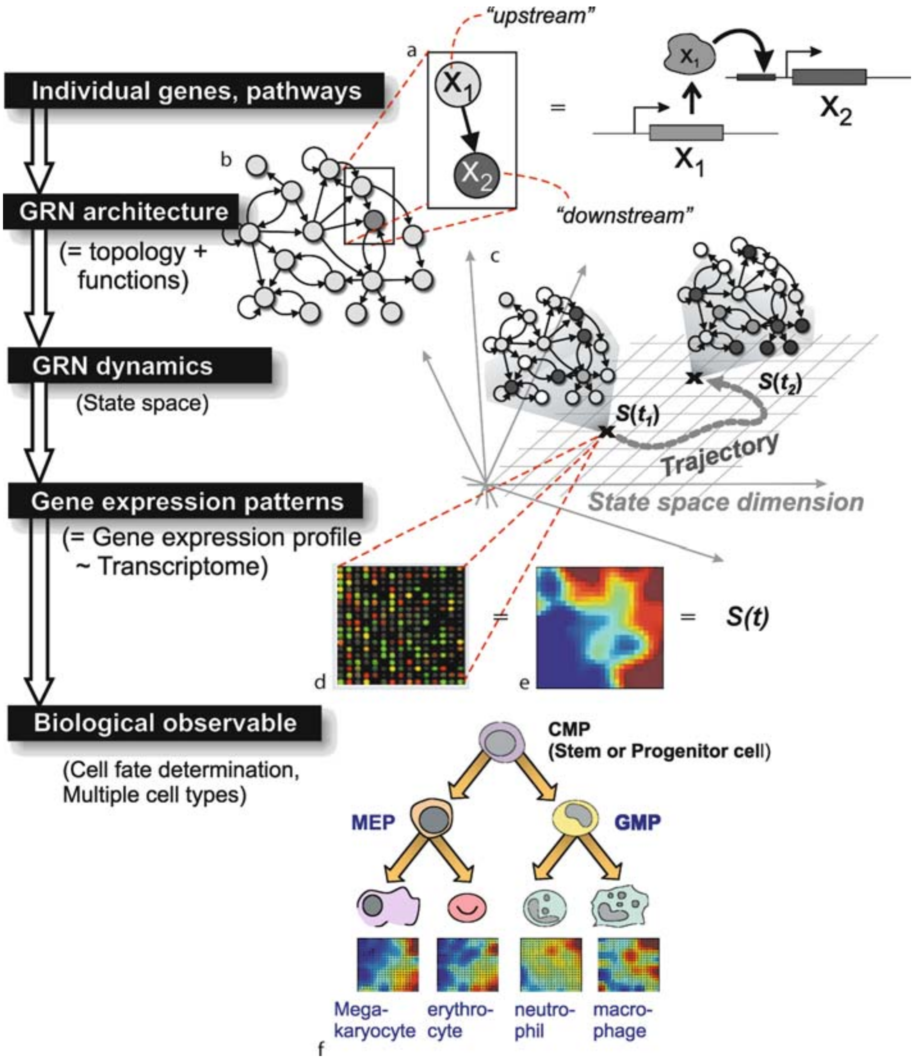
A ***network*** is the most elementary conceptualization of a complex system which is composed of interacting elements and whose behavior as an entity we wish to understand: it formalizes how a set of distinct elements (nodes) influence each other as predetermined by a fixed scheme of their interactions (links between the nodes) and the modality of interactions (rules associated with each node). In this article we focus on the ***gene regulatory network*** (GRN), the network formed by interactions through which genes regulate each other's expression (Fig. 1a), and we ask how they control the global ***behavior*** of the network, and thereby, govern development of cells into the thousands of cell types found in the metazoan body.

Because in a network a node exerts influence onto others, we can further formalize networks as ***directed graphs***, that is, the links are arrows pointing from one node to another (Fig. 1b). The information captured in an undirected or directed graph is the ***network topology*** (see Glossary). In addition, the arrows have a ***modality*** ("sign"), namely, activating or inhibiting their target gene. But since each node can receive several inputs ("upstream regulators"), it is more appropriate to combine modality of interaction together with the way the target gene integrates the various inputs to change its expression behavior (= output). Thus, each node can be assigned a ***function*** that maps all its inputs in a specific way to the output (Fig. 1, top). For instance, "promoter logics" [209] which may dictate that two stimulating inputs, the transcription factors, act synergistically, or that one inhibitory input may, when present, override all other activating inputs, is one way to represent such an input integrating function.

Here we use ***network architecture*** as a term that encompasses both network topology as well as the interaction modalities or the functions. The latter add the ingredients to the topology information that are necessary to describe the ***dynamics*** (***behavior***) of the network.

The Core Gene Regulatory Network (GRN) in Mammals

How complex is the effective GRN of higher, multicellular organisms, such as mammals? Virtually every cell type in mammals contains the ∼25,000 genes in the genome [44,158,178] which could potentially interact with each other. However, in a first approximation, we do not need to deal with all the 25,000 genes but only with those intrinsic regulators that have direct influence on other genes. A subset of roughly 5–10% of the genes in the genome encode ***transcription factors*** (TF) [194] a class of DNA binding pro-



Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 1

Overview: Concepts at different levels for understanding emergence of biological behavior from genes and networks. a Elementary interaction: regulator gene X_1 interacts with target (regulated) gene X_2, simplified as one link of a graph and the corresponding standard 'cartoon' it represents. b Notation of gene regulatory network topology as *directed graph*. c Schematic of a higher dimensional gene expression state space, showing that a network state *S*(*t*) maps onto one point (denoted as *cross*) in the state space. *Dashed curve/arrow* represents state space trajectory. d Spotted DNA microarray for measuring gene expression (mRNA) profiles, representing a network state *S*(*t*) and e the associated "GEDI map" visualization of gene expression pattern. The map is obtained using the program GEDI [57,83]. This program places genes that behave similarly (with respect to their expression in a set of microarray measurements) onto the same pixel (= minicluster of genes) of the map. Similar miniclusters are arranged in nearby pixels in the two-dimensional picture of an *n* × *m* array of pixels. The assignment of genes to a pixel is achieved by a self-organizing map (SOM) algorithm. The *color of each pixel* represents the centroid gene expression level (mRNA abundance) of the genes in the minicluster. For a stack of GEDI maps, all genes are forced to be assigned to the same pixels in the different maps, hence the global coherent patterns of each GEDI map allows for a one-glance 'holistic' comparison of gene expression profiles of different conditions *S*(*t*) (tissue types, time points in a trajectory, etc). f Scheme of branching development of a subset of four different blood cells with their distinct GEDI maps, starting from the multipotent *CMP* = common myeloid progenitor cell [150]. *MEP* = Megakaryocte-Erythorid progenitor cell; *GMP* = Granulocyte-Monocyte progenitor

teins that regulate the expression of genes by binding to the promoter region of the 'target genes' (Fig. 1a). Another few hundred loci in the genome encode microRNAs (miRNA), which are transcribed to RNA but do not encode proteins. miRNAs regulate the expression of genes by interfering with the mRNA of their target genes based on sequence complementarity (for reviews see [41,89,141]). Thus, in this discussion we can assume a regulatory network of around 3000 regulator genes rather than the 25,000 genes of the genome. The genes that do not encode TFs are the "effector genes", encoding the work horses of the cell, including metabolic enzymes and structural genes. In our approximation, we also assume that the effector genes do not directly control other genes (although they may have global effects such as change of pH that may affect the expression of some genes). Further we do not count genes encoding proteins of the signal transduction machinery since they act to mediate external signals (such as hormones) that can be viewed as perturbations to the network and can in a first approximation be left out when discussing cell-intrinsic processes. Thus, the directed graph describing the genome-wide GRN network has a "medusa" structure [112], with a core set of regulators (medusa head) and a periphery of regulated genes (arms) which in a first approximation do not feedback to the core.

The Core GRN as a Graph That Governs Cell Phenotype

The next question is: do the 3000 core regulatory genes form a *connected* graph or rather independent (detached) "modules"? The idea of modularity would have justified the classical paradigm of independent causative pathways and has in fact actively been promoted in an attempt to mitigate the discouragement in view of the unfathomable complexity of the genome [85]. While a systematic survey that would provide a precise number is still not available, we can, based on patchy knowledge from the study of individual TFs, safely assume that a substantial fraction of TFs control the expression of more than one other TF. Many of them also control entire batteries [50] of effector genes, while perhaps a third subset may be specialized in regulating solely the effector genes. In any case, the core regulatory network of ~3000 nodes controls directly or indirectly the entire gene expression profile of 25,000 genes, and hence, the cell phenotype.

Then, assuming that on a average each TF controls at least two (typically more) other TFs [128,188], and considering statistical properties of random evolved networks (graphs) [36], we can safely assume that the core transcriptional network is a connected graph or at least its giant component (largest connected subgraph) covers the vast majority of its nodes. This appears to be the case in GRNs of simple organisms for which more data are available for parts of the network, as in yeast [128], *C. elegans* [54]; sea urchin [50] or, for even more limited subnetworks in mammalian systems [32,183] although studies focused on selected subnetworks may be subjected to investigation bias. However, recent analysis of DNA binding by 'master transcription factors' in mammalian cells using systemic (hence in principle unbiased) chromatin immunoprecipitation techniques [177] show that they typically bind to hundreds if not thousands of target genes [48, 70,105,155], strongly suggesting a global interconnectedness of the GRN. This however, does not exclude the possibility that the genome-wide network may exhibit some "modularity" in terms of weakly connected modules that are locally densely connected [166].

Structure of This Article

The goal of this article is to present both to biologists and physicists a set of basic concepts for understanding how the maps of thousands of interacting genes that systems biology researchers are currently assembling, ultimately control cell phenotypes. In Sect. "Overview: Studies of Networks in Systems Biology" we present an overview to experimental biologists on the history of the analysis of networks in systems biology. In Sect. "Network Architecture" we briefly discuss core issues in studies of network *topology*, before explaining basic ideas of network *dynamics* in Sect. "Network Dynamics" based on two-gene circuits. The central concepts of multi-stability will be explained to biologists, assuming a basic calculus background. Waddington's *epigenetic landscape* will also be addressed in this conjunction. In Sect. "Cell Fates, Cell Types: Terminology and Concepts" we discuss the actual "*biological observable*" that the gene regulatory network controls by introducing to non-biologists central concepts of cell fate regulation. Section "History of Explaning Cell Types" offers a historical overview of various explanations for metazoan cell type diversity, including dynamical, network-based concepts, as well as more 'reductionist' explanations that still prevail in current mainstream biology. Here the formal link between network concepts and Waddington's landscape metaphor will be presented. We then turn from small gene circuits to large, complex networks, and in Sect. "Boolean Networks as Model for Complex GRNs" we introduce the model of random Boolean networks which, despite their simplicity, have provided a useful conceptual framework and paved the way to learning what an integrative understanding of global network dynamics would look like, leading to the first central hypothesis: *Cell types may be high-dimensional attractors of the complex gene regulatory network*. In Sect. "Three Regimes of Behaviors for Boolean Networks" the more fundamental dynamical properties of ordered, critical and chaotic behavior are discussed, leading to the second hypothesis: *Networks that control living cells may be in the critical regime*. In Sect. "Experimental Evidence from Systems Biology" we summarize current experimental findings that lend initial support to these ideas, and in Sect. "Future Directions and Questions" we conclude with an outlook on how these general concepts may impact future biology.

Overview: Studies of Networks in Systems Biology

Cellular networks of biological molecules currently studied by "systems biologists" encompass three large classes: metabolic, protein–protein interaction and gene regulatory (transcriptional) networks (GRN). The formalization of these systems into networks is often taken for granted but is an important issue. It is noteworthy that metabolic network diagrams [193] represent physical networks in that there is a *flow* of information or energy in the links, and thus, the often used metaphors inspired by man-made transport or communication networks ('bottleneck', 'hubs' etc.) are more appropriate than in the other two classes. Moreover, metabolic networks are subjected to the constraint of mass preservation at each node – in compliance with Lavoisier's mass conservation in chemistry. In contrast to such flow networks, protein–protein interaction networks and GRN are abstract notations of "influence networks" in which the nodes influence the behavior of other nodes directly or indirectly – as represented by the links of the graph. There is no actual flow of matter in the links (although of course information is exchanged) and there is no obvious physical law that constrains the architecture, thus allowing for much richer variations of the network structure. The links are hence abstract entities, representing *potential* interactions assembled from independent observations that may not coexist in a particular situation. Thus, in this case the network is rather a convenient graphical representation of a collection of potential interactions.

Studies of such biomolecular influence networks can fundamentally be divided between two levels (Table 1): (A) network architecture and (B) network dynamics (system behavior).

The study of network architecture in molecular biology can further be divided into (A1) efforts to *determine* the actual graph structure that represent the specific interactions of the genes for particular instance (a species) and (A2) the more recent *analysis* of its structure, e. g., for interesting topological features [2]. The determination of the graph structure of the genome-wide GRNs in turn is either achieved (*i*) by direct *experimental demonstration* of the physical interactions, which has in the past decade greatly benefited from novel massively parallel technologies, such as chromatin IP, promoter one-hybrid or protein binding DNA microarrays [35,54,177], or (*ii*) via theoretical *inference* based on observed correlations in gene expression behavior from genome-wide gene expression profiling experiments. Such correlations are the consequence of mutual dependencies of expression in the influence network.

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Table 1
Overview of typical directions and levels in the study of gene regulatory networks

A. Network Architecture
A1. Determination of network architecture
▶ Experimental
▶ Theoretical = ***Network inference***
A2. Analysis of network architecture
Identification/characterization of "interesting" structural features
B. Network Dynamics
For small networks
B1. Modeling: theoretical prediction of behavior
of a real circuit based on (partially known, assumed) architecture + experimental verification
For complex networks (full architecture not known)
B2. Theoretical
Study of generic dynamics of simulated ensembles of networks of particular architecture classes
B3. Experimental
Measurement and analysis of high-dimensional state space trajectories of a real network

The next section provides a concise overview of the study of the network architecture, which due to limitations of available data is mostly an exercise in studying graph topology, and briefly discusses associated problems.

Network Architecture

Network Inference

An emerging challenge is the determination of entire networks, including the connection graph (topology), the direction of interactions ('arrows") and, ideally, the interaction modality and logical function, using systematic inference from genome-wide patterns of gene expression ("inverse problem"). Gene expression profiles (transcriptomes) represent a snap-shot of the mRNA levels in cell or tissue samples. The arrival of DNA microarrays for efficient measuring gene expression profiles for almost all the genes in the genome has stimulated the development of algorithms that address the daunting inverse problem, and the number of proposed algorithms has recently exploded. Most approaches concentrate on inference based on the more readily available static gene expression profiles, although time course microarray experiments, where time evolution of transcriptomes are monitored at close intervals would be advantageous, especially to infer the arrow direction of network links (directed graph) and the interaction functions.

However, because of uncertainties and the paucity of experimental data, systematic network inference faces formidable technical and formal challenges, and most theoretical work has been developed and tested based on simulated networks. A fundamental concern is that microarray based expression levels reflect the average expression of a population of millions of cells that – as recently demonstrated – exhibit vastly diverse behaviors even when they are clonal. Thus, the actual quantity used for inference is not a direct but a convoluted manifestation of network regulation (this issue is discussed in Sect. "Are Cell Types Attractors?"). Moreover, while mRNA levels reflect relatively well the activity status of the corresponding gene promoter, i. e., revealing the *regulated* activity, they are poor indicator of the *regulating* activity of a gene, because of the loose relationship between mRNA level and effective activity of the transcription factor that it encodes. Since the true network architecture ('gold standard') is not known, validation of the theoretical approaches remains unsatisfactory. Nevertheless, the recent availability of large numbers of gene expression profiles and the increasing (although not complete) coverage of gene regulation maps for single-cell organisms (notably E. coli and yeast) open the opportunity to directly study the mapping between gene expression profile and network structure [39,42,61,88,135,190].

Here we refer to [2,131,139] for a survey of inference methods and instead briefly discuss the study of network topology before we move on to network dynamics.

Analyzing Network Structures

Once the network topology is known, even if direction, modality and logics of links are not specified to offer the complete system architecture, it can be analyzed using graph theory tools for the presence of global or local structures (subgraphs) that are "interesting". A potentially interesting feature can be defined as a one which cannot be explained by chance occurrence in some randomized version of the graph (null model), i. e., which departs from what one would expect in a "random" network (see below). Most of the graph theoretical studies have been stimulated by the *protein–protein interaction networks* which have been available for some years, of which the best characterized is that of the yeast *S. cerevisiae* [29]. Such networks represent non-directed graphs, since the links between the nodes (proteins) have been determined by the identification of physical protein interactions (heterodimer or higher complex formation). Here we provide only a cursory overview of this exploding field, while focusing on conceptual issues.

A large array of structural network features, many of them inspired by the study of ecological and social networks [152,160,199] have been found in biomolecular networks. These features include global as well as local features, such as, to mention a few, the scale-free or broad-scale distribution of the connectivity k_i of the nodes i [3,10,20], betweenness of node i [107,208], hierarchical organization [164], modularity [102,121,133,166,206], assortativity [140], and enrichment for specific local topology motifs [147], etc. (for a review of this still expanding field see [2,21,29,152]).

The global property of a scale-free distribution of connectivity k_i, which has attracted most attention early on and quickly entered the vocabulary of the biologist, means that the probability $P(k_i)$ for an arbitrary node (gene) i in the network to have the connectivity k_i has the form $P(k_i) \sim k_i^{\gamma}$ where the characteristic constant γ is the power-law exponent – the slope of the line in a $P(k_i)$ vs. k_i double-logarithmic plot. This distribution implies that there is no characteristic scale, i. e., no stable average value of k: sampling of larger number of nodes N will lead to larger "average" k values. In other words, there is an "unexpectedly" high fraction of nodes which are highly connected ("hubs") while the majority exhibits low connectivity. This property has attracted as much interest as it has stirred controversy because of the connotation of "universality" of scale-freeness on the one hand, and several methodological concerns on the other hand [21,24,73,76,175,181,193]

The Problem of Choosing the Null Model In addition to well-articulated caveats due to incompleteness, bias and noise of the data [29,52], an important general methodological issue in the identification of structural features of interest, especially if conclusion on functionality and evolution is drawn, is the choice of the appropriate "null model" – or "negative control" in the lingo of experimentalists. A total random graph obviously is not an ideal null model, for on its background, any bias due to obvious constraints imposed by the physical reality of the molecular network will appear non-random and hence, be falsely labeled "interesting" even if one is not interested in the physical constraints but in evidence for functional organization [14,93]. For instance, the fact that gene duplication, a general mechanism of genome growth during evolution, will promote the generation of the network motif in which one regulator regulates two target genes, needs to be considered before a "purposeful" enrichment for such a motif due to increased fitness is assumed [93]. Similarly, the rewiring of artificial regulatory networks through reshuffling of *cis* and *trans* regions or the construction of networks based on promoter-sequence information content reveal constraints that lead to bias for particular structures in the absence of selection pressure [18,46]

The problem amounts to the practical question of which structural property of the network should be preserved when randomizing an observed graph to generate a null model. Arbitrarily constrained randomization based on preservation of some a priori graph properties [18,140] thus may not suffice. However, the question of which feature to keep cannot be satisfactorily answered given the lack of detailed knowledge of all physical processes underlying the genesis and evolution of the network. The more one knows of the latter, the more appropriate constraints can be applied to the randomization in order to expose novel biological feature that cannot be (trivially) explained by the neglected elementary construction constrains.

Evolution of Network Structure: Natural Selection or Natural Constraint?

This question goes beyond the technicality of choosing a null model for it reaches into the realm of a deeper questions of evolutionary biology: which features are inevitably linked to the very physical processes through which networks have grown in evolution and which arose due to natural selection because they confer a survival advantage [15,77,78,93,203]? Often this question is taken for granted and an all-mighty selection process is assumed that can produce any structure as long as it contributes sufficiently to fitness. This however, would require that during natural selection the random, mutation-driven reshuffling of nodes and connections has no constrains and that Darwinian evolution explores the entire space of possible architectures. Clearly this is not the case: physical constraints in the rearrangement of DNA (insertions, deletions, duplications, conversion, etc) [16,170,187] as well as graph theoretical considerations channel the possibilities for how one graph can transform into another [93,203]. For instance, growth of the network due to the increase of genome-size (gene number) by gene duplication can, without the invisible hand of selection, give rise to the widely-found scale-free structure [23,186] although it remains to be seen whether this mechanism (and not some more fundamental statistical process) accounts for the ubiquitous (near) scale-free topology. The fact that the scale-free (or at least, broad-scale [10]) architecture has dynamical consequences (see Sect. "Architectural Features of Large Networks and Their Dynamics") raises the question whether properties such as robustness may be inherent to some structure that is "self-organized", rather than sculpted by the invisible hand of natural selection. Thus, one certainly cannot simply argue that the scale-free structure has evolved "because of some functional advantage". Instead, it is reminded here that natural selection can benefit from spontaneous, self-organized structures [116]. This structuralist view [200], in which the physically most likely and not the biologically most functional is realized, need to be considered when analyzing the anatomy of networks [93].

In summary, the choice of the null model has to be made carefully and requires knowledge of the biochemical and physicochemical process that underlies genome evolution. This methodological caveat has its counterpart in the identification of "interesting" nucleotide sequence motifs when analyzing genome sequences [167].

Gene-Specific Information: More Functional Analysis Based on Topology

Beyond pure graph theoretical analysis, there have been attempts to link the topology with functional biological significance. For instance, one question is how the global graph structure changes (such as the size of the giant component) when nodes or links are randomly or selectively (e. g., hubs) removed. It should be noted that the term "robustness" used in such structural studies [5] refer to networks with the aforementioned connotation of transport or communication function with flow in the links and thus, differs fundamentally from robustness in a *dynamical* sense in the influence networks that we will discuss below.

Another approach towards connecting network topology with biological functionality is to employ bioinformatics and consider the *biological identity* of the genes or proteins represented by the nodes. Then one can ask whether some graph-related node properties (e. g., degree of connectivity, betweenness/centrality, contribution to network entropy, etc.) are correlated with known biological properties of the genes, of which the most prosaic is the "essentiality" of the protein, derived from genetic studies [84, 137,207,208]. To mention just a few of the earlier studies of this expanding field, it has been suggested that proteins with large connectivity ("hubs") appear to be enriched for "essential proteins" [104], that hubs evolve slower [68], and that they tend to be older in evolution [58] – in accordance with the model of preferential attachment that generates the scale-free distribution of the connectivity $\underline{k}_i$. However, many of these findings have been contested for statistical or other reasons [27,28,67,106,107]. The conclusion of such functional bioinformatics analysis need to be re-examined when more reliable and complete data become available.

Network Dynamics

While the graph theoretical studies, including the ramifications into core questions of evolution outlined above, are interesting eo ipso, life is breathed into these networks only with the 'dynamics'. Understanding the system-level dynamics of networks is an essential step in our desire to map the static network diagrams which are merely "*anatomical* observables", into the *functional* observable of cell behavior. Dynamics is introduced by considering that a given gene (node) has a time-dependent expression value $x_i(t)$ which in a first approximation represents the activity state of gene *i* (an active gene is expressed and post-translationally activated). Instead of seeing dynamics as a sequence of gene activation, epitomizing the chain of causation of the gene-centric view (as outlined in Sect. "Introduction") a goal in the study of complex systems is to understand the integrated, "emergent" behavior of systems as a holistic entity. We thus define a system (network) state $S(t)$ as $S(t) = [x_1(t), x_2(t), \ldots, x_N(t)]$ which is determined by the expression values x_i of all the N genes of a network which change over time. It is obvious that not all theoretically possible state configurations S can be realized since the individual variables x_i cannot change independently. This global constraint on S imposed by the gene regulatory interactions ultimately determines the "system behavior" which is equivalent to whole-cell phenotype changes, such as switching between cell phenotypes. Thus, one key question is: given a particular network architecture, what is the dynamics of S, and does it reproduce typical cell behaviors, or even predict the particular behavior of a specific cell? Before plunging into complex GRNs, let us introduce basic concepts of modeling dynamics using small circuits (Fig. 2, B1 in Table 1.).

Small Gene Circuits

Basic Formalism The dynamics of the network can be written as a system of ordinary differential equations (ODE) that describe the rate of change of x_i as a function of the state of the all the x_j:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(\mathbf{x}) \,, \tag{1}$$

where $\mathbf{x}$ is the state vector $\mathbf{x}(t) = [x_1, x_2, \ldots, x_N]$ for a network of N genes, and $\mathbf{F}$ describes the interactions (including the interaction matrix), defining how the components influence each other. Concretely, for individual genes in small circuits, e. g. of two genes, $N = 2$:

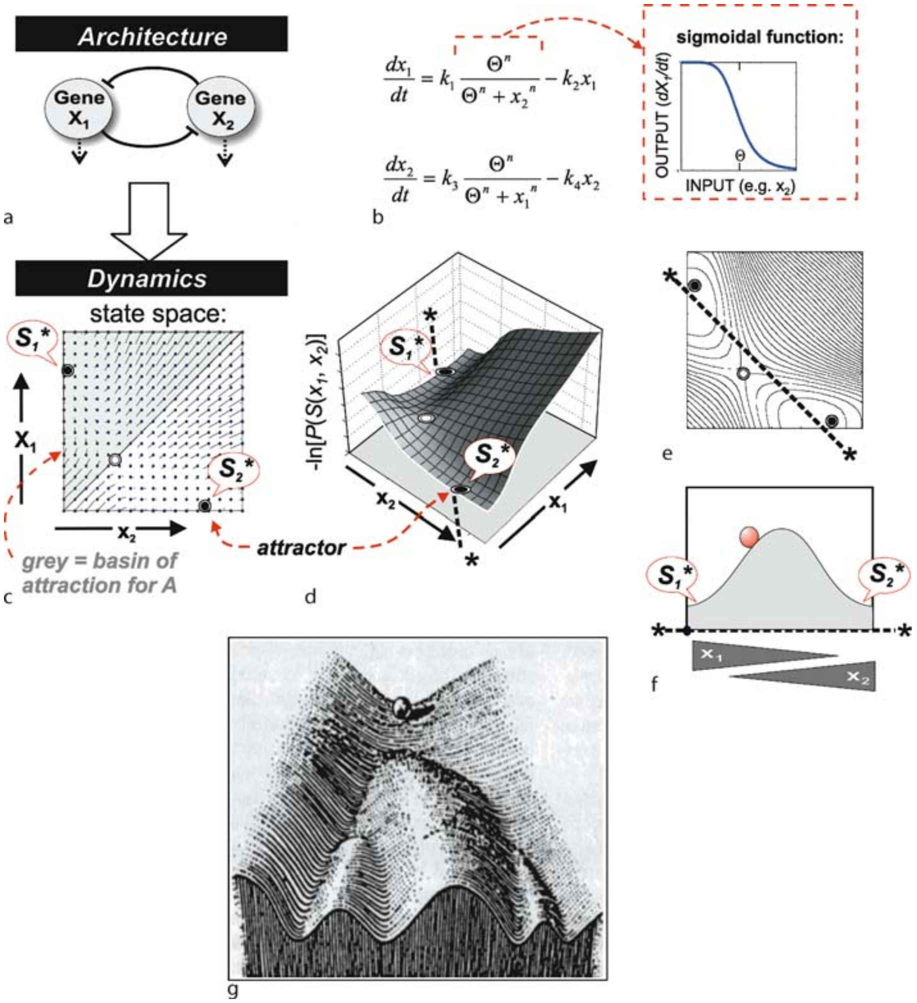
$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= f_1(x_1, x_2) \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= f_2(x_1, x_2) \,. \end{aligned} \tag{1a}$$

Here the functions f_i are part of the network "architecture" that determine how the inputs onto gene *i* jointly determine its dynamical behavior – as defined in Sect. "Introduction". An example of f is shown in Fig. 2b. Its form is further specified by *system parameters* that are constants for the time period of observation. The solution of these system equations is the movement of the state vector $\mathbf{x}$ in the N-dimensional ***gene expression state space*** spanned by x_i.

$$\mathbf{x}(t) = S(t) = [x_1, x_2, \ldots, x_N] \,. \tag{2}$$

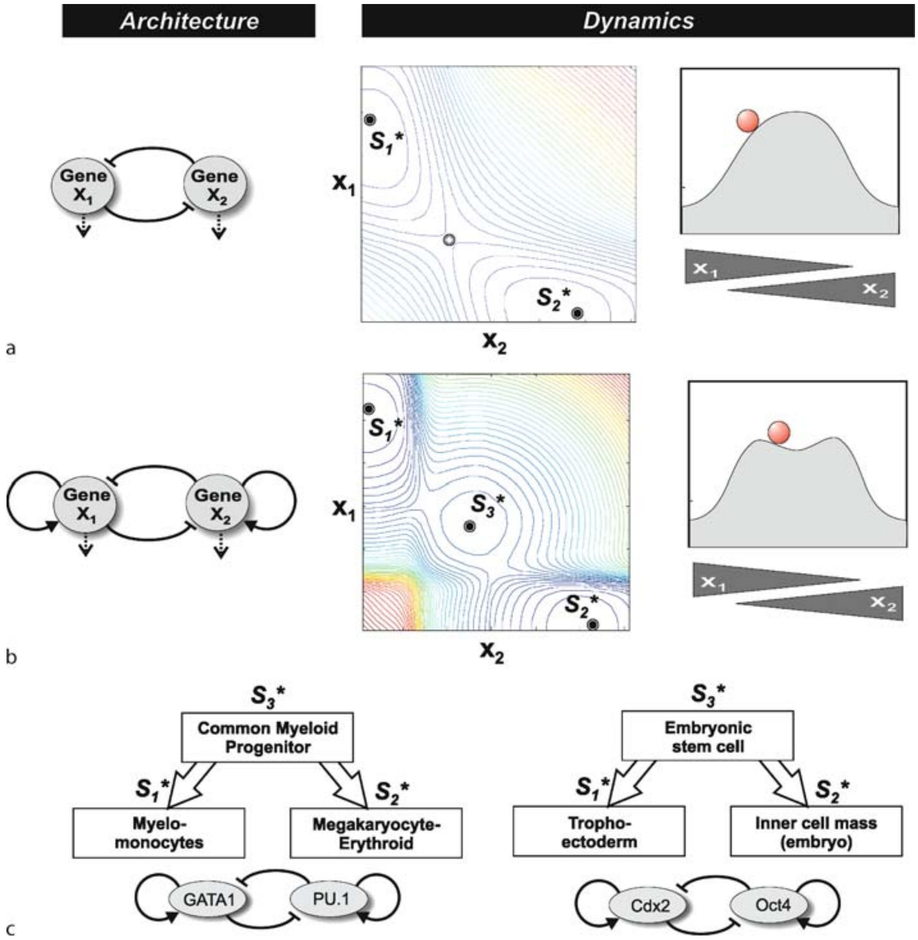
The Network State For a larger number of genes, it is convenient to describe the dynamics of the network with the vector $\mathbf{x}$ which roughly represents the system state S, Eq. (2), that biologists measure using DNA microarrays and is known as the "*gene expression profile*" (see Sect. "Definition of the Subject", Fig. 1). We will thus refer to S instead of the vector x in discussing biological system states. Since Eq. (1) is a first order differential equation, S specifies a system state at time t which for $t = 0$ represents an ***initial condition***. In other words, unlike in macroscopic mechanical systems with inertia, in which the velocity $\mathrm{d}x_i/\mathrm{d}t$ of the "particles" is also important to specify an initial state, in cells a photographic snapshot of all the positions of x_i, that is $S(t)$ itself, specifies the system state. (This will become important for experimental monitoring of network dynamics).

Role of Dynamical Models in Biology Small networks, or more precisely, "*circuits*" of a handful of interacting proteins and genes, have long been the object of modeling efforts in cell biology that use ODEs of the type of (1a) to model the behavior of the circuit [109,192] (Table 1, B1). In contrast, the objects of interest in the study of complex system sciences are large, i. e., "complex" networks of thousands of nodes, such as the 3000-node core GRN, mentioned earlier, and has largely been driven by


Architecture
Gene X1
Gene X2
a
Dynamics
state space:
S1*
X1
S2*
X2
grey = basin of attraction for A
c
$\frac{dx_1}{dt} = k_1 \frac{\Theta^n}{\Theta^n + x_2^n} - k_2 x_1$
$\frac{dx_2}{dt} = k_3 \frac{\Theta^n}{\Theta^n + x_1^n} - k_4 x_2$
b
sigmoidal function:
OUTPUT (dX1/dt)
Θ
INPUT (e.g. x2)
-ln[P(S(x1, x2))]
S1*
S2*
*
x2
x1
attractor
d
e
S1*
S2*
x1
x2
f
g

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 2

From gene circuit architecture to dynamics, exemplified on a bistable two-gene circuit. a Circuit architecture: two mutually inhibitory genes X_1 and X_2 which are expressed at constant rate and inactivated with first-order kinetics ("bistable toggle switch"). b Typical ODE system equations for the circuit in a (see [71,109]). c State space (x_1–x_2-phase-plane). *Each dot* is an example initial state $S_o = [x_1(t = 0), x_2(t = 0)]$ with the emanating tick revealing direction and extent that the state S_o would travel in the next time unit Δt. *Solid circles*, S_1^* and S_2^*, denote stable fixed-points ("attractors"). *Empty circle* denotes unstable fixed-point (saddle-node). d, e, f Various schematic representations of the probability $P(S)$ (for a noisy circuit) to find the system in state $S = [x_1, x_2]$. The "elevation" (z-axis over the x_1–x_2 plane) is calculated as – $\ln(P)$; thus, the most probable = stable states are the lowest in the emerging landscape. g Waddington's metaphoric epigenetic landscape, in the 1957 version [198]



Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 3

Expansion of the bistable two-gene circuit by self-stimulation – creating a central metastable attractor. a The bistable circuit of Fig. 2 and associated epigenetic landscape, now as *contour graph*. b If the two genes exert self-activation, then, for a large range of parameters and additive effect of the two inputs that each gene receives, the epigenetic landscape is altered such that the central unstable fixed-point in a. becomes locally stable, S_3^*, giving rise to robust "tristable" dynamics [99]. c Two specific examples of observed gene circuits that represent the circuit of *C*. Such network motifs are typically found to regulate the branch points where multipotent progenitor cells make binary cell fate decisions [99,162]. The metastable central attractor S_3^* can be modeled as representing the metastable bi-potent progenitor cell which is poised in a state of locally stable indeterminacy between the two prospective fate attractors S_1^* and S_2^*

experimental monitoring of $S(t)$ using microarrays since the function $\boldsymbol{F}$ that represents the architecture of the network is not known (Table 1, B3).

What is the actual aim of modeling a biological network? In mathematical modeling of small gene circuits or of signal transduction pathways, one typically predicts the temporal evolution (time course) of the concentration of the modeled variables $x_1(t)$, or $x_2(t)$, and characterizes critical points, such as stable steady states or oscillations in the low-dimensional state space, e. g. in the $x_1 - x_2$ phase plane, after solving the system equations, as exemplified in Figs. 2 and 3. Unknown parameter values have to be estimated, either based on previous reports or obtained by fitting the modeled $x_i(t)$ to the observed time course. Successful prediction of behaviors also serves to *validate* the assumed circuit architecture. From the ***stability analysis*** of the resulting behavior [151], generalization as to the ***dynamical robustness*** (low sensitivity to noise that cause $x(t)$ to randomly fluctuate) and ***structural robustness*** (preservation of similar dynamical behavior even if network architecture is slightly rewired by mutations) can be made [7,34]. Both types of robustness of the system shall not be confounded with the robustness which is sometimes encountered in the analysis of network topology where preservation of some graph properties (such as global connectedness) upon deletion of nodes or links is examined ("error tolerance") (as discussed in Sect. "Network Architecture").

Although in reality most genes and proteins are embedded in the larger, genome-wide network, small, idealized circuit models which implicitly accept the absence of many unknown links as well as inputs from the nodes in the global network outside the considered circuit often surprisingly well predict the observed kinetics of the variables of the circuit. This not only points to intrinsic structural robustness of the class of natural circuits, in that circuit architectures slightly different from that of the real system can generate the observed dynamics. But it also suggests that obviously, for some not well-understood reasons, local circuits are to some extent dynamically insulated from the larger network in which they are embedded although they are topologically not detached.

Such "functional modularity" may be a property of the particular architecture of the evolved complex GRNs. In fact, in models of complex networks (discussed in the next section), evidence for just such "functional modularity" has been found with respect to network dynamics [116]. In brief preview (see Sect. "Three Regimes of Behaviors for Boolean Networks"), work on random Boolean network models of GRN has revealed three "regimes" of dynamical behaviors: ordered, critical, and chaotic, as described in Sect. "Ordered and Chaotic Regimes of Network Behavior". In the ordered and critical regimes, many or most genes become "frozen" into "active" or "inactive" states, leaving behind *functionally isolated "islands" of genes* which, despite being connected to other islands through the interactions of the network, are free to vary their activities without impacting the behaviors of other functionally isolated islands of genes.

Complex Networks

Despite a recent spate of publications in which the temporal change of expression of individual genes have been predicted based on small circuit diagrams, such predictions do not provide understanding of the integrated cell behavior, such as the change of cell phenotype that may involve thousands genes in the core GRN. Thus, analysis of genome-wide network state is needed to understand the biological observable. In a first approximation, it is plausible that the state of a cell, such as the particular cell (pheno-)type in a multi-cellular organism, is defined by its genome-wide gene expression profile, or *transcriptome* $T = [x_1, x_2, \ldots, x_N]$ with $N = 25,000$. In fact, microarray analysis of various cells and tissues reveals globally distinct, tissue specific patterns of gene expression profiles that can easily be discerned as shown in Fig. 1f. As mentioned above (Sect. "Definition of the Subject"), the gene expression profile across the genome is governed by the core regulatory network of transcription factors (TFs) which enslave the rest of the genome. Thus, in our approximation the network state S of the core transcriptional network of 3000 or so genes essentially controls the entire (genome-wide) gene expression profile.

For clarity of formalization, it is important to note that one genome in principle encodes exactly one fixed network, since the network connections are defined by the specific molecular interactions between the protein structure of TFs and the DNA sequence motif of the cis-regulatory promoter elements they recognize. Both are encoded by the genomic sequence. The often encountered notion that "networks change during development" and that "every cell type has its own network" is in this strict formalism incorrect – the genes absent in one cell type must directly or indirectly have been repressed (and sometimes, continuously kept repressed) by other genes that are expressed. Thus, the genome (of one species) directly maps into one (time-invariant) network architecture which in turn can generate a variety of states $S(t)$. It is the state S that changes in time and is distinct in different cell types or in different functional states within one cell type. Only genetic mutations in the genome will "rewire" the network and change its architecture.

If the network state $S(t)$ of the core GRN directly maps into a state of a cell (the biological observable), then one question is: What is the nature of the integrated dynamics of S in complex, irregularly wired GRN and is it compatible with observable cell behavior?

The study of the dynamics of complex networks (Table 1, B) may at first glance appear to be impeded by our almost complete ignorance about the architecture and interaction modalities between the genes. However, we cannot simply extrapolate from the mindset of studying small circuits with rate equations to the analysis of large networks. This is not only impossible due to the lack of information about the detailed structure of the entire network – the function **F** in Eq. 1 is unknown – but it may also be numerically hard to do. Yet, despite our ignorance about the network architecture, much can be learned if we reset our focus plane to the larger picture of the network. In this regard, the study of the dynamics of complex network can have two distinct goals (see Table 1).

One line of research (Table 1, B2) overcomes the lack of information about the network architecture by taking an ensemble approach [112] and asks: *What is the typical behavior of a statistical ensemble of networks that is* *characterized by an architectural feature (e. g. average connectivity k_i, power-law exponent γ)*? This computationally intense approach typically entails the use of discrete-valued gene networks (Sect. "Boolean Networks as Model for Complex GRNs"). As will be discussed below, such analysis has led to the definition of three broad classes of behaviors: chaotic, critical and ordered.

The second approach (Table 1, B3) to the dynamics of complex GRN is closer to experimental biology and exploits the availability of gene expression profile measurements using DNA microarray technology. Such measurements provide snapshots of the state of the network $S(t)$ over N genes, covering almost the entire transcriptome, and thus, reveals the direct output of the GRN as a distributed control system. Monitoring $S(t)$ and its change in time during biological processes at the whole-cell level will reveal the constraint imposed on the dynamics of $S(t)$ by network interactions and can, in addition to providing data for the inference problem (Sect. "Network Inference"), expose particular dynamical properties of the transcriptome that can be correlated with the biological observable.

Cell Fates, Cell types: Terminology and Concepts

In order to appreciate the meaning of the network state S and how it maps to the biological observable, we will now present (to non-biologists) in more detail the most prosaic biological observable of gene network dynamics: cell fate determination during development.

Stem Cells and Progenitor Cells in Multi-cellular Organism

A hall-mark of multi-cellular organisms is the differentiation of the omnipotent zygote (fertilized egg) via *totipotent* and *pluripotent embryonic stem cells* and *multipotent tissue stem cells* into functionally distinct mature "cell types" of the adult body, such as red blood cell, skin cells, nerve cells, liver cells, etc. This is achieved through a branching (tree-like) scheme of successive specialization into lineages. If the fertilized egg is represented by the main trunk of the "tree of development", then think of cells at the branching points of developmental paths as the stem cells. One example of a multipotent stem cell is the hematopoietic stem cells (HSC) which is capable of differentiating into the entire palette of blood cells, such as red blood cells, platelets and the variety of white blood cells. The last branch points represent *progenitor* cells which have a lesser developmental potential but still can chose between a few cell types (e. g., the common granulocyte-macrophage progenitor (GMP) cell, Fig. 1f). Finally, the outmost branches of the tree represent the mature, terminally differentiated cell types of the body.

A cell that can branch into various lineages is said to be "multipotent". It is a "*stem cell*" when it has the potential to self-renew by cell division, maintaining its differentiation potential, and to create the large family of cells of an entire tissue (e. g., the hematopoietic stem cell). Thus, progenitor cells, which proliferate but cannot infinitively self-renew, are strictly not stem cells. The commitment to a particular cell phenotype (a next generation branch of the tree) is also referred to as a "cell fate" since the cell at the proximal branching point is "fated" to commit to one of its prospective cell types.

Development and Differentiation

The diversification of the embryonic stem cell to yield the spectrum of the thousands [94] or so cell types in the body occurs in a process of successive branching events, at which multipotent cells commit to one fate and which appear to be binary in nature. Thus, multipotent cells make an either-or decision between typically two lineages – although more complex schemes have been proposed [65]. Moreover, it is generally assumed that during natural development there is only ***diversification*** of developmental paths but no *confluence* from different lineages, although recently exceptions to this rule have been reported for the hematopoietic system [1].

As cells develop towards the outer branches of the "tree of development", they become more and more specialized and progressively lose their competence to proliferate and diversify ("potency"). They also develop the phenotype features of a mature cell type; for instance, in the case of red blood cells, they adopt the flat, donut-like shape and synthesize haemoglobin. This process is called ***differentiation***. Most cells then also loose their capacity to divide, that is, to enter into the proliferative state. Thus, mature, terminally differentiated cells are typically quiescent or "*post mitotic*".

The branching scheme of cell types imposes another fundamental property: cell types are *discrete* entities. They are distinct from each other, i. e., they are well separated in the "phenotype space" and are stable [171]. There is thus no continuum of phenotype. As Waddington, a prominent embryologist of the last century, recognized in the 1940s: Cell types are "well-recognisable" entities and "intermediates are rare" [198]. In addition to the (quasi-) discreteness between branches of the same level, discreteness between the stages within one developmental path is also apparent: a multipotent stem cell at a branching point is in a discrete stage and can be identified based on molecular markers, isolated and cultured as such. Hence, a "stem cell" is not just a snapshot of an intermediate stage within a continuous process of development, but a discrete metastable entity.

The flow down the developmental paths, from a stem cell to the terminally differentiated state is, despite the pauses at the various metastable stem and progenitor cell levels, essentially unidirectional and, with a few exceptions, irreversible. In some tissues, the mature cells, as is the case with liver, pancreas or endothelial cells, can upon injury revert to a phenotype similar to that of the last immature (progenitor) stage and resume proliferation to restore the lost cell population, upon which they return to the differentiated, quiescent state.

Biologists often speak in somewhat loose manner of cells "***switching***" their phenotype. This may refer to switching from a progenitor state to a terminally differentiated state (differentiation) or within a progenitor state from the quiescent to other functional states, such as the proliferative or the apoptotic state (apoptosis = programmed cell death). In any case, such intra-lineage switching between different functional states also represents discontinuous, quasi-discrete transitions of whole-cell behaviors. The balance between division, differentiation and death in the progenitor or stem cell compartment of a tissue thus consists of state transitions that entail all-or-none decisions. This balance is at the core of organismal development and tissue homeostasis.

Now we can come back to the network formalism: if the network state S maps directly into the biological observable, what are the properties of the network architecture that confer its ability to produce the properties of the biological observable outlined in this section: discreteness and robustness of cell types, discontinuity of transitions, successive binary diversification and directionality of these processes? Addressing these questions is the long-term goal of a theory of the multicellular organism. In the following we describe the status of research toward this goal.

History of Explaning Cell Types

Waddington's Epigenetic Landscape and Bistable Genetic Circuits

One of the earliest conceptualization of the existence of discreteness of cell types was the work of C. Waddington, who proposed "epigenetic regulation" in the 1940s, an idea that culminated in the famous figure of the "epigenetic landscape" (Fig. 2g). This metaphor, devoid of any formal basis, let alone relationship to gene regulation, captures the basic properties of discrete entities and the instability of intermediates. The term 'epigenetic' was coined by Waddington to describe distinct biological traits that arise from the interplay of the same set of genes and does not require the assumption of a distinct, "causal" gene to be explained [196]. The 1957 version of the epigenetic landscape [198] (Fig. 2g) also implies that multipotent cells are destined to make either-or decisions between two prospective lineages, as embodied in the "watershed" regions (Fig. 2g).

Almost at the same time as Waddington, in 1948 Max Delbrück proposed a generic concept of differentiation into two discrete states in a biochemical system that can be described by equations of the form (3), consisting of two mutual inhibiting metabolites, $x_1(t)$ and $x_2(t)$ that exhibits bistability [53] (For a detailed qualitative explanation, see Fig. 2). The dynamics of such a system is graphically represented in the two dimensional state space spanned by x_1 and x_2 (Fig. 2c-f). ***Bistable dynamics*** implies that there are two steady states S_1^* and S_2^* that satisfy $\mathrm{d}x_1/\mathrm{d}t = \mathrm{d}x_2/\mathrm{d}t = 0$ and are *stable* fixed-points of the system. For the system equations of Fig. 2b, this behavior is observed for a large range of parameters. In a nutshell, the mutual inhibition renders the balanced steady state $S_3^*(x_1 = x_2)$ unstable so that the system settles down in either the steady state $S_1^*(x_1 \gg x_2)$ or $S_2^*(x_1 \ll x_2)$ when kicked out of this unstable fixed point. These two stable steady states and their associated gene activity patterns are discretely separated in the $x_1 - x_2$ state space and have been postulated to represent the differentiated state of cells, thus, corresponding to Waddington's valleys. This was the first conceptualization of a cellular differentiated state as a stable fixed-point of a non-linear dynamical system. Soon after Monod and Jacob discovered the principle of gene regulation they also proposed a circuit of the same architecture as Delbrück [148], but consisting of two mutually suppressing genes instead of metabolites, to explain differentiation in bacteria as a bistable system.

Bistability, Tristability and Multistability

The central idea of bistability is that the very same system, in this case, a gene circuit composed of two genes x_1 and x_2, can under some conditions produce two distinct stable states separated by an unstable stationary-state, and hence, it can switch in almost discontinuous manner from one state (S_1^*) to another (S_2^*) ("toggle switch") (Fig. 2). Bistability is a special (the simplest) case of ***multi-stability***, an elementary (but not necessary) property of systems with non-linear interactions, such as those underlying the gene regulatory influences as detailed in Fig. 2b. Which stable state (in this case, S_1^* or S_2^*) a network occupies depends on the ***initial condition***, the position of $S(t=0)$ in the $x_1 - x_2$ state space (Fig. 2c). The two stable steady-states, S_1^* or S_2^*, are called "***attractors***" since the system in such states will return to the characteristic gene expression pattern in response to small ***perturbations*** to the system by enforced change of the values of x_1 or x_2. Similarly, after an external influence that places the network at an unstable initial state $S'(t=0)$ within the ***basin of attraction*** of attractor S_1^* (gray area in Fig. 2c), and hence cause the network to settle down in the stable state S_1^*, the network will stay there even after the causative influence that initially put it in the state S' has disappeared. Thus, attractor states confer memory of the network state. In contrast, larger perturbations beyond a certain threshold will trigger a transition from one attractor to another, thus explaining the observed discontinuous "switch-like" transitions between two stable states in a continuous system.

After Delbrück, Monod and Jacob, numerous theoretical [74] and, with the rise of systems biology, experimental works have further explored similar simple circuitries that produce bistable behavior (see ref. in [38]). Such circuits and the predicted switch-like behavior have been found in various biological systems, including in gene regulatory circuits in Escherichia coli [157], mammalian cell differentiation [43,99,127,168] as well as in protein signal transduction modules [12,64,205]. Artificial gene regulatory circuits have been constructed using simple recombinant DNA technology to verify model predictions [22,71, 123].

Circuit analyzes have been expanded to cover more complex circuits in mammalian development. It appears that there is a common theme in the circuits that govern cell differentiation (Sect. "Cell Fates, Cell Types: Terminology and Concepts"): interconnected pairs of mutually regulating genes that are often also self-regulatory, as shown in Fig. 3b [43,99,168]. Such circuit diagrams may be crucial in controlling the binary diversification at developmental branch points (Fig. 1f) [99]. In the case where the two mutually inhibiting genes are also are self-stimulatory, as summarized in Fig. 3b, an interesting modification of the bistable dynamics can be obtained. Assuming independent (additive) influence of the self-stimulatory and cross-inhibitory inputs, this circuit will convert the central unstable state (saddle) S_3^* (Fig. 2, 3a) that separate the two stable steady-states into a stable steady state, thus generating ***tristable*** behavior [99]. The third, central stable fixed point has, in symmetrical cases, the gene expression configuration $S_3^*[x_1 \sim x_2]$ (Fig. 3b).

The *promiscuous* expression of intermediate-low levels of x_1 *and* x_2 *in the locally stable state* S_3^* has been associated with a stem or progenitor cell that can differentiate into the cell represented by the attractors $S_1^*(x_1 \gg x_2)$ or $S_2^*(x_1 \ll x_2)$. In fact, the common progenitor cells (Fig. 1f) have been shown to express "promiscuous" expression of genes thought to be specific for either of the lineages in which it will have to commit to, so-to speak providing a "preview" or "multi-lineage priming" of the gene expression of its prospective cell fates [47,59,92]. For instance, the *common myeloid progenitor* (CMP, Fig. 1) which can commit to either the *erythroid* or the *myeloid* lineage, expresses both the erythroid-specific transcription factor GATA1 (= x_1) and the myeloid specific transcription factor PU.1 (= x_2) at intermediate levels (Fig. 3c). The metastable [GATA1~PU.1] configuration generates a state of *indeterminacy* or "suspense" [146] awaiting the signal, be it a biological instruction or stochastic perturbation, that resolves it to become either one of the more stable states [GATA1 $\gg$ PU.1] or [GATA1 $\ll$ PU.1] when cells differentiate into the erythroid or myeloid lineage, respectively. Thus, multi-potency and indeterminacy of a progenitor cell can be defined purely dynamically and does not require a much sought after "stemness" gene (again, a concept that arose from the gene-centered view) [195]. The metastable state also captures the notion of a "higher potential energy" [79] that drives development and hence, may account for the arrow of time in ontogenesis.

A Formalism for Waddington's Epigenetic Landscape Metaphor

Obviously, the valleys in Waddington's epigenetic landscape represent the stable steady states (attractors) while the hill-tops embody the unstable fixed-points, as shown in Fig. 2. How can Waddington's metaphor formally be linked to gene network dynamics and the state space structure? For a one dimensional system $\mathrm{d}x/\mathrm{d}t = f(x)$, this is easily shown with the following analogy: An "energy" landscape represents the cumulative "work" performed/received when "walking" in the state space against / along the vector field of the "force" $f(x)$ (Fig. 2c). Thus, the "potential energy" is obtained by integrating $f(x)$ (the right-hand side of the system equation, Eq. (1)) over the state space variables x.

$$V(x) = -\int f(x)\mathrm{d}x\,. \tag{3}$$

Here the state space dimension x is given the meaning of a physical space, as that pertaining to a landscape, and the integral $V(x)$ is the sum of the "forces" experienced by the network states $S(t) = x(t)$ over a path in x that drive $S(t)$ to the stable states (Fig. 2c). The negative sign establishes the notion of a potential in that the system loses energy as it moves towards the stable steady states which are in the valleys ("lowest energy"). Higher ($N > 1$) dimensional systems (Eq. 1) are in general non-integrable (unless there exists a continuously differentiable (potential) function $V(x_1, x_2, \dots, x_N)$ for which $f_1 dx_1 + f_2 dx_2 + \cdots + f_N dx_N \cdots = 0$ is the *exact total differential*, so that $-grad(V)=\mathbf{F}(\mathbf{x})$ with $\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_N(\mathbf{x})]^T$). Thus, there is in general no potential function that could represent a "landscape"; however, the idea of some sort of "potential landscape" is widely (and loosely) used, especially in the case of two-dimensional systems where the third dimension can be depicted as a cartographic elevation $V(x_1, x_2)$. An elevation function $V(x_1, x_2)$ can be obtained in stochastic systems where $\mathbf{x}(t)$ is subjected to random fluctuations due to gene expression noise [87]. Then $V(S)$ is related to the probability $P(S)$ to find the system in state $S = (x_1, x_2, \dots)$, e.g., $V(S) = -\ln(P(S)$ [118,177]. It should however be kept in mind that the "quasi-potential" V is not a true (conservative) potential, since the vector field is not a conservative field.

The above treatment of the dynamics of the gene regulatory circuit explains the valleys and hills of Waddington's landscape but still lacks the "directionality" of the overall flow on the landscape, depicted by Waddington as the slope from the back to the front of his landscape. (This *arrow or time of development* will briefly be discussed in the outlook Sect. "Future Directions and Questions").

In summary, the epigenetic landscape that Waddington proposed based on his careful observation of cell behavior and that he reshaped over decades [103,179,196,197,198], can now be given both a molecular biology correlate (the gene regulatory networks) and a formal framework (probability landscape of network states S). The landscape idea lies at the heart of the connection between molecular network topology and biological observable.

The Molecular Biology View: "Epigenetic" Marks of Chromatin Modification

Although it is intuitively plausible that stable steady states of the circuits of Delbrück and of Monod and Jacob may represent the stable differentiated state, this explanation of a biological observable in terms of a dynamical system was not popular in the community of experimental molecular biologists and was soon sidelined as molecular biology, and with it the gene-centered view, came to dominate biology. The success in identifying novel genes (and their mutated alleles) and the often straightforward explanation of a phenotype offered by the mere discovery of a (mutant) gene triggered a hunt for such "explanatory genes" to account for biological observables. Gene circuits had to give place to the one-gene-one trait concept, leading to an era in which a new gene-centered epistemological habit, sometimes referred to as "*genetic determinism*", prevailed [180]. Genetic determinism, a particular form of reductionism in biology, was to last for another fifty years after Delbrück's proposal of bistability. Only as the "low hanging fruits" of simple genotype-phenotype relationships seemed to have all been picked, and genome-wide gene expression measurements became possible was the path cleared for the rise of "system biology" and biocomplexity that we witness today.

In genetic determinism, macroscopic observables are reduced in a qualitative manner to the existence of genes or regulatory interactions which at best "form" a linear chain of events. Such "pathways", most lucidly manifest in the arrow-arrow schemes of modern cell biology papers, serve as a mechanistic explanation and satisfied the intellectual longing for causation. It is in light of this thinking that the molecular biologists' explanation of the phenomenon of cell type determination has to be seen. Absent a theory of cell fate diversification and in view of the stunning stability of cell types, a dogma thus came into existence according to which the type identity of cells, once committed to, is irreversibly carved in stone [161]. Rare transdifferentiation events (switch of cell lineages) were regarded as curiosities. The observation that cell types express type-specific genes was explained by the presence of cell-type specific transcription factors. For instance, red blood cells express haemoglobin because of the presence of GATA1 – a lineage specific transcription factor that not only promotes commitment to the erythroid lineage (as discussed above, Sect. "Bistability, Tristability and Multistability") but also controls haemoglobin expression [149]. Conversely, the absence of gene activity, for instance, non-expression of liver-specific genes in erythrocytes, was explained by the silencing of the not needed genes by covalent DNA methylation and histone modifications (methylation, acetylation, etc.) [13,65,69,117,122] which modify chromatin structure, and thereby, control the access of the transcription machinery to the regulatory sites on the DNA. But who controls the controller?

Chromatin modifications [117,122] are thought to confer discrete alterations of gene expression state that are stable and essentially irreversible. This idea of permanent marks on DNA represents the conceptual cousin of mutations (but without altering DNA sequence) and was thus in line with the spirit of genetic determinism. Accordingly, they were readily adopted as an explanation of the apparently irreversible cell type-specific gene inactivation and were given the attribute "epigenetic" to contrast them from the genetic changes that involve alteration of DNA sequences. But enzymes responsible for covalent DNA and chromatin modification are not gene-locus specific, leaving open how the cell type specific gene expression *pattern* is orchestrated.

It is important to mention here a disparity in the current usage of the term "epigenetics" [103]. In modern molecular biology, "epigenetics" is almost exclusively used to refer to DNA methylation and covalent histone modifications; this meaning is taken for granted even among authors who comment on the very usage of this term [13, 26,103], and are unaware that memory effects can arise purely dynamically without a distinct material substrate, as discussed in Sect. "Bistability, Tristability and Multistability". In contrast, biological physicists use "epigenetic" to describe precisely phenomena, such as multi-stability (Sect. "Bistability, Tristability and Multistability") that are found in non-linear systems – a usage that comes closer to Waddington's original metaphor for illustrating the discreteness and stability of cell types [197,198]. The use of Waddington's "epigenetic landscape" has recently seen a revival [75,165] in the modern literature in the context of chromatin modifications but remains loosely metaphoric and without a formal basis.

Rethinking the Histone Code

The idea of methylation and histone modifications as a second code beyond the DNA sequence ("histone code") that cells use to "freeze" their type-specific gene expression pattern relied on the belief that these covalent modifications act like permanent marks (only erased in the germline when an oocyte is formed). This picture is now beginning to change.

First, recent biochemical analysis suggest that the notion of a static, irreversible "histone code" is oversimplified, casting doubt on the view that histone modification is the molecular substrate of "epigenetic memory" [122,126,144,191]. With the accumulating characterization of chromatin modifying enzymes, notably those controlling histone lysine (de)methylation [122,126,144, 191], it is increasingly recognized that the covalent "epigenetic" modifications are bidirectional (reversible) and highly dynamic. Second, cell fate plasticity, most lucidly evident in the long-known but rarely observed transdifferentiation events, or in the artificial reprogramming of cells into embryonic stem cells either by nuclear transfer-mediated cloning [91] or genetic manipulation [143, 156,184,201], confirm that the "epigenetic" marks are reversible – given that the appropriate biochemical context is provided. If what was thought of as permanent molecular marks is actually dynamical and reversible – what then maintains lineage-specific gene expression patterns in an inheritable fashion across cell divisions?

In addition, as mentioned above, there is another, more fundamental question from a conceptual point of view: chromatin-based marking of gene expression status is a generic, not locus – specific process – the same enzymes can apply or remove the covalent marks on virtually any gene in the genome. They are dumb. So what smart system orchestrates the DNA methylation and histone modification machinery at the tens of thousands of loci in the genome so that the appropriate set of genes is (in)activated to generate the cell type-specific patterns of gene expression?

A system-level view avoids the conundrum caused by the mechanistic, proximal explanation [189] of the gene-centered view. A complex systems approach led to the idea that the genome-wide network of transcriptional regulation can under some conditions spontaneously orchestrate, thanks to self-organization, the establishment of lineage- specific gene expression profiles [114], as will be discussed below. In fact, the picture of chromatin modification as primum movens that operates "upstream" of the transcription factors (TFs), controlling their access to the regulatory elements in promoter regions, must be revised in light of a series of new observations. Evidence is accumulating that the controller itself is controlled – namely, by the TFs they are thought to control: TFs may actually take the initiative by recruiting the generic chromatin-modifying enzymes to their target loci [51,80,125, 144,145,182,185]. It is even possible that a mutual, cooperative dynamical interdependence between TFs and chromatin-modifying enzymes may establish locus-specific, switch-like behavior that commands "chromatin status" changes [56,132]. In fact, an equivalent of the indeterminacy state where both opposing lineage specific TFs balance each other (Sect. "Bistability, Tristability and Multistability", Fig. 3b) is found at the level of chromatin modification, in that some promoters exhibit "bivalent" histone modification in which activating and suppressing histone methylations coexist [25]. Such states in fact are associated with TFs expressed at low level – in agreement with the central attractor S_3^* of the tristable model (Fig. 3b). Thus, chromatin modification is at least in part "downstream" of TFs and may thus act to add additional stability to the dynamical states that arise from the network of transcriptional regulation. If correct, such a relationship will allow us to pass primary responsibility for establishing the observable gene expression patterns back to transcription factors. With its genome-wide regulatory connections the

GRN is predestined for the task of coordination and distributed information processing.

But under what conditions can a complex, apparently randomly wired network of 3000 regulators create, with such stunning reliability and accuracy, the thousands of distinct, stable, gene expression profiles that are associated with a meaningful biological state, such as a cell type? Studies of Boolean networks as toy models in the past 40 years have provided important insights.

Boolean Networks as Model for Complex GRNs

General Comment on Simplifying Models

The small-circuit model discussed in Sect. "Small Gene Circuits" represents arbitrarily cut-out fragments of the genome-wide regulatory network. A cell phenotype, however, is determined by the gene expression profile over thousands of genes. How can we study the entire network of 25,000 genes, or at least the core GRN with 3000 transcription factors even if most of the details of interactions remain elusive?

The use of random Boolean networks as a generic network model, independent of a specific GRN architecture of a particular species, was proposed by Kauffman in 1969 – at the time around which the very idea of small gene regulatory circuits were presented by Monod and Jacob [148]. In random Boolean networks the dynamics is implemented by assuming discrete-valued genes that are either ON (expressing the encoded protein) or OFF (silenced). The interaction function (Sect. "Introduction", and **F** in Sect. "Small Gene Circuits") that determines how the ON/OFF status of multiple inputs of a target gene map into its behavior (output status) is a logical (Boolean) function B, and the network topology is randomly-wired, with the exception of some deliberately fixed features. Thus, the work on random Boolean networks allowed the study of the generic behavior of large networks of thousands of genes even before molecular biology could deliver the actual connections and interaction logics of the GRN of living systems (for review, see [8]).

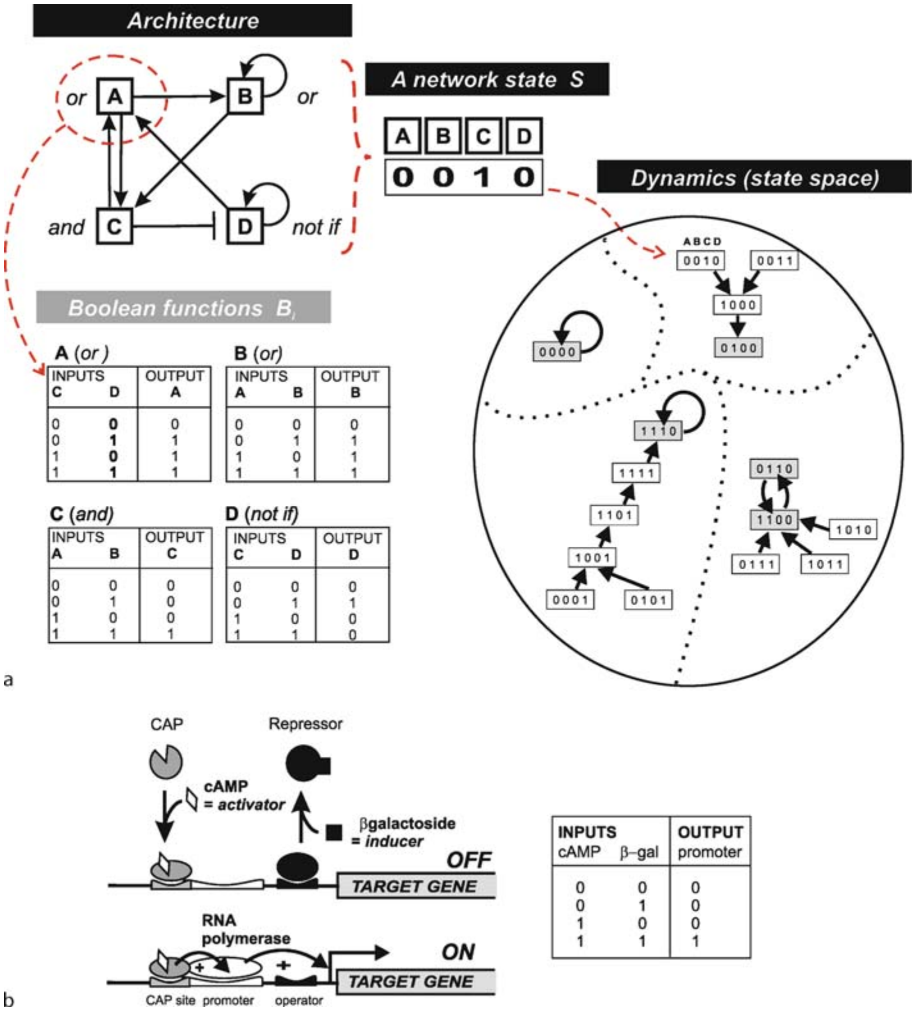
The lack of detailed knowledge about specific genes, their interaction functions and the formidable computational cost for modeling genome-wide networks has warranted a coarse-graining epitomized by the Boolean network approach. But more specifically, in the broader picture of system-wide dynamics, the discretization of gene expression level is also justified because (*i*) the above discussed steep sigmoidal shape of "transfer functions" (Fig. 2b) that describe the influence of one gene onto another's expression rate can be approximated by a step-function and/or (*ii*) the local dynamics produced by such small gene circuit modules is in fact characterized by discontinuous transitions between discrete states as shown in Sect. "Network Dynamics".

In addition, the Boolean network approach offers several advantages over the exhaustive, maximally detailed models favored by engineers who seek to understand a *particular instance* of a system rather than *typical* properties of a *class* of systems. The simplification opens a new vista onto fundamental principles that may have been obscured by the details since, as philosophers and physicists have repeatedly articulated, there is no understanding without simplification, and "less can be more" [11,30, 159]. An important practical advantage of the simplification in the Boolean network approach is the possibility to study *statistical ensembles* of ten thousands of network instances, i. e., entire classes of network architectures, and to address the question of how a particular architecture type maps into a particular type of dynamic behavior. Some of the results obtained in fact are valid for both discrete and continuous behavior of the network nodes [17].

Model Formalism for Boolean Networks

In the Kauffman model of random Boolean networks gene activity values are binary (1 = ON, and 0 = OFF) [111, 115,116] and time is also discretized. Thus, a Boolean network is a generalized form of cellular automata but without the aspect of physical space and the particular neighborhood relations. Then, in analogy to continuous systems, a network of N elements i ($i = 1, 2, \dots, N$), defines a network state S at any given discrete time step t: $\mathbf{S}(t) = [x_1(t), x_2(t), \dots, x_N(t)]$ where x_i is the activity status that now only takes the values 1 or 0. The principles are summarized in Fig. 4. The state space that represents the entire dynamics of the network is finite and contains 2^N states. However, again, not all states are equally likely to be realized and observed, since genes do not behave independently but influence each other's expression status.

Regulation of gene i by its incoming network connections is modeled by the Boolean function B_i that is associated with each gene i and maps the configuration of the activity status (1 or 0) of its input genes (upstream regulators of gene i) into the new value of x_i for the next time point. Thus, the argument of the Boolean function B_i is the input vector $\boldsymbol{R}_i(t) = [x_1(t), x_2(t), \dots, x_{ki}(t)]$, where k_i is the number of inputs that the gene i receives. At each time step, the value of each gene is updated: $x_i(t+1) = B_i[\boldsymbol{R}_i(t)]$. The logical function B_i can be formulated as a "truth table" which is convenient for large k's (Fig. 4a). In the widely studied case where each gene has exactly $k = 2$ inputs, the Boolean function can be one of



A (*or*)

INPUTS C	D	OUTPUT A
0	0	0
0	1	1
1	0	1
1	1	1

B (*or*)

INPUTS A	B	OUTPUT B
0	0	0
0	1	1
1	0	1
1	1	1

C (*and*)

INPUTS A	B	OUTPUT C
0	0	0
0	1	0
1	0	0
1	1	1

D (*not if*)

INPUTS C	D	OUTPUT D
0	0	0
0	1	1
1	0	0
1	1	0

INPUTS cAMP	β-gal	OUTPUT promoter
0	0	0
0	1	0
1	0	0
1	1	1

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 4

Principles of Boolean network models of GRN, exemplified on a $N = 4$ gene network. a From *architecture* to dynamics. The four genes i: A, B, C, D interact as indicated by the directed graph (*top left panel*), and each of the genes is assigned a Boolean function B_i as indicated, with the corresponding "truth table" shown below. A network state is a string of $N = 4$ binary variable, thus there are $2^N = 2^4 = 16$ possible *network states S*. They collectively establish the entire state space and can be arranged in a *state transition map* according to the state transitions imposed by the Boolean functions (*right panel*). The attractor states are colored *gray*. In the example, there are three point attractors and one limit cycle attractor (of period $T = 2$). The dotted lines in the state space denote the attractor boundaries. b Example for capturing the regulation at the promoter of the lactose operon as an "AND" Boolean function. Note that there are many ways to define what constitute an input – in the case shown, the allosteric ligands cAMP ("activator") and a β-galactoside ("inducer"), such as allolactose, give rise to a two-input Boolean function 'AND'

the set of $2^{(2\wedge k)} = 16$ classical Boolean operators, such as AND, OR, NOTIF, etc. [109,116]. Figure 4b shows the example of the well-studied lactose operon and how its regulatory characteristics can be captured as an AND Boolean function for the two inputs.

In the simplest model, all genes are updated in every time step. This synchrony is artificial for it assumes a central clock in the cell which is not likely to exist, although some gating processes from oscillations in the redox potential has been reported [120]. The idealization of synchrony, however, facilitates the study of large Boolean networks, and many of the results that have been found with synchronous Boolean networks carry over to networks with asynchrony of updating that has been implemented in various ways [40,72,81,119]. For synchronous networks the entire network state $\boldsymbol{S}$ can also be viewed as being updated by the updating function U: $\boldsymbol{S}(t+1) = U[\boldsymbol{S}(t)]$, where U summarizes all the N Boolean functions B_i. This facilitates some treatments and is represented in the ***state transition table*** that lists all possible network states in one column and its successor in a second column. This leads to a higher-level, directed graph that represents the entire dynamics of a network as illustrated below.

Dynamics in Boolean Networks

The state transition table captures the entire dynamical behavior of the network and can conveniently be depicted as a ***state transition map***, a directed graph in which a node represents one of the 2^N possible state $S(t)$ (a box with a string or 1, 0s in Fig. 4a. representing the gene expression pattern). Such diagrams are particularly illustrative only for N up to ~ 10, since they display all possible states S of the finite state space [204]. The states are connected by the arrows which represent individual state transitions and collectively depict the ***trajectories*** in the state space (Fig. 4a, right panel).

Since the Boolean functions are *deterministic*, a state $S(t)$ unambiguously transitions into *one* successor state $S(t+1)$. In contrast, a state can have multiple predecessors, since two or more different states can be updated into the same state. Hence, trajectories can converge but not diverge, i. e., there is no branching of trajectories emanating from one state. This property of "losing information about the ancestry" is essential to the robustness of the dynamics of networks.

In updating the network states over time, $S(t)$ can be represented as a walk on the directed graph of the state transition map. Because of the finiteness of the state space in discrete networks, $S(t)$ will eventually encounter an already visited state, and because of the absence of divergence, the set of successive states will be the same as in the previous encounter. Thus, no matter what the initial state is, the network will eventually settle down in either a set of cycling states (which form a *limit cycle*) or in a single stable state that updates onto itself. Accordingly, these states to which all the trajectories are "attracted" to are the ***attractors*** of the network. They are equivalent to stable oscillators or stable fixed-points, respectively, in the continuous description of gene circuits (Sect. "Bistability, Tristability and Multistability"). In other words, because of the regulatory interactions between the genes, the system cannot freely choose any gene expression configuration S. Again, most network states S in the state space are thus unstable and transition to other states to comply with the Boolean rules until the network finds an attractor. And as with the small, continuous systems (Sect. "Bistability, Tristability and Multistability") the set of states S that "drain" to an attractor state constitute its ***basin of attraction***.

However, unlike continuous systems, Boolean network dynamics do not produce unstable steady-states that can represent the indeterminacy of undecided cell states that correspond to stem cells about to make an either-or decision between two lineages (see Sect. "Bistability, Tristability and Multistability"). Instead, basins of attraction are "disjoint" areas of state space.

Attractors as Cell Types Kauffman proposed that the high-dimensional attractor states represent cell types in metazoan organisms – thus expanding the early notion of steady state in small circuits to networks of thousands of genes [111,116] (Sect. "Bistability, Tristability and Multistability"). This provides a natural explanation for the stability of the genome-wide expression profile that is characteristic of and defines a cell type as well as for the stable coordination of genome-wide gene expression oscillations in proliferating cells that undergo the cell division cycle [202]. The correspondence of attractors in large networks with the cell type specific transcriptome is a central hypothesis that links the theoretical treatment of the dynamics of complex networks with experimental cell biology.

Use of Boolean Networks to Model Real Network Dynamics Owing to their simple structure, Boolean networks have been applied in place of differential equations to model real-world networks for which a rudimentary picture of the topology with only few details about the interaction functions is known. Hereby, individual Boolean functions are assigned to the network nodes either based on best guesses, informed by qualitative descriptions from the experimental literature, or randomly. This approach has yielded surprisingly adequate recapitulation of biological behaviors, indicating that the topology itself accounts for a great deal of the dynamical constraints [4,49,60,62,95,130]. Such studies also have been used to evaluate the dynamical regime (discussed in the next section) of real biological networks [19].

Three Regimes of Behaviors for Boolean Networks

The simplicity and tractability of the Boolean network formalism has stimulated a broad stream of investigations that has led to important insights in the fundamental properties of the generic dynamics that large networks can generate [116] even before progress in genomics could even offer a first glimpse on the actual architecture of a real gene regulatory network [112].

Using the ensemble approach, the architectural parameters that influence the global, long-term behavior of complex networks with N up to 100,000 has been determined [114,116]. As mentioned in the introduction, based on the latest count of the genome size and the idea that the core transcriptional network essentially governs the global dynamics of gene expression profiles, networks of $N \sim 3000$ would have sufficed. Nevertheless, a striking result from the ensemble studies was that for a broad class of architectures, even a complex, irregular (randomly wired) network can produce 'ordered' dynamics with stable patterns of gene expression, thus potentially delivering the "biological observable". In general, the global behavior of ensembles of Boolean networks can be divided in three broad regimes [116]: an *ordered* and a *chaotic* regime, and a regime that represents behavior at their common border, the *critical* regime.

Ordered and Chaotic Regimes of Network Behavior

In networks in the ***ordered*** regime, two randomly picked initial states $S_1(t = 0)$ and $S_2(t = 0)$ that are closed to one another in state space as measured by the Hamming distance $H[H(t) = |S_1(t) - S_2(t)| \equiv$ the number of genes whose activities differ in the two states $S_1(t)$ and $S_2(t)]$ will exhibit trajectories that on average will quickly converge (that is, the Hamming distance between the two trajectories will on average decrease with time) [55]. The two trajectories will settle down in one fixed-point attractor or a limit cycle attractor that will have a small period T compared to N and thus, produce very stable system behaviors. Such networks in general have a small number of attractors with typically have small periods T and drain large basins of attraction [116]. Numerical analyzes of large ensembles suggest that the average period length scales with $\sqrt{N}$. The state transition map, as shown for $N = 4$ in Fig. 4a, will show that trajectories converge onto attractor states from many different directions and are in general rather short [Maliackal, unpublished] so that attractor basins appear as compact, with high rotational symmetry and hence, "bushy".

It is important to stress here that if a cell type is an attractor, then different cell types are different attractors, and, in the absence of the unstable steady states present in continuous dynamical systems (see Sect. "Bistability, Tristability and Multistability") differentiation consists in perturbations that move a system state from one attractor into the basin of another attractor from which it flows to the new attractor state that encodes the gene expression profile of the new cell type. Examination of the bushy basins in the ordered regime makes it clear, as do numerical investigations and experiments (Sect. "Are Cell Types Attractors?"), that multiple pathways can lead from one attractor to another attractor – a property that meets resistance in the community of pathway-centered biologists.

In contrast, in networks in the ***chaotic*** regime, two randomly placed initial states $S_1(t = 0)$ and $S_2(t = 0)$ that are initially close to one another (in terms of Hamming distance) will generate trajectories that on average will *diverge* and either end with high likelihood in two different attractors, or they may appear to "wander" aimlessly in state space. This happens because the attractor is a limit cycle attractor with very long period T – on the scale of 2^N so that in the worst case a trajectory may visit most if not all 2^N possible network configurations S. For a small network of just $N = 200$, this is a limit cycle in the order of length $2^{100} \approx 10^{30}$ time steps. As a point of comparison, the universe is some 10^{17} seconds old. Given the hyperastronomic size of this number, this "limit cycle" will appear as an aperiodic and endless stream of uncorrelated state transitions, as if the system is on a "permanent transient". Thus, networks in the chaotic regime are not stable, trajectories tend to diverge (at least initially), and their behavior is sensitive to the initial state. In the state transition map, the small attractors typically receive trajectories with long transients that arrive from a few state space directions and hence, in contrast to the bushy attractors of the ordered regime, the basins have long thin branches and appear "tree-like".

The definition of "chaos" for discrete networks given here is distinct from that of (deterministic) chaos in continuous systems, where time evolution of infinitesimally closed initial states can be monitored and shown to diverge. Nevertheless, the degree of chaos in discrete networks, as qualitatively outlined above, is well-defined and can be quantified based on the slope of the curve in the so-called Derrida plot which assesses how a large number of random pairs of initial states evolves in one time step [55].

More recently, it was shown that it is possible to determine the behavior class from the architecture, without simulating the state transitions and determining the Derrida plot, simply by calculating the expected *average sensitivity* from all the N Boolean functions B_i [173]. In addition, a novel distance measure that uses normalized compression distances (NCD) – which captures the complexity in the difference between two states $S(t)$ better than the Hamming distance used in the Derrida plot – has been proposed to determine the regime of networks [153].

Critical Networks: Life at Edge of Chaos?

Critical networks are those which exhibit a dynamical behavior just at the edge between that of the ordered and the chaotic regime, and have been postulated to be optimally poised to govern the dynamics of living cells [114, 116]. *Ordered* behavior would be too rigid, in that most perturbations of a stable attractor would be absorbed and the network would return to the same attractor state, minimizing the possibility for a network to change its internal equilibrium state in response to external signals (which are modeled as flipping individual genes from ON to OFF or vice versa). *Chaotic* behavior, on the other hand, would be too sensitive to such perturbations because trajectories diverge – so that the network would wander off in state space and fail to exhibit robust behavior. *Critical* networks may represent the optimal mix between stability and responsiveness, hence conferring robustness to random perturbations (noise) and adaptability in response to specific signals. Critical network have several remarkable features:

First, consider the need of cells to plausibly make the maximum number of reliable discriminations, and to act on them in the presence of noise with maximum reliability. Then 'deep' in the ordered regime, convergence of trajectories in state space is high, hence, as explained above (Sect. "Dynamics in Boolean Networks") information is constantly discarded. In this "lossy" regime, information about past discriminations is thus easily lost. Conversely, in the chaotic regime and with only a small amount of noise, the system will diverge and hence cannot respond reliably to external signals (perturbations). It seems plausible that optimal capacity to categorize and act reliably is found in critical networks, or networks slightly in the ordered regime.

Second, it has recently been shown that a measure of the correlation of pairs of genes with respect to their altering activities, called "***mutual information***" (MI) is maximized for critical networks [154]. The MI measures the mutual dependence of two variables (vectors), such as two genes based on their expression in a set of states $S(t)$. Consider two genes in a synchronous Boolean network, A and B. The mutual information between x_{A} and x_{B} is defined as $\mathrm{MI(A,B)} = H(x_{\mathrm{A}}) + H(x_{\mathrm{B}}) - H(x_{\mathrm{A}}, x_{\mathrm{B}})$. Here, $H(x)$ is the entropy of the variable x, and $H(x, y)$ is the joint entropy of x and y. Mutual information is 0 if either gene A or B is unchanging, or if A and B are changing in time in an uncorrelated way. Mutual information is greater than 0 and bounded by 1.0, if A and B are fluctuating in a correlated way. Thus, *critical networks maximize the correlated changing behavior of variables in the genetic network*. This new result strongly suggests, at least in the ensemble of random Boolean networks, that critical networks can coordinate the most complex organized behavior.

Third, the "basin entropy" of a Boolean network, which characterizes the way the state space is partitioned in to the disjoint basins of attraction of various sizes (Fig. 4a) also exposes a particular property of critical networks [124]. If the size or "weight", W_i, of a basin of attraction i is the fraction of all the 2^N states that flow to that attractor, then the basin entropy is defined as $\Sigma W_i \log(W_i)$. The remarkable result is that *only* critical networks have the property that this basin entropy continues to increase as the size of the network increases. By contrast, ordered and chaotic networks have basin entropies that first increase, but then stop increasing with network size [124]. If one thinks of basins of attraction and attractors not only as cell fates, or cell types, but as distinct specific cellular programs encoded by the network, then only critical networks appear to have the capacity to expand the diversity of what "a cell can do" with increasing network size. Again, this strongly suggests that critical networks can carry out the most complex coordinated behaviors. Thus, GRN may have evolved under natural selection (or otherwise – see Sect. "Evolution of Network Structure: Natural Selection or Natural Constraint?") to be critical.

Finally, it is noteworthy that while Boolean networks were invented to model GRNs, the variables can equally be interpreted as any kind of two-valued states of components in a cell, and the Boolean network becomes a *causal network* concerning events in a cell, including GRN as a subset of such events. This suggests that not only GRN but the entire network of processes in cells, including signal transduction and metabolic processes, that is, information, mass and energy flow, may optimally be coordinated if the network is critical.

If life is poised to be in the critical regime, then the questions follow: which architecture produces ordered, critical and chaotic behaviors, and are living cells in fact the critical regime?

Architectural Features of Large Networks and Their Dynamics

As outlined in Sect. "Network Architecture", the recent availability of data on gene regulation in real networks, although far from complete, has triggered the study of complex, irregular network topologies as static graphs. This line of investigation is now beginning to merge with the study of the dynamics of generic Boolean networks. Below we summarize some of the interesting architecture features and their significance for global dynamics in term of the three regimes. First, studies of generic dynamics in ensembles of Boolean networks have established the following major structure-dynamics relationships:

(1) *The average input connectivity per node, k.* Initial studies on Boolean networks by Kauffman assumed a homogenous distribution of inputs k. It was found

that $k = 2$ networks are in the ordered/critical regime (given other parameters, see below) [116]. Above a critical k_c value (which depends on other parameters, see below) networks behave chaotically. Analysis of continuous, linearized models also suggest that in general, sparsity of connections is more likely to promote ordered dynamics – or stability [142].

(2) The *distribution of the connectivity (degree)* over the individual network genes. As mentioned in Sect. "Network Architecture", the global topology of many complex molecular networks appears to have a connectivity distribution that approximates a power-law. For directed graphs such as the GRN, one needs to distinguish between the distribution of the input and output connectivities. Analyses of the dynamics of random Boolean networks with either scale-free input [66] or output connectivity distribution suggest that this property favors the ordered regime for a given value of the parameter p ("internal homogeneity", see below) [6]; Specifically, if the slope of the scale free distribution (power-law exponent γ) is greater than 2.5, the corresponding Boolean network is ordered, regardless of the value of the parameter p (see below). For values of p approaching 1.0, $\gamma > 2.0$ suffices to assure ordered dynamics.

(3) The nature of the Boolean functions is an important aspect of the network architecture that also influences the global dynamics. In the early studies, Kauffman characterized Boolean functions with respect to these two important features [116]:

(a) "*internal homogeneity* p". The parameter p $(0.5 < p < 1)$ is the proportion of either 1s or 0s in the output column of the truth table of the Boolean function (Fig. 4). Thus, a function with $p = 0.5$ has equal numbers of 1s and 0s for the output of all input configurations. Boolean functions with p-values close to 1 or 0 are said to exhibit high internal homogeneity.

(b) "*Canalizing function*". A Boolean function B_i of target gene i is said to be canalizing if at least one of its inputs has one value (either 1 or 0) that determines the output of gene i (1 or 0), independently of the values of the other components of the input vector. If the two values of the input j determines *both* output values of gene i [a "fully canalizing" function, e. g., if $x_j(t) = 1$ (or 0, respectively), then $x_i(t + 1) = 1$ (or 0, respectively)], then the other inputs have no influence on the output at all and the "*effective* input connectivity" of i, $k_{i\text{eff}}$ is smaller than the "*nominal*" k_i. For instance, for Boolean functions with $k = 2$, only two of the 16 possible functions, XOR and XNOR, are not canalizing. Four functions are "fully canalizing", i. e., are effectively $k = 1$ functions (TRANSFER, COMPLEMENT).

From ensemble studies it was found that both a high internal homogeneity p and a high proportion of canalizing functions contribute to ordered behavior [116].

Do Real Networks Have Architecture Features That Suggest They Avoid Chaos?

Only scant data is available for transcriptional networks, and it must be interpreted with due caution since new data may, given sampling bias and artefacts, especially for nearly scale-free distributions, affect present statistics. In any case, existing data indicate that in fact the average input connectivity of GRNs is rather low, and far from $k \sim N$ which would lead to chaotic behavior. Specifically, analysis of available (partial) transcriptional networks suggest that the *input* degrees approximates an *exponential* distribution while the *output* degree distribution seems to be *scale-free* although the number of nodes are rather small to reliably identify a power-law distribution [54,82].

GRN data from *E. coli*, for which the most complete, hand-curated maps of genome-wide gene regulation exists [169] and from partial gene interaction networks from yeast, obtained mostly by chromatin-precipitation/microarray (ChIP-chip) [128], indicate that the average input connectivity (which is exponentially distributed) is below 4 [82,134,188]. A recent work on the worm *C. elegans* using the Yeast-one-hybrid system on a limited set of 72 promoters found an average of 4 DNA-binding proteins per promoter [54]. Thus, while such analyzes await correction as coverage increase, clearly, real GRNs for microbial and lower metazoan are sparse, and hence more likely to be in or near the ordered regime.

In contrast to the input, the output connectivity appears to be power-law distributed for yeast and bacteria, and perhaps also for *C. elegans*, if the low coverage of the data available so far can be trusted [54,82]. The paucity of data points precludes reliable estimates of the power-law exponents. However, the scale-free property, if confirmed for the entire GRN, may well also contribute to avoiding chaotic and increasing the ordered or critical regime [6]. Nevertheless, It is reminded here that the deeper meaning of the scale-freeness per se and its genesis (natural selection due to functionality or not) are not clear – it may be an inevitable manifestation of fundamental statistics rather than evolved under natural selection (Sect. "Evolution of Network Structure: Natural Selection or Natural Constraint?").

As for the use of Boolean functions, analysis of a set of 150 experimentally verified regulatory mechanism of well-studied promoters revealed an enrichment for canalizing functions, again, in accordance with the architecture criteria associated with ordered dynamics [86]. Similarly, when canalizing functions were randomly imposed onto the published yeast protein–protein interaction network topology to create a architecture whose dynamics was then simulated, ordered behavior was observed [113].

In this conjunction it is interesting to mention ***microRNAs*** (miRNAs), a recently discovered class of non-coding RNA which act by inhibiting gene expression at the post-transcriptional level through sequence complementarity to mRNA [41,89]. Hence, they suppress a gene independent of the TF constellation at the promoter. Thus, miRNAS epitomize the most simple and powerful molecular realization of a canalizing function. Their existence may therefore shift the network behavior from the chaotic towards the critical or ordered regime. Interestingly, RNA-based gene regulation is believed to have appeared before the metazoan radiation 1000 million years ago, and microRNAs are thought to have evolved in ancestors of Bilateria [141]. In fact, many miRNA play key roles in cell fate determination during tissue specification of development of vertebrates [41,172], and a composite feedback loop circuit involving both microRNA and TFs has been described in neutrophil differentiation [63]. The shift of network behavior from the chaotic towards the critical or ordered regime may indeed enable the coordination of complex gene expression patterns during ontogeny of multicellular systems which requires maximal information processing capacity to ensure the coexistence of stability and diversity.

There are, as mentioned in Sect. "Analyzing Network Structures", many more global and local topological features that have been found in biomolecular networks that appear to be interesting, as defined in the sense in Sect. "Network Architecture" (enriched above some null model graph). It would be interesting to test how they contribute to producing chaotic, critical or ordered behavior. The impact of most of these topology features on the global dynamics, notably the three regimes of behaviors, remains unknown, since most functional interpretation of network motifs have focused on local dynamics [9,136]. The low quality and availability of experimental data of GRN architectures opens at the moment only a minimal window into the dynamical regimes of biological networks. The improvement with respect to coverage and quality of real GRNs, to be expected in the coming years, is mostly driven by a reductionist agenda whose intrinsic aim is to exhaustively enumerate all the "pathways" that may serve "predictive modeling" of gene behaviors, as detailed in Sect. "Small Gene Circuits". However, with the concepts introduced here, a framework now exists that warrants a deeper, genome-wide analysis of the relationship between structure and biological function. Such analysis should also address the fundamental dualism between inevitable self-organization (due to intrinsic constraints from physical laws) and natural selection (of random mutants for functional advantages) [46,93,203] to ask whether criticality is self-organized (see Sect. "Evolution of Network Structure: Natural Selection or Natural Constraint?").

Experimental Evidence from Systems Biology

The experimental validation of the central concepts that were erected based on theoretical analysis of network dynamics, notably the ensemble approach of Boolean networks, amounts to addressing the following two questions:

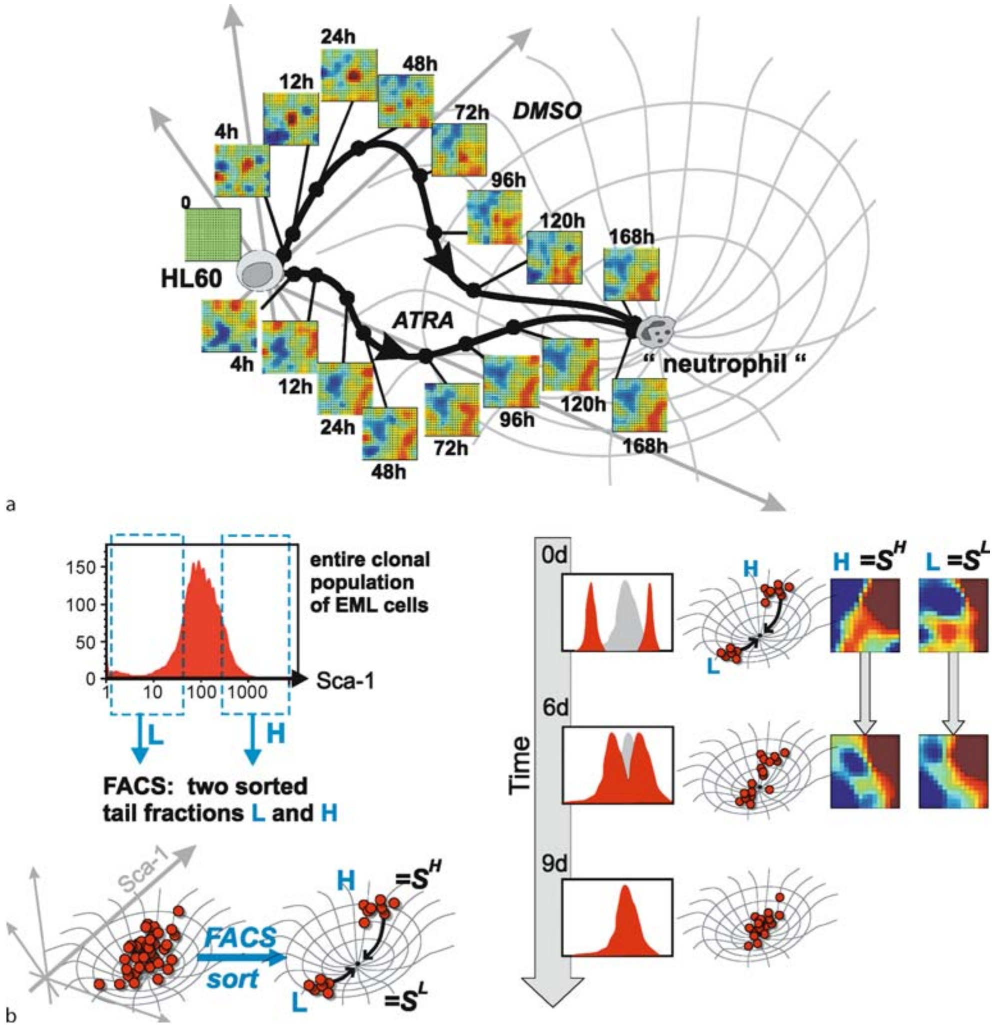
1. Are cell types attractors?
2. Is the dynamical behavior of the genomic GRN in the critical regime?

As experimental systems biology begins to reach beyond the systematic characterization of genes and their interactions it has already been possible to design experiments to obtain evidence to address these questions.

Are Cell Types Attractors?

Obviously, the observable gene expression profiles S^*, as shown in Fig. 1e, are stationary (steady) states and characteristic of a cell type. But "steady state" ($\mathrm{d}\mathbf{x}/\mathrm{d}t = 0$ in Eq. (1)) does not necessary imply a stable (self-stabilizing) attractor state that attracts nearby states. In the absence of knowledge of the architecture of the underlying GRN (we do not know the function **F** in Eq. (1)) we cannot perform the standard formal analysis, such as linear stability analysis around S^*, to determine whether a stationary state is stable or not; however, use of microarray-based expression profiling not to identify individual genes as in the gene-centered view, but in a novel integrated manner (Table 1, B3) for the analysis of $S(t)$ provides a way to address the question. The qualitative properties of an attractor offer a handle, in that a high-dimensional attractor state S^*, be it a fixed-point or a small limit cycle or even a strange attractor in the state space, requires that the volume of the states around it contracts to the attractor, i. e., $\mathrm{div}(\mathbf{F}) < 0$. [90].

Convergence of Trajectories Thus, one consequence is that trajectories emanating from states around (and near) S^* converge towards it from most (ideally, all) directions



Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 5
Experimental evidence that a cell fate (differentiated state, cell type) is a high-dimensional stable attractor in gene expression state space. Evidence is based on convergence (**a**) from different directions of high-dimensional trajectories to the gene expression pattern of the differentiated state, or on the relaxation of states placed near the border of the basin of attraction back to the "center" of the attractor state (**b**). **a** Gene expression profile dynamics of HL60 cells treated with ATRA (all-trans retinoid acid) or DMSO (dimethylsulfoxide) at 0h to differentiate them into neutrophil-like cells. Selected gene expression profile snapshots along the trajectories shown as GEDI maps (as in Fig. 1), schematically placed along the state space trajectories. GEDI maps show the convergence of $N \sim 2800$ genes towards very similar patterns at 168h [98]. **b** Heterogeneity of a population of clonal (genetically identical) cells is exploited to demonstrate the relaxation towards the attractor state. The histogram (*top, left panel*) from flow cytometric measurement shows the inherent heterogeneity (spread) of cells with respect to stem cell marker Sca-1 expression in EML cells which cannot be attributed solely to measurement or gene expression noise but reflect metastable cell-individuality [37]. Two spontaneous "outlier" subfractions, expressing either low (L) or high (H) levels of the Sca-1, were sorted using FACS (fluorescence-activated cell sorting) and cultured independently. They represent "small perturbations" of the attractor state. Each sorted subpopulation will over a period of 5–9 days restore the parental distribution (*right panel, schematic*). Gene expression profiling (shown as as GEDI maps) reveal that the H and L subpopulations are distinct with respect not only to Sca-1 expression levels but also to that of multiple other genes. Thus, the two outlier cell fractions are at distinct states S^H and S^L in the high-dimensional state space of which Sca-1 is only one distinguishing dimension. The restoration of the parental distribution of Sca-1 is accompanied by the approaching of the two distinct gene expression profiles S^H and S^L of the spontaneously perturbed cells to become similar to each other (and to that of the parental population), indicating a high-dimensional attractor state

and in all dimensions of the state space. It is technically challenging to sample multiple high-dimensional initial states near the attractor states and demonstrate that they all converge towards S^*over time. However, the fact that the promyeloid precursor cells HL60 (a leukemic cell line) can be triggered to differentiate into mature neutrophil-like cells by an array of distinct chemical agents can be exploited [45]. This historical observation itself, without analysis of $S(t)$ by gene expression profiling, already suggests that the neutrophil state is an attractor, because it reflects robustness and means that the detailed history of how it is reached doesn't matter.

But the advent of microarray-based gene expression profiling technologies has recently opened up the possibility to show that (at least two) distinct trajectories converge towards gene expression profile of the neutrophil state S_{Neutr} if S_{Neutr} is an attractor state [98]. Thus, HL60 cells were treated with either one of two differentiation-inducing reagents, all-trans-retinoid acid (ATRA) and dimethyl sulfoxide (DMSO), two chemically unrelated compounds, and the changes of the transcriptome over time in response to either treatment were measured at multiple time intervals to monitor the two trajectories, $S^{\text{ATRA}}(t)$ and $S^{\text{DMSO}}(t)$, respectively (see Fig. 5a for details). In fact the two trajectories first diverged to an extent that the two state vectors lost most of their correlation (i. e. the two gene expression profiles $S^{\text{ATRA}}(24\,\text{h})$ and $S^{\text{DMSO}}(24\,\text{h})$ were maximally different at $t = 24$ hours after stimulation). But subsequently, they converged to very similar $S(t)$ values when the cells reached the differentiated neutrophil state for both treatments (Fig. 5a). The convergence was not complete, but quite dramatic relative to the maximally divergent state at 24 h, and was contributed by around 2800 of the 3800 genes monitored (for details see [98]). Thus, it appears that at least the artificial drug-induced differentiated neutrophil state is so stable that it can apparently orchestrate the expression of thousands of genes to produce the appropriate cell-type defining expression pattern S from quite distinct perturbed cellular states. Although only two trajectories have been measured rather than an entire state space volume, the convergence with respect to 2800 state space dimensions is strongly indicative of a high-dimensional attractor state.

Relaxation After Small Perturbations and the Problem of Cell Heterogeneity A second way to expose qualitative properties of a high-dimensional attractor S^* is to perform a weak perturbation of S^*, into a state S' near the edge of (but within) the basin of attraction (where S' should differ from S^* with respect to as many genes x_i as possible), and to observe its return of the network to S^*. This more intuitive property of an attractor state was difficult to measure because of a phenomenon often neglected by theorists and experimentalists alike: cell population heterogeneity due to "gene expression noise". Microarray measurements, as is the case with many biochemical analysis methods, require the material of millions of cells, thus the measured $S(t)$ is actually a population average. This can be problematic because the population is heterogeneous [33]: gene expression level of gene i can typically differ by as much as 100 or more fold between two individual cells within a clonal (genetically identical) population [37]. Thus, virtually all genes i exhibit a broad (log-normal) histogram (Fig. 5b, inset) when the expression level of an individual gene x_i is measured at the single-cell level across a population [38,101,129]. Such cell-to-cell variability is often explained by "gene expression noise" caused by random temporal fluctuations due to low copy numbers of specific molecules in the cell [108]. However, there may be other possibly deterministic sources that generate metastable variant cells [33,176]. Such non-genetic, enduring individuality of cells translates into the picture of a cell population forming a cloud of points in state space around the attractor S^*, in which each cell represents a single point. Application of a low dose perturbation intended to allow a system to relax back to the attractor state then will be interpreted by individual cells differently: Those positioned at the border of the cloud may be kicked out of the basin of attraction, and move to another attractor, thus masking the trajectory of relaxation. In fact, single-cell resolution measurements of the response to low dose stimulation in cell populations confirmed this picture of heterogeneity, in that a differentiation inducer given at low dose to trigger a partial response (weak perturbation) produced a bimodal distribution of gene expression of differentiation markers: Some cells differentiated, other did not [38].

However, the spontaneous heterogeneity eo ipso, consisting of transient but persistent variant cells within a population [37], allows us to demonstrate the relaxation to the attractor when single cell-level manipulation and analysis are performed: Physical isolation of the population fractions at two opposite the edge of the cloud (basin of attraction), based on one single state space dimension x_k can substitute for the weak perturbation that places cells to the border of a basin (see Fig. 5b for details). Indeed such "outlier" population fractions exhibited not only distinct levels of x_k expression but also globally distinct gene expression profiles (despite being members of the same clonal population). Cells of both outlier fractions eventually "flowed back" to populate the state space region around the attractor state and restored the original distribution (shape of the cloud) [37]. The time scale of this relaxation (>5 days) was similar to that for the HL60 cells to converge to the attractor of the differentiated neutrophils. This result is summarized in Fig. 5b. Again, the spontaneous regeneration of the distinct gene expression profile of a macroscopically observable cell phenotype, consisting of thousands of genes, supports the notion of a high-dimensional attractor in gene expression space that maps into a distinct, observable cell type.

Are Gene Regulatory Networks in Living Cells Critical?

The second central question we ask in this article is whether GRNs produce a global dynamics that is in the critical regime (Sect. "Three Regimes of Behaviors for Boolean Networks"). This question is not as straightforward to address experimentally. First, the notion of order, chaos and criticality are defined in the models as properties of network ensembles. Second, it is cumbersome to sample a large number of pairs of initial states and monitor their high-dimensional trajectories to determine whether they on average converge, diverge or stay "parallel".

As mentioned earlier (Sect. "Dynamics in Boolean Networks"), one approach for obtaining a first glimpse of an idea as to whether real networks are critical or not was recently proposed by Aldana and coworkers [19] and refs. herein): They "imposed" a dynamics onto real biological networks, for which only the topology but not the interaction functions are known, by treating them as Boolean networks, whereby the Boolean functions were guessed or randomly assigned according to some rules. Such studies suggest that these networks, given their assumed topologies, are in the critical regime.

To more directly characterize the observed dynamics of natural systems in terms of the three regimes, several indirect approaches have been taken. These strategies are based on novel measurable quantities computed from several schemes of microarray experiments that are now available but were not originally generated with an intention to answer this question. By first determining in simulated networks whether that quantity is associated also with criticality in silico, inference is then made as to what regime of system behavior the observed system resembles. Three such pieces of evidence provide a first hint that perhaps, living cells may indeed be in the critical regime:

(i) Gene expression profile changes during Hela cell cycle progression was compared with the detailed temporal structure of the updating of network states in Boolean networks. Tithe discretized real gene expression data of cells progressing in the cell cycle were compared with that of simulated state cycles of random Boolean networks in the three regimes in terms of the Lempev-Zip complexity of time series. This led to the conclusion that the dynamical behavior of thousands of genes was most consistent with either ordered or critical behavior, but not chaotic behavior [174].

(ii) A striking property predicted from analysis of simulated critical Boolean networks is that if a randomly selected gene is deleted and the number of other genes that change their expression as a consequence of that single-gene deletion ("avalanche size") is measured, and such single-gene deletion experiments is repeated many times, the avalanche-sizes will exhibit a power law distribution with a slope of $-\gamma = 1.5$. This specific behavior is only seen in critical networks. Analysis of just such data for over 200 single deletion mutants in yeast [163] from the experiments reported by Hughes et al. [100]was recently performed. It was found that the distribution of the avalanche sizes not only approached a power law, but that the slope was also -1.5 [163]. This result was insensitive to altering the criterion for defining a "change in gene activity" from two-fold to five fold in calculating the avalanche size.

(iii) A more direct determination of the regime of network dynamics was recently reported, in which an analogous analysis as the Derrida plot (see Sect. "Ordered and Chaotic Regimes of Network Behavior") was performed [153]. Macrophage gene expression profiles were measured at various time points in response to various perturbations (= stimulation of Toll-like receptors with distinct agents), offering a way to measure the time evolution of a large number of similar initial states. Here, instead of the Hamming distance the normalized compression distance NCD (as mentioned in Sect. "Ordered and Chaotic Regimes of Network Behavior") was used to circumvent the problems associated with Derrida curves. The results were consistent with critical dynamics, in that on average, for many pairs of initial states, their distance at time t and $t + \Delta t$ was on average equal, thus neither trajectory convergence nor divergence took place.

Future Directions and Questions

The ideas of a state space representing the dynamics of the network, and of its particular structure that stems from the constraints imposed by the regulatory interactions and can be epitomized as an 'epigenetic landscape', serve as a conceptual bridge linking network architecture with the observable biological behavior of a cell. In this framework, the attractors (valleys) are disjoint regions in the state space landscape and represent cell fates and cell types. One profound and bold hypothesis is that for networks to have a landscape with attractors that optimally convey cell lineage identity and robustness of their gene expression profile, yet allow enough flexibility for cell phenotype switches during development, the networks must be poised at the boundary between the chaotic and the ordered regime. Such critical networks may be a universal property of networks that have maximal information processing capacity. But what can we naturally do with this conceptual framework presented here? And what are the next questions to ask for the near future?

Clearly, new functional genomics analysis techniques will soon advance the experimental elucidation of the architecture of the GRN of various species, including that of metazoan organisms. This will drastically expand the opportunities for theoretical analysis of the architecture of networks and finally afford a much closer look at the dynamics without resorting to simulated network ensembles.

However, beyond network analysis in terms of mathematical formalisms, the concepts of integrated network dynamics presented here should also pave the way for a formal rather than descriptive understanding of tissue homeostasis and development as well as diseases, such as cancer. One corollary of the idea that cell types are attractors is that cancer cells are also attractors – which are lurking somewhere in state space (near embryonic or stem cell attractors) and are normally avoided by the physiological developmental trajectories. They become accessible in pathological conditions and trap cells in them, preventing terminal differentiation (this idea is discussed in detail in [33,96]).

Such biological interpretation of the concepts presented here will require that these concepts be expanded to embrace the following aspects of dynamical biological systems that are currently not well understood but can already be framed:

Attractor Transitions and Development

If cell fates are attractors, then development is a flow in state space trajectories which then represent the developmental trajectories. But how do cells, e. g., a progenitor cell committing to a particular cell fate, move from one attractor to another? Currently, two models are being envisioned for the "flow between attractors" that constitutes developmental paths in the generation of the multi-cellular organisms:

(i) In one model, stimulated by the studies in Boolean networks, the cell "jumps" from one attractor to another in response to a distinct perturbation (e. g., developmental signal) which is represented as the imposed alteration of the expression status of a set of genes ("bit-flipping" in binary Boolean networks). This corresponds to the displacement of the network state $S(t)$ by an external influence to a new position in the basin of another attractor.

(ii) The second model posits that the landscape changes, and has its roots in the classical modeling of nonlinear dynamical systems: For instance, the attractor of the progenitor cell (valley) may be converted into an unstable steady-state point (hill top) at which point the network (cell) will spontaneously be attracted by either of the two attractors on each side of the newly formed hill. This is exemplified in a model for fate decision in bipotent progenitor cells in hematopoiesis [99]. In this model, a change of the landscape structure ensues from a change in the network architecture caused by the slow alteration of the value of a system parameter that controls the interaction strength in the system equations (see Sect. "Small Gene Circuits"). Thus, the external signal that triggers the differentiation exerts its effect by affecting the system parameters. Increasing the decay rates of x_1 and x_2 in the example of Fig. 3b will lead to the disappearance of the central attractor and convert the tristable landscape of Fig. 3b to the bistable one of Fig. 3a [99]. Such qualitative changes that occur during the slow increase or decrease of a system parameter are referred to as a ***bifurcation***. Note that this second model takes a narrower view, assuming that the network under study is not the global, universal and fixed network of the entire genome, as discussed in Sect. "Complex Networks", but rather a subnetwork that can change its architecture based on the presence or absence of gene products of genes that are outside of the subnetwork.

Noisy Systems

In a bistable switching system, it is immediately obvious that random fluctuations in x_i due to "gene expression noise" may trigger a transition between the attractors and thus, explain the stochastic phenotype transitions, as has been recently shown for several micro-organismal systems and discussed for mammalian cell differentiation [87,97, 101,108]. The observed heterogeneity of cells in a nominally identical cell population,caused either by "gene expression noise" or other diversifying processes, such as the random partitioning of molecules at cell divisions, implies that cell states cannot be viewed as deterministic points and trajectories in the state space, but rather as moving "clouds" – held together by the attractors.

More recently, on the basis of such bistable switches, it has even been proposed and shown in synthetic networks that environmental bias in fluctuation magnitude may explicitly control the switch to the physiologically favorable attractor state – because gene expression noise may be higher (relative to the deterministically controlled expression levels) when cells are in the attractor that dictates a gene expression pattern that is incompatible with a given environment [110]. Such biased gene expression noise thus may drive the flow between the attractors, and hence, (on average) guide the system through attractor transitions to produce various cell fates in a manner that is commensurate with development of the tissue. On the other hand, it may also explain the local stochasticity of cell fate decisions observed for many stem and progenitor cells.

Directionality

The concepts of cell fates and cell types as attractors, as well as any mechanisms that explains attractor transitions during development and differentiation, do not explain the overall directionality of development, or the "*arrow of time*" in ontogenesis: Why is development essentially a one-way process, i. e., irreversible in time, given that the underlying regulatory events, the switching ON or OFF of genes, are fully reversible? Where does the overall slope (from back to front) in Waddington's epigenetic landscape (Fig. 2g) come from? One idea is that "gene expression noise" may play the role of thermal noise (heat) in thermodynamics in explaining the irreversibility of processes. Alternatively, the network could be specifically wired, perhaps through natural selection so that attractor transitions are highly biased for one direction.

Beyond Cell Autonomous Processes

To understand development we need of course to open our view beyond the cell autonomous dynamics of GRNs. Some of the genes expressed in particular states $S(t)$ encode secreted proteins that affect the gene expression hence, state $S(t)$ of neighboring cells. Such inter-cell communication establishes a network at a higher level, with its own constrained dynamics that need to be incorporated in the models of developmental trajectories in gene expression state space.

Evolution

As discussed several times in this article, when networks are studied in the light of evolution, a central question arises: how do mutations, which essentially rewire the network by altering the nature of regulatory interactions, give rise to the particular architecture of the GRN that we find today? What are the relative roles in shaping particular network architectures, of (*i*) constraints due to physical (graph-theoretical) laws and self-organization vs. that of (*ii*) natural selection for functionality? If selection plays a major role, can it select for such features as the global landscape structure, or even for criticality? Or can the latter even "self-organize" without adaptive selection [31]? Such questions go far beyond the current analysis of Darwinian mechanisms of robustness and evolvability of networks. They are at the heart of the quest in biocomplexity research for fundamental principles of life, of which the process of natural selection itself is a just subset.

Regulatory networks offer a accessible and formalizable object of study to begin to ask these questions.

Bibliography

Primary Literature

1. Adolfsson J, Mansson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, Bryder D, Yang L, Borge OJ, Thoren LA, Anderson K, Sitnicka E, Sasaki Y, Sigvardsson M, Jacobsen SE (2005) Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell 121:295–306
2. Aittokallio T, Schwikowski B (2006) Graph-based methods for analysing networks in cell biology. Brief Bioinform 7:243–55
3. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118:4947–57
4. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. J Theor Biol 223:1–18
5. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–82
6. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci USA 100:8710–4
7. Aldana M, Balleza E, Kauffman S, Resendiz O (2007) Robustness and evolvability in genetic regulatory networks. J Theor Biol 245:433–48
8. Aldana M, Coppersmith S, Kadanoff LP (2003) Boolean dynamics with random couplings. In: Kaplan E, Marsden JE, Sreenivasan KR (eds) Perspectives and problems in nonlinear science. A celebratory volume in honor of Lawrence Sirovich. Springer, New York
9. Alon U (2003) Biological networks: the tinkerer as an engineer. Science 301:1866–7

10. Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97:11149–52
11. Anderson PW (1972) More is different. Science 177:393–396
12. Angeli D, Ferrell JE Jr., Sontag ED (2004) Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc Natl Acad Sci USA 101: 1822–1827
13. Arney KL, Fisher AG (2004) Epigenetic aspects of differentiation. J Cell Sci 117:4355–63
14. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L (2004) Comment on Network motifs: simple building blocks of complex networks and Superfamilies of evolved and designed networks. Science 305:1107; author reply 1107
15. Autumn K, Ryan MJ, Wake DB (2002) Integrating historical and mechanistic biology enhances the study of adaptation. Q Rev Biol 77:383–408
16. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14:283–91
17. Bagley RJ, Glass L (1996) Counting and classifying attractors in high dimensional dynamical systems. J Theor Biol 183:269–84
18. Balcan D, Kabakcioglu A, Mungan M, Erzan A (2007) The information coded in the yeast response elements accounts for most of the topological properties of its transcriptional regulation network. PLoS ONE 2:e501
19. Balleza E, Alvarez-Buylla ER, Chaos A, Kauffman A, Shmulevich I, Aldana M (2008) Critical dynamics in genetic regulatory networks: examples from four kingdoms. PLoS One 3:e2456
20. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–12
21. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113
22. Becskei A, Seraphin B, Serrano L (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. EMBO J 20:2528–2535
23. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4:51
24. Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2:E9
25. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125:315–26
26. Bird A (2007) Perceptions of epigenetics. Nature 447:396–8
27. Bloom JD, Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. BMC Evol Biol 3:21
28. Bloom JD, Adami C (2004) Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. BMC Evol Biol 4:14
29. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14:292–9
30. Bornholdt S (2005) Systems biology. Less is more in modeling large genetic networks. Science 310:449–51
31. Bornholdt S, Rohlf T (2000) Topological evolution of dynamical networks: global criticality from local dynamics. Phys Rev Lett 84:6114–7
32. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122:947–56
33. Brock A, Chang H, Huang SH Non-genetic cell heterogeneity and mutation-less tumor progression. Manuscript submitted
34. Brown KS, Hill CC, Calero GA, Myers CR, Lee KH, Sethna JP, Cerione RA (2004) The statistical mechanics of complex signaling networks: nerve growth factor signaling. Phys Biol 1: 184–195
35. Bulyk ML (2006) DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol 17:422–30
36. Callaway DS, Hopcroft JE, Kleinberg JM, Newman ME, Strogatz SH (2001) Are randomly grown graphs really random? Phys Rev E Stat Nonlin Soft Matter Phys 64:041902
37. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature 453:544–547
38. Chang HH, Oh PY, Ingber DE, Huang S (2006) Multistable and multistep dynamics in neutrophil differentiation. BMC Cell Biol 7:11
39. Chang WC, Li CW, Chen BS (2005) Quantitative inference of dynamic regulatory pathways via microarray data. BMC Bioinformatics 6:44
40. Chaves M, Sontag ED, Albert R (2006) Methods of robustness analysis for Boolean models of gene control networks. Syst Biol (Stevenage) 153:154–67
41. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8: 93–103
42. Chen KC, Wang TY, Tseng HH, Huang CY, Kao CY (2005) A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. Bioinformatics 21:2883–90
43. Chickarmane V, Troein C, Nuber UA, Sauro HM, Peterson C (2006) Transcriptional dynamics of the embryonic stem cell switch. PLoS Comput Biol 2:e123
44. Claverie JM (2001) Gene number. What if there are only 30,000 human genes? Science 291:1255–7
45. Collins SJ (1987) The HL-60 promyelocytic leukemia cell line: proliferation, differentiation, and cellular oncogene expression. Blood 70:1233–1244
46. Cordero OX, Hogeweg P (2006) Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23:1931–6
47. Cross MA, Enver T (1997) The lineage commitment of haemopoietic progenitor cells. Curr Opin Genet Dev 7: 609–613
48. Dang CV, O'Donnell KA, Zeller KI, Nguyen T, Osthus RC, Li F (2006) The c-Myc target gene network. Semin Cancer Biol 16:253–64
49. Davidich MI, Bornholdt S (2008) Boolean network model predicts cell cycle sequence of fission yeast. PLoS ONE 3:e1672
50. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. Science 311:796–800
51. de la Serna IL, Ohkawa Y, Berkes CA, Bergstrom DA, Dacwag CS, Tapscott SJ, Imbalzano AN (2005) MyoD targets chromatin

remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex. Mol Cell Biol 25:3997–4009

52. Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1:349–56
53. Delbrück M (1949) Discussion. In: Unités biologiques douées de continuité génétique Colloques Internationaux du Centre National de la Recherche Scientifique. CNRS, Paris
54. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, Walhout AJ (2006) A gene-centered C. elegans protein-DNA interaction network. Cell 125:1193–205
55. Derrida B, Pomeau Y (1986) Random networks of automata: a simple annealed approximation. Europhys Lett 1:45–49
56. Dodd IB, Micheelsen MA, Sneppen K, Thon G (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129:813–22
57. Eichler GS, Huang S, Ingber DE (2003) Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles. Bioinformatics 19:2321–2322
58. Eisenberg E, Levanon EY (2003) Preferential attachment in the protein network evolution. Phys Rev Lett 91:138701
59. Enver T, Heyworth CM, Dexter TM (1998) Do stem cells play dice? Blood 92:348–51; discussion 352
60. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16:2923–39
61. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5:e8
62. Faure A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics 22:e124–31
63. Fazi F, Rosa A, Fatica A, Gelmetti V, De Marchis ML, Nervi C, Bozzoni I (2005) A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. Cell 123:819–31
64. Ferrell JE Jr., Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280:895–8
65. Fisher AG (2002) Cellular identity and lineage choice. Nat Rev Immunol 2:977–82
66. Fox JJ, Hill CC (2001) From topology to dynamics in biochemical networks. Chaos 11:809–815
67. Fraser HB, Hirsh AE (2004) Evolutionary rate depends on number of protein–protein interactions independently of gene expression level. BMC Evol Biol 4:13
68. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296:750–2
69. Fuks F (2005) DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev 15: 490–495
70. Gao H, Falt S, Sandelin A, Gustafsson JA, Dahlman-Wright K (2007) Genome-wide identification of estrogen receptor α binding sites in mouse liver. Mol Endocrinol 22:10–22
71. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403:339–342
72. Gershenson C (2002) Classification of random Boolean networks. In: Standish RK, Bedau MA, Abbass HA (eds) Artificial life, vol 8. MIT Press, Cambridge, pp 1–8
73. Gisiger T (2001) Scale invariance in biology: coincidence or footprint of a universal mechanism? Biol Rev Camb Philos Soc 76:161–209
74. Glass L, Kauffman SA (1972) Co-operative components, spatial localization and oscillatory cellular dynamics. J Theor Biol 34:219–37
75. Goldberg AD, Allis CD, Bernstein E (2007) Epigenetics: a landscape takes shape. Cell 128:635–8
76. Goldstein ML, Morris SA, Yen GG (2004) Problems with fitting to the power-law distribution. Eur Phys J B 41:255–258
77. Goodwin BC, Webster GC (1999) Rethinking the origin of species by natural selection. Riv Biol 92:464–7
78. Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc Lond B Biol Sci 205:581–98
79. Graf T (2002) Differentiation plasticity of hematopoietic cells. Blood 99:3089–101
80. Grass JA, Boyer ME, Pal S, Wu J, Weiss MJ, Bresnick EH (2003) GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. Proc Natl Acad Sci USA 100:8811–6
81. Greil F, Drossel B, Sattler J (2007) Critical Kauffman networks under deterministic asynchronous update. New J Phys 9:373
82. Guelzim N, Bottani S, Bourgine P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 31:60–3
83. Guo Y, Eichler GS, Feng Y, Ingber DE, Huang S (2006) Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers. J Biomed Biotechnol 2006:69141
84. Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–6
85. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47–52
86. Harris SE, Sawhill BK, Wuensche A, Kauffman SA (2002) A model of transcriptional regulatory networks based on biases in the observed regulation rules. Complexity 7:23–40
87. Hasty J, Pradines J, Dolnik M, Collins JJ (2000) Noise-based switches and amplifiers for gene expression. Proc Natl Acad Sci USA 97:2075–80
88. Haverty PM, Hansen U, Weng Z (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acids Res 32:179–88
89. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 5:522–31
90. Hilborn R (1994) Chaos and nonlinear dynamics: An introduction for scientists and engineers, 2 edn. Oxford University Press, New York
91. Hochedlinger K, Jaenisch R (2006) Nuclear reprogramming and pluripotency. Nature 441:1061–7
92. Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C, Enver T (1997) Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev 11:774–85

93. Huang S (2004) Back to the biology in systems biology: what can we learn from biomolecular networks. Brief Funct Genomics Proteomics 2:279–297
94. Huang S (2007) Cell fates as attractors – stability and flexibility of cellular phenotype. In: Endothelial biomedicine, 1st edn, Cambridge University Press, New York, pp 1761–1779
95. Huang S, Ingber DE (2000) Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. Exp Cell Res 261:91–103
96. Huang S, Ingber DE (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Dis 26:27–54
97. Huang S, Wikswo J (2006) Dimensions of systems biology. Rev Physiol Biochem Pharmacol 157:81–104
98. Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. Phys Rev Lett 94:128701
99. Huang S, Guo YP, May G, Enver T (2007) Bifurcation dynamics of cell fate decision in bipotent progenitor cells. Dev Biol 305:695–713
100. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell 102:109–26
101. Hume DA (2000) Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. Blood 96:2323–8
102. Ihmels J, Bergmann S, Barkai N (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics 20:1993–2003
103. Jablonka E, Lamb MJ (2002) The changing concept of epigenetics. Ann N Y Acad Sci 981:82–96
104. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42
105. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genomewide mapping of in vivo protein-DNA interactions. Science 316:1497–502
106. Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1
107. Joy MP, Brock A, Ingber DE, Huang S (2005) High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol 2005:96–103
108. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6:451–64
109. Kaplan D, Glass L (1995) Understanding Nonlinear Dynamics, 1st edn. Springer, New York
110. Kashiwagi K, Urabe I, Kancko K, Yomo T (2006) Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. PLoS One, 1:e49
111. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224:177–8
112. Kauffman S (2004) A proposal for using the ensemble approach to understand genetic regulatory networks. J Theor Biol 230:581–90
113. Kauffman S, Peterson C, Samuelsson B, Troein C (2003) Random Boolean network models and the yeast transcriptional network. Proc Natl Acad Sci USA 100:14796–9
114. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467
115. Kauffman SA (1991) Antichaos and adaptation. Sci Am 265:78–84
116. Kauffman SA (1993) The origins of order. Oxford University Press, New York
117. Khorasanizadeh S (2004) The nucleosome: from genomic organization to genomic regulation. Cell 116:259–72
118. Kim KY, Wang J (2007) Potential energy landscape and robustness of a gene regulatory network: toggle switch. PLoS Comput Biol 3:e60
119. Klemm K, Bornholdt S (2005) Stable and unstable attractors in Boolean networks. Phys Rev E Stat Nonlin Soft Matter Phys 72:055101
120. Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. Proc Natl Acad Sci USA 101:1200–5
121. Kloster M, Tang C, Wingreen NS (2005) Finding regulatory modules through large-scale gene-expression data analysis. Bioinformatics 21:1172–9
122. Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705
123. Kramer BP, Fussenegger M (2005) Hysteresis in a synthetic mammalian gene network. Proc Natl Acad Sci USA 102: 9517–9522
124. Krawitz P, Shmulevich I (2007) Basin entropy in Boolean network ensembles. Phys Rev Lett 98:158701
125. Krysinska H, Hoogenkamp M, Ingram R, Wilson N, Tagoh H, Laslo P, Singh H, Bonifer C (2007) A two-step, PU.1-dependent mechanism for developmentally regulated chromatin remodeling and transcription of the c-fms gene. Mol Cell Biol 27:878–87
126. Kubicek S, Jenuwein T (2004) A crack in histone lysine methylation. Cell 119:903–6
127. Laslo P, Spooner CJ, Warmflash A, Lancki DW, Lee HJ, Sciammas R, Gantner BN, Dinner AR, Singh H (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. Cell 126:755–66
128. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298:799–804
129. Levsky JM, Singer RH (2003) Gene expression and the myth of the average cell. Trends Cell Biol 13:4–6
130. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci USA 101:4781–6
131. Li H, Xuan J, Wang Y, Zhan M (2008) Inferring regulatory networks. Front Biosci 13:263–75
132. Lim HN, van Oudenaarden A (2007) A multistep epigenetic switch enables the stable inheritance of DNA methylation states. Nat Genet 39:269–75
133. Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH (2007) Modular organization of protein interaction networks. Bioinformatics 23:207–14

134. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431:308–12
135. MacCarthy T, Pomiankowski A, Seymour R (2005) Using large-scale perturbations in gene network reconstruction. BMC Bioinformatics 6:11
136. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA 100:11980–5
137. Manke T, Demetrius L, Vingron M (2006) An entropic characterization of protein interaction networks and cellular robustness. JR Soc Interface 3:843–50
138. Marcotte EM (2001) The path not taken. Nat Biotechnol 19:626–627
139. Margolin AA, Califano A (2007) Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci 1115:51–72
140. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–3
141. Mattick JS (2007) A new paradigm for developmental biology. J Exp Biol 210:1526–47
142. May RM (1972) Will a large complex system be stable? Nature 238:413–414
143. Meissner A, Wernig M, Jaenisch R (2007) Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nat Biotechnol 25:1177–1181
144. Mellor J (2006) Dynamic nucleosomes and gene transcription. Trends Genet 22:320–9
145. Metzger E, Wissmann M, Schule R (2006) Histone demethylation and androgen-dependent transcription. Curr Opin Genet Dev 16:513–7
146. Mikkers H, Frisen J (2005) Deconstructing stemness. Embo J 24:2715–9
147. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–7
148. Monod J, Jacob F (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harb Symp Quant Biol 26:389–401
149. Morceau F, Schnekenburger M, Dicato M, Diederich M (2004) GATA-1: friends, brothers, and coworkers. Ann N Y Acad Sci 1030:537–54
150. Morrison SJ, Uchida N, Weissman IL (1995) The biology of hematopoietic stem cells. Annu Rev Cell Dev Biol 11:35–71
151. Murray JD (1989) Mathematical biology, 2nd edn (1993). Springer, Berlin
152. Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45:167–256
153. Nykter M, Price ND, Aldana M, Ramsey SA, Kauffman SA, Hood L, Yli-Harja O, Shmulevich I (2008) Gene expression dynamics in the macrophage exhibit criticality. Proc Natl Acad Sci USA 105:1897–900
154. Nykter M, Price ND, Larjo A, Aho T, Kauffman SA, Yli-Harja O, Shmulevich I (2008) Critical networks exhibit maximal information diversity in structure-dynamics relationships. Phys Rev Lett 100:058702
155. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science 303:1378–81
156. Okita K, Ichisaka T, Yamanaka S (2007) Generation of germline-competent induced pluripotent stem cells. Nature 448:313–7
157. Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A (2004) Multistability in the lactose utilization network of Escherichia coli. Nature 427:737–740
158. Pennisi E (2003) Human genome. A low number wins the GeneSweep Pool. Science 300:1484
159. Picht P (1969) Mut zur utopie. Piper, München
160. Proulx SR, Promislow DE, Phillips PC (2005) Network thinking in ecology and evolution. Trends Ecol Evol 20:345–53
161. Raff M (2003) Adult stem cell plasticity: fact or artifact? Annu Rev Cell Dev Biol 19:1–22
162. Ralston A and Rossant J (2005) Genetic regulation of stem cell origins in the mouse embryo. Clin Genet 68:106–12
163. Ramo P, Kesseli J, Yli-Harja O (2006) Perturbation avalanches and criticality in gene regulatory networks. J Theor Biol 242:164–70
164. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–5
165. Reik W, Dean W (2002) Back to the beginning. Nature 420:127
166. Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, Gutierrez-Rios RM, Martinez-Antonio A, Avila-Sanchez C, Collado-Vides J (2005) Modular analysis of the transcriptional regulatory network of E. coli. Trends Genet 21:16–20
167. Robins H, Krasnitz M, Barak H, Levine AJ (2005) A relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. J Bacteriol 187:8370–4
168. Roeder I, Glauche I (2006) Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. J Theor Biol 241:852–65
169. Salgado H, Santos-Zavaleta A, Gama-Castro S, Peralta-Gil M, Penaloza-Spinola MI, Martinez-Antonio A, Karp PD, Collado-Vides J (2006) The comprehensive updated regulatory network of Escherichia coli K-12. BMC Bioinformatics 7:5
170. Samonte RV, Eichler EE (2002) Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3:65–72
171. Sandberg R, Ernberg I (2005) Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). Proc Natl Acad Sci USA 102:2052–7
172. Shivdasani RA (2006) MicroRNAs: regulators of gene expression and cell differentiation. Blood 108:3646–53
173. Shmulevich I, Kauffman SA (2004) Activities and sensitivities in boolean network models. Phys Rev Lett 93:048701
174. Shmulevich I, Kauffman SA, Aldana M (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc Natl Acad Sci USA 102:13439–44
175. Siegal ML, Promislow DE, Bergman A (2007) Functional and evolutionary inference in gene networks: does topology matter? Genetica 129:83–103
176. Smith MC, Sumner ER, Avery SV (2007) Glutathione and Gts1p drive beneficial variability in the cadmium resistances of individual yeast cells. Mol Microbiol 66:699–712
177. Southall TD, Brand AH (2007) Chromatin profiling in model organisms. Brief Funct Genomic Proteomic 6:133–40
178. Southan C (2004) Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics 4:1712–26

179. Stern CD (2000) Conrad H. Waddington's contributions to avian and mammalian development, 1930–1940. Int J Dev Biol 44:15–22
180. Strohman R (1994) Epigenesis: the missing beat in biotechnology? Biotechnology (N Y) 12:156–64
181. Stumpf MP, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc Natl Acad Sci USA 102:4221–4
182. Suzuki M, Yamada T, Kihara-Negishi F, Sakurai T, Hara E, Tenen DG, Hozumi N, Oikawa T (2006) Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. Oncogene 25:2477–88
183. Swiers G, Patient R, Loose M (2006) Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. Dev Biol 294:525–40
184. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–76
185. Tapscott SJ (2005) The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development 132:2685–95
186. Taylor JS, Raes J (2004) Duplication and divergence: The evolution of new genes and old ideas. Annu Rev Genet 38: 615–643
187. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36:492–6
188. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 20:433–40
189. Tinbergen N (1952) Derived activities; their causation, biological significance, origin, and emancipation during evolution. Q Rev Biol 27:1–32
190. Toh H, Horimoto K (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics 18:287–97
191. Trojer P, Reinberg D (2006) Histone lysine demethylases and their impact on epigenetics. Cell 125:213–7
192. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
193. van Helden J, Wernisch L, Gilbert D, Wodak SJ (2002) Graph-based analysis of metabolic networks. Ernst Schering Res Found Workshop:245–74
194. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19:479–84
195. Vogel G (2003) Stem cells. 'Stemness' genes still elusive. Science 302:371
196. Waddington CH (1940) Organisers and genes. Cambridge University Press, Cambridge
197. Waddington CH (1956) Principles of embryology. Allen and Unwin Ltd, London
198. Waddington CH (1957) The strategy of the genes. Allen and Unwin, London
199. Watts DJ (2004) The "new" science of networks. Ann Rev Sociol 20:243–270
200. Webster G, Goodwin BC (1999) A structuralist approach to morphology. Riv Biol 92:495–8
201. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature 448:318–24
202. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13:1977–2000
203. Wilkins AS (2007) Colloquium Papers: Between "design" and "bricolage": Genetic networks, levels of selection, and adaptive evolution. Proc Natl Acad Sci USA 104 Suppl 1:8590–6
204. Wuensche A (1998) Genomic regulation modeled as a network with basins of attraction. Pac Symp Biocomput:89–102
205. Xiong W, Ferrell JE Jr. (2003) A positive-feedback-based bistable 'memory module' that governs a cell fate decision. Nature 426:460–465
206. Xu X, Wang L, Ding D (2004) Learning module networks from genome-wide location and expression data. FEBS Lett 578:297–304
207. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. Trends Genet 20:227–31
208. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3:e59
209. Yuh CH, Bolouri H, Davidson EH (2001) Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development 128:617–29

Books and Reviews

Huang S (2004) Back to the biology in systems biology: what can we learn from biomolecular networks. Brief Funct Genomics Proteomics 2:279–297
Huang S (2007) Cell fates as attractors – stability and flexibility of cellular phenotype. In: Endothelial biomedicine, 1st edn. Cambridge University Press, New York, pp 1761–1779
Huang S, Ingber DE (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Dis 26:27–54
Kaneko K (2006) Life: An introduction to complex systems biology, 1edn. Springer, Berlin
Kauffman SA (1991) Antichaos and adaptation. Sci Am 265:78–84
Kauffman SA (1993) The origins of order. Oxford University Press, New York
Kauffman SA (1996) At home in the universe: the search for the laws of self-organization and complexity. Oxford University Press, New York
Laurent M, Kellershohn N (1999) Multistability: a major means of differentiation and evolution in biological systems. Trends Biochem Sci 24:418–422
Wilkins AS (2007) Colloquium papers: Between "design" and "bricolage": Genetic networks, levels of selection, and adaptive evolution. Proc Natl Acad Sci USA 104 Suppl 1:8590–6

Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Introduction to

William H. K. Lee
US Geological Survey (Retired), Menlo Park, USA

Article Outline

This Introduction is intended to serve as a 'road map' for readers to navigate through the 42 Encyclopedia articles on earthquakes, tsunamis, and volcanoes. Selecting the topics and authors was somewhat subjective, as it is not possible to cover the vast existing literature with only 42 articles. They are, however, representative of the wide range of problems investigated in connection with these natural phenomena. I will introduce these articles by grouping them into sections and then into subsections. However, some articles belong to more than one section or one subsection, reflecting the inter-related nature of earthquakes, tsunamis and volcanoes. For the benefit of the readers, I will point to certain issues discussed in some of the articles which, in my view, have not been settled completely. I have also taken the liberty of quoting or paraphrasing sentences from many of these articles when introducing them, but I do not claim to be accurate. It is best for these articles to speak for themselves.

I wish to thank Bernard Chouet for helping me in planning and reviewing the manuscripts of the volcanoes section. I am grateful to Bernard Chouet, Edo Nyland, Jose Pujol, Chris Stephens, and Ta-liang Teng for their helpful comments that greatly improved this manuscript.

Introduction

Earthquakes, tsunamis, and volcanic eruptions are complex and often inter-related natural phenomena with disastrous impact to society rivaling those caused by the worst floods or storms. The 1556 Huaxian earthquake in the Shansi province of China claimed over 830,000 lives. The total economic loss of the 1995 Kobe earthquake in Japan was estimated at US $200 billion. The 2004 Indian Ocean tsunami (triggered by the Sumatra–Andaman earthquake of December 26) bought devastation thousands of miles away with fatalities exceeding 280,000. The 79 AD eruption of Mount Vesuvius near Naples, Italy buried the towns of Pompeii and Herculaneum. The 1902 eruption of Mount Pelée, Martinique, totally destroyed the town of St. Pierre.

Insurance companies classify major natural catastrophes as storms, floods, or earthquakes (including tsunamis, and volcanic eruptions). Since 1950, about 2.5 million people have died due to these catastrophes and overall economic losses have totaled about US $2 trillion in current dollar values. Earthquakes, tsunamis, and volcanic eruptions have accounted for about half of the fatalities and more than one third of the total economic losses. Geoscientists have attempted to predict these events, but with limited success. There are many reasons for such slow progress: (1) systematic monitoring of earthquakes, tsunamis and volcanoes requires large capital investment for instruments and very long-term support for operation and maintenance; (2) catastrophic earthquakes, tsunamis and volcanic eruptions occur rarely, and (3) politicians and citizens are quick to forget these hazards in the face of other more frequent and pressing issues. But with continuing rapid population growth and urbanization, the loss potential from these natural hazards in the world is quickly escalating.

With advances in nonlinear dynamics and complexity studies, geoscientists have applied modern nonlinear techniques and concepts such as chaos, fractal, critical phenomena, and self-organized criticality to the study of earthquakes, tsunamis and volcanoes. Here we sample these efforts, mainly in seismicity modeling for earthquake prediction and forecast, along with articles that review recent progress in studying earthquakes, tsunamis, and volcanoes. Although predictability is desirable, it is also possible to reduce these natural hazards with more practical approaches, such as early warning systems, hazard analysis, engineering considerations, and other mitigation efforts. Several articles in this Encyclopedia discuss these practical solutions.

Earthquakes

When a sudden rupture occurs in the Earth, seismic waves are generated. When these waves reach the Earth's surface, we may feel them as a series of vibrations, which we call an earthquake. Instrumental recordings of earthquakes have been made since the latter part of the 19th century by seismographic stations and networks from local to global scales. The observed data have been used, for example, (1) to compute the source parameters of earthquakes, (2) to determine the physical properties of the Earth's interior, (3) to test the theory of plate tectonics, (4) to map active faults, (5) to infer the nature of damaging ground shaking, (6) to carry out seismic hazard analyzes, and (7) to predict and forecast earthquakes. A satisfactory theory of the complex earthquake process has not yet been achieved, and realistic equations for modeling earthquakes do not exist at present. There is, however, good progress towards a physical foundation for the earthquake source process, partly as a result of research directed toward earthquake prediction.

Earthquake Monitoring, and Probing the Earth's Interior

Earthquakes are complex natural phenomena, and their monitoring requires an interdisciplinary approach, including using tools from other scientific disciplines and engineering. In ► Earthquake Monitoring and Early Warning Systems, W.H.K. Lee and Y.M. Wu presented a summary of earthquake monitoring, a description of the products derived from the analysis of seismograms, and a discussion of the limitations of these products. The basic results of earthquake monitoring are summarized in earthquake catalogs, which are lists of origin time, hypocenter location, and magnitude of earthquakes, as well as other source parameters. Lee and Wu describe the traditional earthquake location method formulated as an inverse problem. In ► Earthquake Location, Direct, Global-Search Methods, Lomax et al. review a different approach using direct-search over a space of possible locations, and discuss other related algorithms. Direct-search earthquake location is important because, relative to the traditional linearized method, it is both easier to apply to more realistic Earth models and is computational more stable. Although it has not been widely applied because of its computational demand, it shows great promise for the future as computer power is advancing rapidly.

The most frequently determined parameter after 'location' is 'magnitude', which is used to characterize the 'size' of an earthquake. A brief introduction to the quantification of earthquake size, including magnitude and seismic moment, is given in ► Earthquake Monitoring and Early Warning Systems by Lee and Wu. Despite its various limitations, magnitude provides important information concerning the earthquake source. Magnitude values have an immense practical value for realistic long-term disaster preparedness and risk mitigation efforts. A detailed review, including current practices for magnitude determinations, appears in ► Earthquake Magnitude by P. Bormann and J. Saul.

Besides computing earthquake source parameters, earthquake monitoring also provides data that can be used to probe the Earth's interior. In ► Tomography, Seismic, J. Pujol reviews a number of techniques designed to investigate the interior of the Earth using arrival times and/or waveforms from natural and artificial sources. The most common product of a tomographic study is a seismic velocity model, although other parameters, such as attenuation and anisotropy, can also be estimated. Seismic tomography generally has higher resolution than that provided by other geophysical methods, such as gravity and magnetics, and furnishes information (1) about fundamental problems concerning the internal structure of the Earth on a global scale, and (2) for tectonic and seismic hazard studies on a local scale.

In ► Seismic Wave Propagation in Media with Complex Geometries, Simulation of, H. Igel et al. present the state-of-the-art in computational wave propagation. They point to future developments, particularly in connection with the search for efficient generation of computational grids for models with complex topography and faults, as well as for the combined simulation of soil-structure interactions. In addition to imaging subsurface structure and earthquake sources, 3-D wave simulations can forecast strong ground motions from large earthquakes. In the absence of deterministic prediction of earthquakes, the calculation of earthquake scenarios in regions with sufficiently well-known crustal structures and faults will play an important role in assessing and mitigating potential damage, particularly those due to local site effects.

In addition to the classical parametrization of the Earth as a layered structure with smooth velocity perturbation, a new approach using scattered waves that reflect Earth's heterogeneity is introduced by H. Sato in his article on ► Seismic Waves in Heterogeneous Earth, Scattering of. For high-frequency seismograms, envelope characteristics such as the excitation level and the decay gradient of coda envelopes and the envelope broadening of the direct wavelet are useful for the study of small-scale inhomogeneities within the Earth. The radiative transfer theory with scattering coefficients calculated from the Born approximation and the Markov approximation for the parabolic wave equation are powerful mathematical tools for these analyzes. Studies of the scattering of high-frequency seismic waves in the heterogeneous Earth are important for understanding the physical structure and the geodynamic processes that reflect the evolution of the solid Earth.

Earthquake Prediction and Forecasting

A fundamental question in earthquake science is whether earthquake prediction is possible. Debate on this question has been going on for decades without clear resolution. Are pure observational methods without specific physical understanding sufficient? Earthquakes have been instrumentally monitored continuously for about 100 years (although not uniformly over the Earth), but reliable and detailed earthquake catalogs cover only about 50 years. Consequently, it seems questionable that earthquakes can be predicted solely on the basis of observed seismicity patterns, given that large earthquakes in a given region have recurrence intervals ranging from decades to centuries or longer. Despite progress made in earthquake physics, we are still not able to write down all the governing equations for these events and lack sufficient information about the Earth's properties. Nevertheless, many attempts have been and are being made to predict and forecast earthquakes. In this section, several articles based on empirical and physics-based approaches will be briefly introduced.

In ▶ Geo-complexity and Earthquake Prediction, V. Keilis-Borok et al. present an algorithmic prediction method for individual extreme events having low probability but large societal impact. They show that the earthquake prediction problem is necessarily intertwined with problems of disaster preparedness, the dynamics of the solid Earth, and the modeling of extreme events in hierarchical complex systems. The algorithms considered by Keilis-Borok et al. are based on premonitory seismicity patterns and provide alarms lasting months to years. Since the 1990s, these alarms have been posted for use in testing such algorithms against newly occurred large earthquakes. Some success has been achieved, and although the areas for the predicted earthquakes are very large and the predicted time windows are very long, such predictions can be helpful for the officials and the public to undertake appropriate preparedness.

Stochastic models are a practical way of bridging the gap between the detailed modeling of a complex system and the need to fit models to limited data. In ▶ Earthquake Occurrence and Mechanisms, Stochastic Models for, D. Vere-Jones presents a brief account of the role and development of stochastic models of seismicity, from the first empirical studies to current models used in earthquake probability forecasting. The author combines a model of the physical processes generating the observable data (earthquake catalogs) with a model for the errors, or uncertainties, in our ability to predict those observables.

D.A. Yuen et al. propose the use of statistical approaches and data-assimilation techniques to earthquake forecasting in their article on ▶ Earthquake Clusters over Multi-dimensional Space, Visualization of. The nature of the spatial-temporal evolution of earthquakes may be assessed from the observed seismicity and geodetic measurements by recognizing nonlinear patterns hidden in the vast amount of seemingly unrelated data. The authors endeavor to bring across the basic concept of clustering and its role in earthquake forecasting, and conclude that the clustering of seismic activity reflects both the similarity between clusters and their correlation properties.

In ▶ Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, G. Zoeller et al. present a combined approach to understanding seismicity and the emergence of patterns in the occurrence of earthquakes based on numerical modeling and data analysis. The discussion and interpretation of seismicity in terms of statistical physics leads to the concept of 'critical states', i. e. states in the seismic cycle with an increased probability for abrupt changes involving large earthquakes. They demonstrate that numerical fault models are valuable for understanding the underlying mechanisms of observed seismicity patterns, as well as for practical estimates of future seismic hazard.

D. Sornette and M.J. Werner in ▶ Seismicity, Statistical Physics Approaches to stress that the term 'statistical' in 'statistical physics' has a different meaning than as used in 'statistical seismology'. Statistical seismology has been developed as a marriage between probability theory, statistics, and earthquake occurrences without considerations of earthquake physics. In statistical physics approaches to seismicity, researchers strive to derive statistical models from microscopic laws of friction, damage, rupture, etc. Sornette and Werner summarize some of the concepts and tools that have been developed, including the leading theoretical physical models of the space-time organization of earthquakes. They then present several examples of the new metrics proposed by statistical physicists, underlining their strengths and weaknesses. They conclude that a holistic approach emphasizing the interactions between earthquakes and faults is promising, and that statistical seismology needs to evolve into a genuine physically-based statistical physics of earthquakes.

In ▶ Earthquake Networks, Complex, S. Abe and N. Suzuki discuss the construction of a complex earthquake network obtained by mapping seismic data to a growing stochastic graph. This graph, or network, turns out to exhibit a number of remarkable physical and mathematical behaviors that share common traits with many other complex systems. The scale-free and small-world natures are typical examples in complex earthquake networks.

Electromagnetic phenomena associated with earthquakes, such as earthquake light have been reported throughout almost all human history. Until rather recently, however, most such observations were unreliable and best described as folklore. In ▶ Earthquakes, Electromagnetic Signals of, S. Uyeda et al. summarize the scientific search for electromagnetic precursors for earthquake prediction. The presumption is that since earthquakes occur when slowly increasing tectonic stress in the Earth's crust reaches a critical level; the same stress may give rise to some electromagnetic phenomena. Research on possible relationships was initiated in several countries around the world in the 1980s. Two main approaches are (1) the monitoring of possible emissions from focal regions in a wide range of frequency from DC to VHF, and (2) the monitoring of possible anomalies in the transmission of man-made electromagnetic waves of various frequencies over focal regions. Despite much circumstantial evidence, earthquake-related electromagnetic signals, in particular those at a pre-seismic stage are not yet widely accepted to be associated with earthquakes.

Observational programs focused on searching for reliable precursory phenomena in seismicity, seismic velocities, tilt and strain, electromagnetic signals, chemical emissions and animal behavior, claim some successes but no systematic precursors have been identified. In ▶ Earthquake Forecasting and Verification, J.R. Holliday et al. stress that reliable earthquake forecasting will require systematic verification. They point out that although earthquakes are complex phenomena, systematic scaling laws such as the Gutenberg–Richter frequency-magnitude relation have been recognized. The Gutenberg–Richter relation is given by: $\log N(M) = a - bM$, where M is the earthquake magnitude, $N(M)$ is the number of earthquakes with magnitude greater than or equal to M, and a and b are constants. Since $b \approx 1$, this means that the number of earthquakes increase tenfold for each decrease of one magnitude unit. This suggests that *large* earthquakes occur in regions where there are large numbers of *small* earthquakes. On this basis, the regions where large earthquakes will occur can be forecast with considerable accuracy, but the Gutenberg–Richter relation provides no information about the precise occurrence times.

Earthquake Engineering Considerations and Early Warning Systems

Since seismic hazards exist in many regions of the world, three major strategies are introduced to reduce their societal impacts: (1) to avoid building in high seismic-risk areas, (2) to build structures that can withstand the effects of earthquakes, and (3) to plan for earthquake emergencies. The first strategy is not very practical because, with rapid population growth, many economically productive activities are increasingly located in high seismic-risk areas. However, by mapping active faults and by studying past earthquakes, we may estimate the risk potential from earthquakes and plan our land use accordingly. The second strategy depends on the skills of engineers, and also requires seismologists to provide realistic estimates of the ground motions resulting from expected earthquakes. The third strategy includes attempting to predict earthquakes reliably well in advance to minimize damage and casualties, and also requires the cooperation of the entire society. Although we are far from being able to predict earthquakes reliably, earthquake early warning systems can provide critical information to reduce damage and causalities, as well as to aid rescuing and recovery efforts.

Accurate prediction of the level and variability of near-source strong-ground motions in future earthquakes is one of the key challenges facing seismologists and earthquake engineers. The increasing number of near-source recordings collected by dense strong-motion networks exemplifies the inherent complexity of near-field ground shaking, which is governed by a number of interacting physical processes. Characterizing, quantifying, and modeling ground-motion complexity requires a joint investigation of (1) the physics of earthquake rupture, (2) wave-propagation in heterogeneous media, and (3) the effects of local site conditions. In ▶ Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, P.M. Mai discusses briefly the beginnings of strong-motion seismology and the recognition of ground-motion complexity. Using two well-recorded recent earthquakes, the author introduces the observational aspects of near-field ground shaking and describes the basic mathematical tools used in the computation of ground motion. The key elements for characterizing and modeling ground-motion complexity are also explained, supplemented by a concise overview of the underlying physical processes.

With increasing urbanization worldwide, earthquake hazards pose ever greater threats to lives, property, and livelihoods in populated areas near major active faults on land or near offshore subduction zones. Earthquake early-warning (EEW) systems can be useful tools for reducing the impact of earthquakes, provided that the populated areas are favorably located with respect to earthquake sources and their citizens are properly trained to respond to the warning messages. Under favorable conditions, an EEW system can forewarn an urban population of impending strong shaking with lead times that range from a few seconds to a few tens of seconds. A lead time is the time interval between issuing a warning and the arrival of the *S*- and surface waves, which are the most destructive due to their large amplitudes. Even a few seconds of advance warning is useful for pre-programmed emergency measures at various critical facilities, such as the deceleration of rapid-transit vehicles and high-speed trains, the orderly shutoff of gas pipelines, the controlled shutdown of some high-technological manufacturing operations, the safe-guarding of computer facilities, and bringing elevators to a stop at the nearest floor.

Recent advances in early warning methodologies are summarized by W.H.K Lee and Y.M. Wu in the second part of their article, ► Earthquake Monitoring and Early Warning Systems. In ► Earthquake Early Warning System in Southern Italy, A. Zollo et al. analyze and illustrate the main scientific and technological issues related to the implementation and management of the earthquake early warning system under development in the Campania region of southern Italy. The system is designed to issue alerts to distant coastal targets using data from a dense seismic network deployed in the Apennine belt region. The authors note that earthquake early warning systems can also help mitigate the effects of earthquake-induced disasters such as fires, explosions, landslides, and tsunamis. Earthquake early warning systems can be installed at relatively low cost in developing countries, where even moderate-size earthquakes can cause damage comparable to that caused by much larger earthquakes in developed countries.

Nonlinear problems in structural earthquake engineering deal with the dynamic response of meta-stable, man-made buildings subjected to strong earthquake shaking. During earthquakes, structures constructed on soft sediments and soils deform together with the underlying soil. Strong shaking forces the soil-structure systems to evolve through different levels of nonlinear response, with continuously changing properties that depend upon the time history of excitation and on the progression and degree of damage. In ► Earthquake Engineering, Non-linear Problems in, M.D. Trifunac first briefly discuss the literature on complex and chaotic dynamics of simple mechanical oscillators, and then introduces the dynamic characteristics and governing equations of the meta-stable structural dynamics in earthquake engineering. He describes the nature of the solutions of the governing equations in terms of both the vibrational and the wave representations. The author also addresses the dynamic instability, material and geometric nonlinearities, and complexities of the governing equations associated with nonlinear soil-structure interaction.

Structural health monitoring and structural damage detection refers to the processes of determining and tracking the structural integrity and assessing the nature of damage in a structure. An important and challenging problem is being able to detect the principal components of damage in structures (as they occur during or soon after the earthquake) before physical inspection. In the article, ► Earthquake Damage: Detection and Early Warning in Man-Made Structures, M.I. Todorovska focuses on global methods and intermediate-scale methods, which can point to the parts of the structure that have been damaged. Recently, structural identification and health monitoring of buildings based on detecting changes in wave travel time through the structure has received renewed attention and has proven to be very promising.

Earthquake Physics

Brittle deformation, which is the primary mode of deformation of the Earth's crust in response to tectonic stress, is manifested by faulting at the long timescale and by earthquakes at the short timescale. It is one of the best-known examples of a system exhibiting self-organized criticality. A full understanding of this system is essential for evaluating earthquake hazards, but our current understanding is sketchy. In ► Brittle Tectonics: A Non-linear Dynamical System, C.H. Scholz shows that an earthquake dynamic system has two characteristic length scales, W^* and W^{**}. An earthquake nucleates within the seismogenic zone and initially propagates in all directions along its perimeter, acting as a 3D crack. When its dimension exceeds W^*, the rupture is restricted to propagating in the horizontal direction, acting as a 2D crack. Thus a symmetry breakage occurs at the dimension W^*. *Small* earthquakes, with dimensions smaller than W^*, are not self-similar with *large* earthquakes, those with lengths larger than W^*. The same occurs for suprafaults at the dimension W^{**} (a suprafault is the shear relaxation structure that includes a fault and its associated ductile shear zone).

Earthquake prediction is desirable for reducing seismic hazards, but we lack an understanding of how and why earthquakes begin and grow larger or stop. Theoretical and laboratory studies show that a quasi-static rupture growth precedes dynamic rupture. Thus, detecting the quasi-static rupture growth may lead to forecasting the subsequent dynamic rupture. In ► Earthquake Nucleation Process, Y. Iio reviews studies that analyze the early portions of observed waveforms, and summarizes what we presently understand about earthquake nucleation process. An earthquake initiates over a small patch of a fault, and then their rupture fronts expand outward until they stop. Some large earthquakes have a rupture extent greater than 1000 km, while fault lengths of small microearthquakes range over only a few meters. Surprisingly, the concept that earthquakes are self-similar is widely accepted despite fault length ranging over 6 orders of magnitude. One example of such similarity is the proportionality of average fault slip to fault length, which implies a constant static stress drop, independent of earthquake size.

The self-similarity law raises a fundamental question, namely what is the difference between large and small earthquakes? One end-member model represents earthquakes as ruptures that grow randomly and then terminate at an earlier stage for smaller earthquakes, but continue longer for larger earthquakes. This type of model has been proposed mainly to explain the frequency–magnitude distribution of earthquakes (the Gutenberg–Richter relation), but it implies that it is impossible to forecast the final size of an earthquake at the time the rupture initiates. However, the other end-member model predicts that larger earthquakes have a larger 'seed' than smaller earthquakes, and that large and small earthquakes are different even at their beginnings.

Geoscientists have long sought an understanding of how earthquakes interact. Can earthquakes trigger other earthquakes? The answer is clearly yes over short time and distance scales, as in the case of mainshock–aftershock sequences. Over increasing time and distance scales, however, this question is more difficult to answer. In ▶ Earthquakes, Dynamic Triggering of, S.G. Prejean and D.P. Hill explore the most distant regions over which earthquakes can trigger other earthquakes. This subject has been the focus of extensive research over the past twenty five years, and offers a potentially important key to improving our understanding of earthquake nucleation. In this review, the authors discuss physical models and give a description of documented patterns of remote dynamic triggering.

Models of the earthquake source have been successfully used in predicting many of the general properties of seismic waves radiated from earthquakes. These general properties can be derived from a simple omega-squared spectral shape. In ▶ Earthquake Scaling Laws, R. Madariaga derives general expressions for energy, moment and stress in terms of measured spectral parameters, and shows that earthquake sources can be reduced to a single family with the three parameters of moment, corner frequency and radiated energy. He suggests that most of the properties of the seismic spectrum and slip distribution can be explained by a simple crack model. Whether an earthquake is modeled as a simple circular crack or as a complex distribution of such cracks, the result is the same.

In ▶ Earthquake Source: Asymmetry and Rotation Effects, R. Teisseyre presents a consistent theory describing an elastic continuum subjected to complex internal processes, considers all the possible kinds of the point-related motions and deformations, and defines a complex rotation field including spin and twist. Also included in the discussion is a new description of the source processes, including the role of rotation in source dynamics, an explanation of co-action of the slip and rotation motions, and a theory of seismic rotation waves. Rotational seismology is an emerging field, and a progress report is provided in the Appendix in ▶ Earthquake Monitoring and Early Warning Systems by W.H.K. Lee and Y.M. Wu.

Some New Tools to Study Earthquakes

The Global Positioning System (GPS) is a space-based Global Navigation Satellite System. Using signals transmitted by a constellation of GPS satellites, the positions of ground-based receivers can be calculated to high precision, making it possible to track relative movements of points on the Earth's surface over time. Unlike older geodetic surveying methods (which involved periodically but infrequent measuring angles, distances, or elevations between points), GPS can provide precise 3-D positions over a range of sampling rates and on a global scale. GPS equipment is easy to use and can be set up to collect data continuously. Since its early geophysical applications in the mid-1980s, this versatile tool, which can be used to track displacements over time periods of seconds to decades, has become indispensable for crustal deformation studies, leading to many important insights and some surprising discoveries. In ▶ GPS: Applications in Crustal Deformation Monitoring, J. Murray-Moraleda focuses on applications of GPS data to the studies of tectonic, seismic, and volcanic processes. The author presents an overview of how GPS works and how it is used to collect data for geophysical studies. The article also describes a variety of ways in which GPS data have been used to measure crustal deformation and investigate the underlying processes.

The concept of a seismic cycle involves processes associated with the accumulation and release of stress on seismogenic faults, and is commonly divided into three intervals: (1) the coseismic interval for events occurring during an earthquake, (2) the postseismic interval immediately following an earthquake, and (3) the interseismic period in between large earthquakes. In ▶ Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, R. Lohman explores how we can draw conclusions about fault zone slip at depths far greater than are directly accessible to us, based on how the Earth's surface deforms during, before, and after earthquakes.

Atmospheric sound can be radiated by the displacement or rupture of the Earth's surface induced by earthquakes, tsunamis, and volcanoes, and by the flow and excitation of fluids during volcanic eruptions. These complex and potentially cataclysmic phenomena share some common physics, yet represent different ways of converting energy into atmospheric sound. In ▶ Infrasound from Earthquakes, Tsunamis and Volcanoes, M. Garces and A. LePichon discuss some of the signal features unique to earthquakes, tsunamis, and volcanoes captured by the present generation of infrasound arrays. They also discuss contemporary methods for the analysis, interpretation, and modeling of these diverse signals, and consider some of the associated geophysical problems that remain unsolved.

Tsunamis

Tsunamis are oceanic gravity waves generated by seafloor deformation due to earthquakes, volcanic eruptions, landslides, or asteroid impacts. Earthquake tsunamis, such as the 2004 Indian Ocean tsunami (caused by the Sumatra–Andaman earthquake of December 26), are the most frequent type of tsunamis. However, large volcanic eruptions, such as the 1883 Krakatau eruption (in the Sunda strait between the islands of Java and Sumatra) also cause ocean-wide tsunamis. Landslides (which are often triggered by earthquakes) cause large tsunamis locally, but their effects are usually limited to the immediate vicinity of the source.

Modeling: Forward and Inverse Approaches

Forward-modeling of a tsunami starts from given initial conditions, computes its propagation in the ocean, and calculates the tsunami arrival times and/or water run-up heights along the coasts. Once the initial conditions are provided, the propagation and coastal behavior can be numerically computed for an actual bathymetry. These calculations are useful for early tsunami warning and for detailed hazard estimations. However, the initial conditions associated with tsunami generation processes are still poorly known, because large tsunamis are rare and the tsunami generation in the open ocean is not directly observable. Currently, the tsunami source is estimated indirectly, mostly on the basis of seismological analysis, but a more direct estimation of the tsunami source is essential to better understand the tsunami generation process and to more accurately forecast the effects of a tsunami along the coasts.

In ▶ Tsunamis, Inverse Problem of, K. Satake reviews inverse methods used in the quantification of tsunami sources from the observations. The author describes the tsunami generation by earthquakes, with an emphasis on the fault parameters and their effects on tsunami propagation, including shallow water theory and numerical computation. The author then summarizes the tsunami observations, including instrumental sea-level data and run-up height estimates for modern, historical and prehistoric tsunamis. He also describes methods for modeling and quantifying a tsunami source, and for analyzing tsunami travel times, amplitudes and waveforms. He concludes with an estimation of earthquake fault parameters derived from waveform inversion of tsunami data, and a discussion of heterogeneous fault motion and its application for tsunami warning.

Tsunami inundation is the one of the final stages of tsunami evolution, when the wave encroaches upon and floods dry land. It is during this stage that a tsunami is most destructive and takes the vast majority of its victims. To gauge the near-shore impact of tsunami inundation, engineers and scientists rely primarily on three different methods: (1) field survey of past events, (2) physical experimentation in a laboratory, and (3) numerical modeling. In ▶ Tsunami Inundation, Modeling of, P.J. Lynett focuses on numerical simulations. He reviews tsunami generation and open ocean propagation, and discusses the physics of near-shore tsunami evolution, hydrodynamic modeling of tsunami evolution, moving shoreline algorithms, and effect of topographical features on inundation.

Tsunami Forecasting and Warning

The original definition of 'tsunami earthquake' was given by H. Kanamori (Phys Earth Planet Inter 6:346–359, 1972) as "an earthquake that produces a large-size tsunami relative to the value of its surface wave magnitude (M_S)". The true damage potential that a tsunami earthquake represents may not be recognized by conventional near real-time seismic analysis methods that utilize measurements of relatively high-frequency signals, and thus the threat may only become apparent upon the arrival of the tsunami waves on the local shores. Although tsunami earthquakes occur relatively infrequently, the effect on the local population can be devastating, as was most recently illustrated by the July 2006 Java tsunami earthquake, which was quickly followed by tsunami waves two to seven meters high, traveling as far as two kilometers inland and killing at least 668 people.

It is important to note that the definition of 'tsunami earthquake' is distinct from that of 'tsunamigenic earthquake'. A tsunamigenic earthquake is any earthquake that excites a tsunami. Tsunami earthquakes are a specific subset of tsunamigenic earthquakes. In ▶ Tsunami Earthquakes, J. Polet and H. Kanamori describe the characteristics of tsunami earthquakes and the possible factors involved in the anomalously strong excitation of tsunamis by these events. They also discuss a possible model for these infrequent, but potentially very damaging events.

Tsunamis are among nature's most destructive hazards. Typically generated by large, underwater shallow earthquakes, tsunamis can cross an ocean basin in a matter of hours. Although difficult to detect, and not dangerous while propagating in open ocean, tsunamis can unleash awesome destructive power when they reach coastal areas. With advance warning, populations dwelling in coastal areas can be alerted to evacuate to higher ground and away from the coast, thus saving many lives.

Tsunami travels at about the same speed of a commercial airliner, however, seismic waves can travel at speeds more than 40 times greater. Because of this large disparity in speed, scientists rely on seismic methods to detect the possibility of tsunami generation and to warn coastal populations of an approaching tsunami well in advance of its arrival. The seismic P-wave for example, travels from Alaska to Hawaii in about 7 min, whereas a tsunami will take about 5.5 hours to travel the same distance. Although over 200 sea-level stations reporting in near-real time are operating in the Pacific Ocean, it may take an hour or more, depending on the location of the epicenter, before the existence (or not) of an actual tsunami generation is confirmed. In other ocean basins where the density of sea-level instruments reporting data in near real-time is less, the delay in tsunami detection is correspondingly longer. However, global, regional, and local seismic networks, and the infrastructure needed to process the large amounts of seismic data that they record, are well in place around the world. For these reasons, tsunami warning centers provide initial tsunami warnings to coastal populations based entirely on the occurrence of a large shallow offshore earthquake. It is well-known, however, that large shallow offshore earthquakes may or may not be tsunamigenic.

In ► Tsunami Forecasting and Warning, O. Kamigaichi discusses the complexity problem in tsunami forecasting for large local events, and describes the Tsunami Early Warning System in Japan. Tsunami disaster mitigation can be achieved effectively by the appropriate combination of software and hardware countermeasures. Important issues for disaster mitigation includes: (1) improving people's awareness of the tsunami hazards, (2) imparting the necessity of spontaneous evacuation when people notice an imminent threat of tsunami on their own (feeling strong shaking near the coast, seeing abnormal sea level change, etc), (3) giving clear directions on how to respond to the tsunami forecast, and (4) conducting tsunami evacuation drills. The author notes that in tsunami forecasting, a trade-off exists between promptness and accuracy/reliability.

In ► Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, B. Hirshorn and S. Weinstein describe the basic method used by the Pacific Tsunami Warning Center (PTWC) mainly for large teleseismic events. Software running at the PTWC processes in real time seismic signals from over 150 seismic stations worldwide provided by various seismic networks. Automatic seismic event detection algorithms page the duty scientists for any earthquake occurring worldwide over about Magnitude 5.5. Other automatic software locates these events, and provides a first estimate of their magnitude and other source parameters in near real time. Duty scientists then refine the software's automated source parameter estimates and issue a warning if necessary. The authors also describe their ongoing efforts to improve estimates of earthquake source parameters.

Wedge Mechanics, Submarine Landslides and Slow Earthquakes

A study of the mechanics of wedge-shaped geological bodies, such as accretionary prisms in subduction zones and fold-and-thrust belts in collision zones, is interesting because they enable us to use the observed morphology and deformation of these bodies to constrain properties of the thrust faults underlying them. The fundamental process described in wedge mechanics is how gravitational force, in the presence of a sloping surface, is balanced by basal stress and internal stress. The internal state of stress depends on the rheology of the wedge. The most commonly assumed wedge rheology for geological problems is perfect Coulomb plasticity, and the model based on this rheology is referred to as the Coulomb wedge model.

The connection between wedge mechanics and great earthquakes and tsunamis at subduction zones is an emerging new field of study. In their article, ► Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Wang et al. cover the topics of stable and critical Coulomb wedges, dynamic Coulomb wedge, stress drop and increase in a subduction earthquake, and tsunamigenic coseismic seafloor deformation. Better constraints are needed to quantify how stresses along different downdip segments of the subduction fault evolve with time throughout an earthquake cycle and how the evolution impacts wedge and seafloor deformation. Submarine monitoring in conjunction with land-based monitoring at subduction zones that are currently in different phases of earthquake cycles will allow us to better understand the evolution of fault and wedge stresses during the interseismic period. In this regard, cabled seafloor monitoring networks including borehole observatories being designed or implemented at different subduction zones will surely yield valuable data in the near future.

The term 'submarine landslide' encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. The term ‘slow earthquake’ describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Recently, a GPS network on the south flank of Kilauea volcano, Hawaii, recorded multiple slow earthquakes on the subaerial portion of a large landslide system that extends primarily into the submarine environment. Since catastrophic failure of submarine landslides can cause a tsunami they represent significant hazards to coastal zones. Because submarine landslide systems are among the most active as well as spatially confined deforming areas on Earth, they are excellent targets for understanding the general fault failure process. In ► Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, B.A. Brooks et al. present a review of this interdisciplinary topic of interest in geodesy, seismology, tsunamis, and volcanology.

Volcanoes

About 1,500 volcanoes have erupted one or more times during the past 10,000 years, and since A.D. 1600, volcanic disasters have killed about 300,000 people and resulted in property damage and economic loss exceeding hundreds of millions of dollar. Articles in this section are intended to summarize recent research in: (1) volcano seismology, (2) physical processes involved in volcanoes, and (3) modeling volcanic eruptions and hazards warning.

Volcano Seismology

Magma transport in a volcano is episodic due to the inherent instability of magmatic systems at all time scales. This episodicity is reflected in seismic activity, which originates in dynamic interactions between gas, liquid and solid along magma transport paths that involve complex geometries. The description of the flow processes is governed by the nonlinear equations of fluid dynamics. In volcanic fluids, further complexity arises from the strong nonlinear dependence of magma rheology on temperature, pressure, and water and crystal content, and nonlinear characteristics of associated processes underlying the physico-chemical evolution of liquid-gas mixtures constituting magma.

In ► Volcanoes, Non-linear Processes in, B. Chouet presents a brief review of volcano seismology and addresses basic issues in the quantitative interpretation of processes in active volcanic systems. Starting with an introduction of the seismic methodology used to quantify the source of volcano seismicity, the author then focuses on sources originating in the dynamics of volcanic fluids. A review of some of the representative source mechanisms of Long-Period (LP) and Very Long-Period (VLP) signals is followed by a description of a mesoscale computational approach for simulating two-phase flows of complex magmatic fluids. Refined understanding of magma and hydrothermal transport dynamics therefore requires multidisciplinary research involving detailed field measurements, laboratory experiments, and numerical modeling. Such research is fundamental to monitoring and interpreting the subsurface migration of magma that often leads to eruptions, and thus would enhance our ability to forecast hazardous volcanic activity.

Volcano seismicity produces a wide variety of seismic signals that provide glimpses of the internal dynamics of volcanic systems. Quantitative approaches to analysis and interpret volcano-seismic signals have been developed since the late 1970s. The availability of seismic equipments with wide frequency and dynamic ranges since the early 1990s has revealed a variety of volcano-seismic signals over a wide range of periods. Quantification of the sources of volcano-seismic signals is crucial to achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems. In ► Volcano Seismic Signals, Source Quantification of, H. Kumagai provides the theoretical basis for a quantification of the sources of volcano-seismic signals. The author focuses on the phenomenological representation of seismic sources, waveform inversion to estimate source mechanisms, spectral analysis based on an autoregressive model, and physical properties of fluid-solid coupled waves.

Among various eruptive styles, Strombolian activity is easier to study because of its repetitive behavior. Since Strombolian activity offers numerous interesting seismic signals, a growing attention has been devoted to the application of waveform inversion for imaging conduit geometry and retrieving eruption dynamics from seismological recordings. Quantitative models fitting seismological observations are a powerful tool for interpreting seismic recordings from active volcanoes. In ► Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, L. D’Auria and M. Martini discuss the mechanism of generation of Very-Long Period (VLP) signals accompanying Strombolian explosions. This eruptive style, occurring at many basaltic volcanoes worldwide, is characterized by the ascent and the bursting of large gas slugs. The mechanism of formation, ascent and explosion of bubbles and slugs and their relation with eruptive activity has been studied theoretically and by analogue simulations. The authors report results from numerical simulations, focusing on the seismic signals generated by pressure variations applied to the conduit walls.

Physical Processes in Volcanoes

The dynamics of solid-liquid composite systems are relevant to many problems, including how melts or aqueous fluids migrate through the mantle and crust toward the surface, how deformation and fracture in these regions are influenced by the existence of fluids, and also how these fluids can be observed in seismic tomographic images. In ▶ Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Y. Takei introduces a general continuum mechanical theory for macroscopic dynamics of solid-liquid composite systems, and emphasizes on how such interactions with pore geometry can be studied. The author then discusses the determinability of porosity and pore geometry from seismic tomographic images, and presents a practical method to assess porosity and pore geometry from tomographic V_P and V_S images.

A volcano consists of solids, liquids, gases, and intermediate materials of any two of these phases. Mechanical and thermodynamical interactions between these phases are essential in the generating a variety of volcanic activities. In particular, the gas phase is mechanically distinct from the other phases and plays an important role in dynamic phenomena in volcanoes. In ▶ Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, M. Ichihara and T. Nishimura discuss several bubble dynamics phenomena from the viewpoint that a bubbly fluid acts as an impulse generator of observable signals, such as earthquakes, ground deformations, airwaves, and an eruption itself. The authors focus on the notion that the impulse is excited by non-linear coupling between internal processes in a bubbly fluid and an external perturbation. The importance of these processes has recently become noticed as a possible triggering mechanism of eruptions, earthquakes, and inflation of a volcano.

Our capability to mitigate volcano hazards relies in large part on forecasting explosive events, a process which requires a high degree of understanding of the physicochemical factors operating during explosive volcanism. The approaches taken to gain an understanding of explosive volcanism have relied on a combination of field observations, theoretical models and laboratory models of materials and mechanisms. In ▶ Volcanic Eruptions, Explosive: Experimental Insights, S.J. Lane and M.R. James first review aspects of the volcanic materials literature, with the aim of illustrating the nature of molten rock, the complexity of which underpins most explosive volcanic processes. Experimental modeling of these processes can then build on the materials understanding. Such experiments involve investigation of the behavior of natural volcanic products at laboratory time and length scales, including the response of magma samples to rapid changes in pressure and temperature, the fall behavior of silicate particles in the atmosphere, and the generation and separation of electrostatic charge during explosive eruptions.

In ▶ Volcanic Eruptions: Cyclicity During Lava Dome Growth, O. Melnik et al. consider the process of slow extrusion of very viscous magma that forms lava domes. Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows generated by dome collapses, explosive eruptions, and volcanic blasts. These eruptions commonly display fairly regular alternations between periods of high and low or no activity with time scales from hours to years. Usually hazardous phenomena are associated with periods of high magma discharge rate. Hence, understanding the causes of pulse activity during extrusive eruptions is an important step towards forecasting volcanic behavior, and especially the transition to explosive activity when magma discharge rate increases by a few orders of magnitude. In recent years the risks have escalated because the population density in the vicinity of many active volcanoes has increased.

Modeling Volcanic Eruptions and Hazards Warning

While a wide range of complex deterministic models exists to model various volcanic processes, these provide little in the way of information about future activity. Being the (partially) observed realization of a complex system, volcanological data are inherently stochastic in nature, and need to be modeled using statistical models. In ▶ Volcanic Eruptions: Stochastic Models of Occurrence Patterns, M.S. Bebbington considers models of eruption occurrence, omitting techniques for forecasting the nature and effect of the eruption. As the track record of a potentially active volcano provides the best method of assessing its future long-term hazards, the author first briefly reviews the provenance and characteristics of the data available, and then discusses various taxonomies for stochastic models. The examples of Mount Etna and Yucca Mountain are selected for more detailed examination partly because many, somewhat contradictory, results exist. Different models make different assumptions, and vary in how much information they can extract from data. In addition, the data used often varies from study to study, and the sensitivity of models to data is important, but too often ignored.

In ▶ Volcanic Hazards and Early Warning, R.I. Tilling highlights the range in possible outcomes of volcano unrest and reviews some recent examples of the actual outcomes documented for several well-monitored volcanoes. The author also discusses the challenge for emergency-management authorities, as well as challenges in achieving refined predictive capability. To respond effectively to a developing volcanic crisis, timely and reliable early warnings are absolutely essential; they can be achieved only by a greatly improved capability for eruption prediction. This in turn depends on the quantity and quality of volcano-monitoring data and the diagnostic interpretation of such information.

Discussions

The terms 'Prediction' and 'forecasting' are often used interchangeably. However, the commonly accepted definition of an 'earthquake prediction' is a concise statement, in advance of the event, of the time, location, and magnitude of a future earthquake. To be practically useful to the society, the time window must be short (in days or months), the location extent small (within tens of kilometers), and the magnitude precise (± 0.5 unit). A 'forecast', on the other hand, is more loosely defined as the probability of an occurrence of a large earthquake in a given region (e. g., southern California) during the coming decades or centuries. The broader time intervals associated with forecasts allows society to consider and implement mitigation efforts over a large region.

As an observational seismologist and on a personal note, I am skeptical that earthquakes can be reliably predicted before (1) we have collected accurate data of their occurrences over a sufficiently long period of time, and (2) we have a good understanding of the physical processes that create them. Although earthquakes have been known from antiquity, accurate earthquake catalogs exist only since about the 1960s. Since the recurrence of a damaging earthquake in a given area is often more than 100 years (some even thousands of years), it is obvious that we lack the necessary observed data. C. Lanczos (*Linear Differential Operators*, Van Nostrand-Reinhold, 1961) said it well in general: "a lack of information cannot be remedied by any mathematical trickery". Nevertheless, we must apply new concepts and tools to extract as much useful information as possible from the existing data. Indeed, we must thank many pioneers for enlightening us with many interesting and tentative results about earthquakes, tsunamis and volcanoes that they managed to extract from inadequate and insufficient data.

Because most tsunamis are generated by earthquakes, successes in predicting tsunamis depend on predicting earthquakes and recognizing them as tsunamigenic. Predicting a volcanic eruption is a little easier, as the location is known and there are often some observable phenomena preceding it. However, the exact time and the intensity and extent of an eruption are difficult to predict because the volcanic processes are very complex involving gas, liquid and solid phases.

Fortunately, earthquake, tsunami and volcano hazards can be reduced by employing sound engineering practices, early warning systems, and hazard analysis, using many of the tools and concepts that were developed for prediction. Since fatalities and economic loss from a single catastrophic event can reach 100,000 or more, and $100 billion or more, respectively, it is imperative that governments should support long-term monitoring with modern instruments and research, including complexity studies of earthquakes, tsunamis and volcanoes.

Complexity and Non-linearity in Autonomous Robotics, Introduction to

WARREN E. DIXON
Department of Mechanical and Aerospace Engineering,
Gainesville, USA

Historically, robotic systems have played a key role in manufacturing, hazardous material handling, exploration and surveillance, search and rescue, and military applications. For most of these applications, single robot systems with traditional sensing and actuation are able to execute predetermined tasks in a well defined environment. The ability to develop a robotic system capable of executing tasks in an environment that is not well defined or is dynamic has been a daunting task. The development of such a capability has been limited by the complexity required in sensing, cognitive decision making, new actuation capabilities, and robustness and responsiveness of the control system. The complexity of these challenges increases when groups of robots are required to interact with each other or with people in a social or physical manner. This volume of work is a collection of recent and emerging efforts at the dawn of a new millennium to develop enabling technologies for new frontiers where robots are able to execute tasks in dynamic and unstructured environments.

Sensing is a fundamental requirement for an autonomous system to efficiently interact with its environment. Vision is arguably the primary environmental sensor used by human beings and many other animals to understand how to interact with dynamic and unstructured environments. While traditional sensor modalities will continue to be an integral part of emerging robotic systems, new research efforts will be driven by the desire for human-like interpretation and reaction to image information. However, the use of image-based feedback adds complexity and new challenges. One challenge is how an autonomous system can find and track objects of interest in an image of dense, cluttered environments (see ▶ Motion Prediction for Continued Autonomy). Once a target(s) of interest is determined, how can a robot use a set of two dimensional images to interpret its relative position and orientation with respect to different objects (see ▶ Image Based State Estimation). Another challenge is how to use reconstructed and estimated information from an image to develop a stable closed-loop error system (see ▶ Adaptive Visual Servo Control).

Given feedback from vision sensors, traditional sensors (e. g., sonar, radar, lidar, encoders), or a fusion of these sensors, another complexity is how to best exploit sensor and other information discovered during the execution of a plan, including autonomous selection of which behavior(s) to invoke, in what sequence, and by what method. The ability to organize and apply situational knowledge to macro-level planning and decision-making can be enabled through a scaffolding provided by a computer architecture that provides interfaces for different sources of information (see ▶ Software Architectures for Autonomy). As robotic systems operate in increasingly more complicated environments, more intelligence is required by the autonomous system. The ultimate goal for an autonomous system is to enable a level of cognition with human-like intelligence in perception, motor control, and high-level cognition. Beyond sensing and reaction, cognition enables a robotic system to reason about an environment without direct interaction, enabling the capability to plan for selecting between competing goals, coping with multiple sensory inputs, and completing multiple tasks (see ▶ Cognitive Robotics). Computer architectures that provide a backbone for organizing and interfacing with different sensing and actuation components and the ability for autonomous systems to exhibit increasing levels of intelligence are especially important as robotics move from behind chained off work cells in an assembly line to direct emotional and physical interaction with people (see ▶ Human Robot Interaction).

In addition to new frontiers in sensing, computer architectures, and autonomous reasoning, advances are also required in the construction and actuation of an autonomous robotic system. For example, since people spend a significant portion of their time inside buildings, the building itself could be designed as a morphing robot surface that is articulated, programmable, and embedded with integrated digital technologies (see ▶ Continuum Robots). Or, as robots are forecasted to take an increasing role for in-home care, robots will need new methods of physically interacting with people and the environment. For example, many biological systems have shown that significant articulation can be achieved through tongues, trunks, and tentacles. Robots inspired by these systems could also morph their shape to their environment to access difficult-to-reach areas, and to perform adaptive grasping using whole arm manipulation (see also ▶ Continuum Robots).

New actuation and manipulation methods place new challenges on control systems. One challenge includes the lack of precise mathematical models of the environment and its interaction with the robot through actuators and sensors. As autonomous robots interact with people, the robot may be required to interpret events that a person describes using linguistic terms. These challenges motivate the need for new control designs that may be inspired by the way that humans tend to work with vague or imprecise concepts (see ▶ Neuro-fuzzy Control of Autonomous Robotics).

Given the complexity that arises from developing a single autonomous system capable of interacting with uncertain and dynamic environments, motivation exists to develop technologies for groups of robotic systems. Distributed robotics hold the promise to enable groups of robots to collaborate to solve complex tasks such as monitoring a vast environment, manipulating large objects, or building advanced structures. One challenge to enable groups of robots to collaborate efficiently is how to develop path planning and motion coordination methods so that teams of autonomous mobile robots can share the same workspace while avoiding interference with each other, and/or while achieving group motion objectives (see ▶ Multiple Mobile Robot Teams, Path Planning and Motion Coordination in). Obtaining an optimal solution for the coordination of multiple robot vehicles is a challenge because an integrated approach for designing communication, sensing, and control systems must all be considered as constraints on the system performance (see ▶ Distributed Controls of Multiple Robotic Systems, An Optimization Approach). The design of a single robot requires the balancing of many factors (e. g., cost, sensing, processing capabilities), many of which are conflicting. In addition to these factors, distributed robotics adds the complexity of multiple robots. For example, there are different advantages and disadvantages to designing a homogeneous versus a heterogeneous team (see ▶ Distributed Robotic Teams: A Framework for Simulated and Real–World Modeling).

Despite different design tradeoffs for teams of robots, some advanced teams hold the potential for unprecedented versatility and robustness at a low cost. For example, modular self-reconfigurable (MSR) robot teams are composed of a large number of repeated modules that can rearrange their connectedness to form a large variety of structures to suit the task (see ▶ Modular Self-Reconfigurable Robots). In contrast to self-reconfigurable robotics, self-replication utilizes an original unit to actively assemble an exact copy of itself from passive components. This has the potential to result in an exponential growth in the number of robots available to perform a job, thus drastically shortening the original unit's task time (see ▶ Self--replicating Robotic Systems). To enable self replication, a robot must exploit resources in the environment such as raw materials and energy. Insights to address this challenge can be provided from nature, where biological systems must forage for resources. Foraging robots are mobile robots capable of searching for and, when found, transporting objects to one or more collection points. Foraging robots may be single robots operating individually, or multiple robots operating collectively (see ▶ Foraging Robots).

New technologies are emerging in sensing, actuation, reasoning, and control. As these technologies mature, new levels of autonomy are made possible. As robotic systems are able to operate in increasingly more dynamic and uncertain environments, an exciting new era for robotics is evident.

Complexity in Systems Level Biology and Genetics: Statistical Perspectives

David A. Stephens
Department of Mathematics and Statistics,
McGill University, Montreal, Canada

Article Outline

Glossary

Systems biology The holistic study of biological structure, function and organization.

Probabilistic graphical model A probabilistic model defining the relationships between variables in a model by means of a graph, used to represent the relationships in a biological network or pathway.

MCMC Markov chain Monte Carlo – a computational method for approximating high-dimensional integrals using Markov chains to sample from probability distributions, commonly used in Bayesian inference.

Microarray A high-throughput experimental platform for collecting functional gene expression and other genomic data.

Cluster analysis A statistical method for discovering subgroups in data.

Metabolomics The study of the metabolic content of tissues.

Definition of the Subject

This chapter identifies the challenges posed to biologists, geneticists and other scientists by advances in technology that have made the observation and study of biological systems increasingly possible. High-throughput platforms have made routine the collection vast amounts of structural and functional data, and have provided insights into the working cell, and helped to explain the role of genetics in common diseases. Associated with the improvements in technology is the need for statistical procedures that extract the biological information from the available data in a coherent fashion, and perhaps more importantly, can quantify the certainty with which conclusions can be made. This chapter outlines a biological hierarchy of structures, functions and interactions that can now be observed, and detail the statistical procedures that are necessary for analyzing the resulting data. The chapter has four main sections. The first section details the historical connection between statistics and the analysis of biological and genetic data, and summarizes fundamental concepts in biology and genetics. The second section outlines specific mathematical and statistical methods that are useful in the modeling of data arising in bioinformatics. In sections three and four, two particular issues are discussed in detail: functional genomic via microarray analysis, and metabolomics. Section five identifies some future directions for biological research in which statisticians will play a vital role.

Introduction

The observation of biological systems, their processes and inter-reactions, is one of the most important activities in modern science. It has the capacity to provide direct insight into fundamental aspects of biology, genetics, evolution, and indirectly will inform many aspects of public health. Recent advances in technology – high-throughput measurement platforms, imaging – have brought a new era of increasingly precise methods of investigation. In parallel to this, there is an increasingly important focus on statistical methods that allow the information gathered to be processed and synthesized. This chapter outlines key statistical techniques that allow the information gathered to be used in an optimal fashion.

Although its origin is dated rather earlier, the term *Systems Biology* (see, for example, [1,2,3]) has, since 2000, been used to describe the study of the operation of biological systems, using tools from mathematics, statistics and computer science, supplanting *computational biology* and *bioinformatics* as an all-encompassing term for quantitative investigation in molecular biology. Most biological systems are hugely complex, involving chemical and mechanical processes operating at different scales. It is important therefore that information gathered is processed coherently, according to self-consistent rules and practices, in the presence of the uncertainty induced by imperfect observation of the underlying system. The most natural framework for coherent processing of information is that of probabilistic modeling.

Statistical Versus Mathematical Modeling

There is a great tradition of mathematical and probabilistic modeling of biology and genetics; see [4] for a thorough review. The mathematization of biology, evolution and heredity began at the end of the nineteenth century, and continued for the first half of the twentieth century, by far pre-dating the era of molecular biology and genetics that culminated at the turn of the last millennium with the human genome project. Consequently, the mathematical models of, say, evolutionary processes that were developed by Yule [5] and Fisher and Wright [6,7,8], and classical models of heredity, could only be experimentally verified and developed many years after their conception. It could also be convincingly argued that through the work of F. Galton, K. S. Pearson and R. A. Fisher, modern statistics has its foundation in biology and genetics.

In parallel to statistical and *stochastic* formulation of models for biological systems, there has been a more recent focus on the construction of *deterministic* models to describe observed biological phenomena. Such models fall under the broad description *Mathematical Biology*, and have their roots in applied mathematics and dynamical systems; see, for example, [8,9] for a comprehensive treatment. The distinction between stochastic and deterministic models is important to make, as the objectives and tools used often differ considerably. This chapter will restrict attention to stochastic models, and the processing of observed data, and thus is perhaps more closely tied to the immediate interests of the scientist, although some of the models utilized will be inspired by mathematical models of the phenomena being observed.

Fundamental Concepts in Biology and Genetics

To facilitate the discussion of statistical methods applied to systems biology, it is necessary to introduce fundamental concepts from molecular biology and genetics; see the classic text [10] for full details. Attention is restricted to eukaryotic, organisms whose cells are constructed to contain a nucleus which coding information is encapsulated.

- The cell nucleus is a complex architecture containing several nuclear domains [11] whose organization is not completely understood, but the fundamental activity that occurs within the nucleus is the production and distribution of proteins.
- Dioxyribonucleic acid (DNA) is a long string of nucleotides that encodes biological information, and that is copied or *transcribed* into ribonucleic acid (RNA), which in turn enables the formation of proteins. Specific segments of the DNA, genes, encode the proteins, although non-coding regions of DNA – for example, promoter regions, transcription factor binding sites – also have important roles. Genetic variation at the nucleotide level, even involving a single nucleotide, can disrupt cellular activity. In humans and most other complex organisms, DNA is arranged into chromosomes, which are duplicated in the process of mitosis. The entire content of the DNA of an organism is termed the *genome*.
- Proteins are macromolecules formed by the *translation* of RNA, comprising amino acids arranged (in primary structure) in a linear fashion, comprising domains with different roles, and physically configured in three dimensional space. Proteins are responsible for all biological activities that take place in the cell, although proteins may have different roles in different tissues at different times, due to the *regulation* of transcription.
- Proteins interact with each other in different ways in different contexts in interaction networks that may be dynamically organized. Genes are also regarded as having indirect interactions through gene regulatory networks.
- Genetic variation amongst individuals in a population is due to mutation and selection, which can be regarded as stochastic mechanisms. Genetic information in the form of DNA passes from parent to offspring, which

promulgates genetic variation. Individuals in a population are typically related in evolutionary history. Similarly, proteins can also thought to be related through evolutionary history.

- Genetic disorders are the result of genetic variation, but the nature of the genetic variation can be large- or small-scale; at the smallest scale, variation in single nucleotides (*single nucleotide polymorphisms* (SNPs)) can contribute to the variation in observed traits.

Broadly, attention is focused on the study of *structure* and *function* of DNA, genes and proteins, and the nature of their *interactions*. It is useful, if simplistic, to view biological activities in terms of an organizational hierarchy of inter-related chemical reactions at the level of DNA, protein, nucleus, network and cellular levels. A holistic view of mathematical modeling and statistical inference requires the experimenter to model simultaneously actions and interactions of all the component features, whilst recognizing that the component features cannot observed directly, and can only be studied through separate experiments on often widely different platforms. It is the role of the bioinformatician or systems biologist to synthesize the data available from separate experiments in an optimal fashion.

Mathematical Representations of the Organizational Hierarchy

A mathematical representation of a biological system is required that recognizes, first, the complexity of the system, secondly, its potentially temporally changing nature, and thirdly the inherent uncertainties that are present. It is the last feature that necessitates the use of probabilistic or stochastic modeling.

An aphorism commonly ascribed to D.V. Lindley states that "*Probability is the language of uncertainty*"; probability provides a coherent framework for processing information in the presence of imperfect knowledge, and through the paradigm of Bayesian theory [12] provides the mathematical template for statistical inference and prediction. In the modeling of complex systems, three sorts of uncertainty are typically present:

- *Uncertainty of Structure:* Imperfect knowledge of the connections between the interacting components is typically present. For example, in a gene regulatory network, it may be possible via the measurement of gene co-expression to establish which genes interact within the network, but it may not be apparent precisely how the organization of regulation operates, that is which genes regulate the expression of other genes.
- *Uncertainty concerning Model Components:* In any mathematical or probabilistic model of a biological system, there are model components (differential equations, probability distributions, parameter settings) that must be chosen to facilitate implementation of the model. These components reflect, but are not determined by, structural considerations.
- *Uncertainty of Observation:* Any experimental procedure carries with it uncertainty induced by the measurement of underlying system, that is typically subject to random measurement error, or noise. For example, many biological systems rely on imaging technology, and the extraction of the level of signal of a fluorescent probe, for a representation of the amount of biological material present. In microarray studies (see Sect. "Microarrays"), comparative hybridization of messenger RNA (mRNA) to a medium is a technique for measuring gene expression that is noisy due to several factors (imaging noise, variation in hybridization) not attributable to a biological cause.

The framework to be built must handle these types of uncertainty, and permit inference about structure and model components.

Models Derived from Differential Equations

A deterministic model reflecting the dynamic relationships often present in biological systems may be based on the system of ordinary differential equations (ODEs)

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) \qquad (1)$$

where $\boldsymbol{x}(t) = (x_1(t), \dots, x_d(t))^{\mathsf{T}}$ represent the levels of d quantities being observed $\dot{\boldsymbol{x}}(t)$ represents time derivative, and $\boldsymbol{g}$ is some potentially non-linear system of equations, that may be suggested by biological prior knowledge or prior experimentation. The model in Eq. (1) is a classical "Mathematical Biology" model, that has been successful in representing forms of organization in many biological systems (see, for example [13] for general applications). Suppressed in the notation is a dependence on system parameters, $\boldsymbol{\theta}$, a k-dimensional vector that may be presumed fixed, and "tuned" to replicate observed behavior, or estimated from observed data. When data representing a partial observation of the system are available, inferences about $\boldsymbol{\theta}$ can be made, and models defined by ODE systems are of growing interest to statisticians; see, for example, [15,16,17].

Equation (1) can be readily extended to a *stochastic differential equation* (SDE) system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) + \boldsymbol{dz}(t) \qquad (2)$$

where $\boldsymbol{z}(t)$ is some stochastic process that renders the solution to Eq. (2) a stochastic process (see, for example, [18] for a comprehensive recent summary of modeling approaches and inference procedures, and a specific application in [19]). The final term $\boldsymbol{dz}(t)$ represents the infinitesimal stochastic increment in $\boldsymbol{z}(t)$. Such models, although particularly useful for modeling activity at the molecular level, often rely on simplifying assumptions (linearity of $\boldsymbol{g}$, Gaussianity of $\boldsymbol{z}$) and the fact that the relationship structure captured by $\boldsymbol{g}$ is known. Inference for the parameters of the system can be made, but in general require advanced computational methods (Monte Carlo (MC) and Markov chain Monte Carlo (MCMC)).

Probabilistic Graphical Models

A simple and often directly implementable approach is based on a *probabilistic graphical model*, comprising a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, described by a series of nodes $\mathcal{N}$, edges $\mathcal{E}$, and a collection of random variables $\boldsymbol{X} = (X_1, \dots, X_d)^\mathsf{T}$ placed at the nodes, all of which may be dynamically changing. See, for example [20] for a recent summary, [14] for mathematical details and [22] for a biological application.

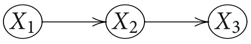
The objective of constructing such a model is to identify the joint probability structure of $\boldsymbol{X}$ given the graph $\mathcal{G}$, which possibly is parametrized by parameters $\boldsymbol{\phi}$, $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G})$. In many applications, $\boldsymbol{X}$ is not directly observed, but is instead inferred from observed data, $\boldsymbol{Y}$, arising as noisy observations derived from $\boldsymbol{X}$. Again, a k-dimensional parameter vector $\boldsymbol{\theta}$ helps to characterize the stochastic dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ by parametrize the conditional probability density $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$. The joint probability model encapsulating the probabilistic structure of the model is

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) = f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{3}$$

The objectives of inference are to learn about $\mathcal{G}$ (the uncertain structural component) and parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ (the uncertain model parameters and observation components).

The graph structure $\mathcal{G}$ is described by $\mathcal{N}$ and $\mathcal{E}$. In holistic models, $\mathcal{G}$ represents the interconnections between interacting modules (genomic modules, transcription modules, regulatory modules, proteomic modules, metabolic modules etc.) and also the interconnections within modules in the form of sub graphs. The nodes $\mathcal{N}$ (and hence $\boldsymbol{X}$) represent influential variables in the model structure, and the edges $\mathcal{E}$ represent dependencies. The edge connecting two nodes, if present, may be *directed* or *undirected* according to the nature of the influence; a directed edge indicates the direction of *causation*, an undirected edge indicates a *dependence*.

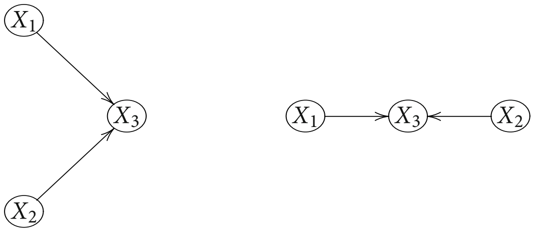
Causality is a concept distinct from dependence (association, co-variation or correlation), and represents the influence of one node on one or more other nodes (see, for example, [23] for a recent discussion of the distinction with examples, and [24,25] for early influential papers discussing how functional dependence may be learned from real data). A simple causal relationship between three variables X_1, X_2, X_3 can be represented



which encodes a *conditional independence relationship* between X_1 and X_3 given X_2, and a factorization of the joint distribution

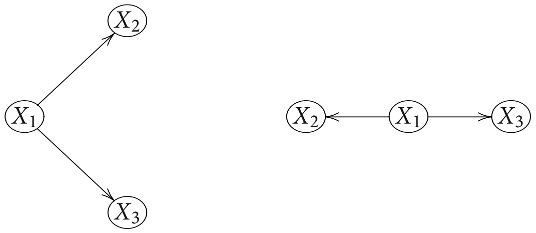
$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2).$$

Similarly, the equivalent graphs



encode $\quad p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$

whereas the graphs for conditional independence of X_2 and X_3 given X_1 are



encoding $\quad p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1). \quad (4)$

Such simple model assumptions are the building blocks for the construction of highly complex graphical representations of biological systems. There is an important difference between analysis based purely on simultaneous observation of all components of the system, which can typically on yield inference on *dependencies* (say, covariances measured in the joint probability model $p(\boldsymbol{x})$ – see, for example [26,27,28] and analysis based on *interventions*-genomic knock–out experiments, chemical or biological challenges, transcriptional/translational perturbation such as RNA interference (RNAi) – that may yield information on casual links; see, for example [29,30].

Bayesian Statistical Inference

Given a statistical model for observed data such as Eq. (3), inference for the parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and the graph structure $\mathcal{G}$ is required. The optimal coherent framework is that of *Bayesian statistical inference* (see for example [31]), that requires computation of the *posterior distribution* for the unknown (or *unobservable*) quantities given by

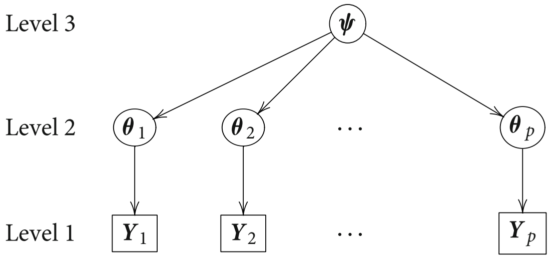
$$\begin{aligned}\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y}) &\propto f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \\ &= L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y})p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \qquad (5)\end{aligned}$$

a probability distribution from which can be computed parameter estimates with associated uncertainties, and predictions from the model. The terms $L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})$ are termed *likelihood* and *prior probability distribution* respectively. The likelihood reflects the observed data, and the prior distribution encapsulates biological prior knowledge about the system under study. If the graph structure is known in advance, the prior distribution for that component can be set to be degenerate. If, as in many cases of probabilistic graphical models, the $\boldsymbol{x}$ are unobserved, then the posterior distribution incorporates them also,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}, \boldsymbol{x}|, \boldsymbol{y}) \propto f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) f_X(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \quad (6)$$

yielding a *latent* or *state-space* model, otherwise interpreted as a *missing data* model.

The likelihood and prior can often be formulated in a hierarchical fashion to reflect believed causal or conditional independence structures. If a graph $\mathcal{G}$ is separable into two sub graphs $\mathcal{G}_1, \mathcal{G}_2$ conditional on a connecting node η, similar to the graph in Eq. (4), then the probability model also factorizes into a similar fashion; for example, X_1 might represent the amount of expressed mRNA of a gene that regulates two separate functional modules, and X_2 and X_3 might be the levels of expression of collections of related proteins. The hierarchical specification also extends to parameters in probability models; a standard formulation of a Bayesian hierarchical model involves specification of conditional independence structures at multiple levels of within a graph. The following three-level hierarchical model relates data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ at level 1, to a population of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^{\mathsf{T}}$ at level 2, to hyper parameters $\boldsymbol{\psi}$ at level 3



(7)

yielding the factorization of the Bayesian full joint distribution as

$$\begin{aligned}&f_{X,Y,\boldsymbol{\psi},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &\qquad = p(\boldsymbol{\psi}) \left\{ \prod_{i=1}^{p} p(\boldsymbol{\theta}_i|\boldsymbol{\psi}) \right\} \left\{ \prod_{i=1}^{p} p(\boldsymbol{Y}_i|\boldsymbol{\phi}_i) \right\} .\end{aligned}$$

Bayesian Computation

The posterior distribution is, potentially, a high-dimensional multivariate function on a complicated parameter space. The proportionality constant in Eq. (5) takes the form

$$f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}) = \int f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\phi}\,\mathrm{d}\mathcal{G} \qquad (8)$$

and in Eq. (6) takes the form

$$f_Y(\boldsymbol{y}) = \int f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\phi}\,\mathrm{d}\mathcal{G}\,\mathrm{d}\boldsymbol{x} \quad (9)$$

and is termed the *marginal likelihood* or *prior predictive distribution* for the *observable* quantities $\boldsymbol{x}$ and $\boldsymbol{y}$. In formal Bayesian theory, it is the representation of the distribution of the observable quantities through the paradigm of *exchangeability* that justifies the decomposition in Eq. (8) into likelihood and prior, and justifies, via asymptotic arguments, the use of the posterior distribution for inference (see Chaps. 1–4 in [13] for full details). It is evident from these equations that exact computation of the posterior distribution necessitates high-dimensional integration, and in many cases this cannot be carried out analytically.

Numerical Integration Approaches Classical numerical integration methods, or analytic approximation methods are suitable only in low dimensions. Stochastic numerical integration, for example Monte Carlo integration, approximates expectations by using empirical averages of functional of samples obtained from the target distribution; for probability distribution $\pi(\boldsymbol{x})$, the approximation of $\mathrm{E}_\pi[g(\boldsymbol{X})]$,

$$\mathrm{E}_\pi[g(\boldsymbol{X})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$$

is achieved by randomly sampling $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ (N large) from $\pi()$, and using the estimate

$$\widehat{\mathrm{E}}_\pi[g(\boldsymbol{X})] = \frac{1}{N}\sum_{i=1}^{n} g(\boldsymbol{x}_i).$$

An adaptation of the Monte Carlo method can be used if the functions g and π are not "similar" (in the sense that g is large in magnitude where π is not, and *vice versa*); *importance sampling* uses the representation

$$\mathrm{E}_\pi[g(\boldsymbol{X})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \frac{g(\boldsymbol{x})\pi(\boldsymbol{x})}{p(\boldsymbol{x})}p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

for some pdf $p()$ having common support with π, and constructs an estimate from a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from $p()$ of the form

$$\widehat{\mathrm{E}}_\pi[g(\boldsymbol{X})] = \frac{1}{N}\sum_{i=1}^{n} \frac{g(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}.$$

Under standard regularity conditions, the corresponding estimators converge to the required expectation. Further extensions are also useful:

- *Sequential Monte Carlo:* Sequential Monte Carlo (SMC) is an adaptive procedure that constructs a sequence of improving importance sampling distributions. SMC is a technique that is especially useful for inference problems where data are collected sequentially in time, but is also used in standard Monte Carlo problems (see [32]).
- *Quasi Monte Carlo:* Quasi Monte Carlo (QMC) utilizes uniform **but not random** samples to approximate the required expectations. It can be shown that QMC can produce estimators with lower variance than standard Monte Carlo.

Markov Chain Monte Carlo Markov chain Monte Carlo (MCMC) is a stochastic Monte Carlo method for sampling from a high-dimensional probability distribution $\pi(\boldsymbol{x})$, and using the samples to approximate expectations with respect to that distribution. An ergodic, discrete-time Markov chain is defined on the support of π in such a way that the stationary distribution of the chain exists, and is equal to π. Dependent samples from π are obtained by collecting realized values of the chain after it has reached its stationary phase, and then used as the basis of a Monte Carlo strategy.

The most common MCMC algorithm is known as the *Metropolis–Hastings* algorithm which proceeds as follows. If the state of the d-dimensional chain $\{\boldsymbol{X}_t\}$ at iteration t is given by $\boldsymbol{X}_t = \boldsymbol{u}$, then a candidate state $\boldsymbol{v}$ is generated from conditional density $q(\boldsymbol{u}, \boldsymbol{v}) = q(\boldsymbol{v}|\boldsymbol{u})$, and accepted as the new state of the chain (that is, $\boldsymbol{X}_{t+1} \overset{\mathrm{def}}{=} \boldsymbol{v}$) with probability $\alpha(\boldsymbol{u}, \boldsymbol{v})$ given by

$$\alpha(\boldsymbol{u}, \boldsymbol{v}) = \min\left\{1, \frac{\pi(\boldsymbol{v})q(\boldsymbol{v}, \boldsymbol{u})}{\pi(\boldsymbol{u})q(\boldsymbol{u}, \boldsymbol{v})}\right\}.$$

A common MCMC approach involves using a *Gibbs sampler strategy* that performs iterative sampling with updating from the collection of *full conditional* distributions

$$\begin{aligned}\pi(\boldsymbol{x}_j|\boldsymbol{x}_{(j)}) &= \pi(\boldsymbol{x}_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \boldsymbol{x}_d) \\ &= \frac{\pi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)}{\pi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \boldsymbol{x}_d)}, \quad j = 1, \ldots, d\end{aligned}$$

rather than updating the components of $\boldsymbol{x}$ simultaneously. There is a vast literature on MCMC theory and applications; see [33,34] for comprehensive treatments.

MCMC re-focuses inferential interest from computing posterior analytic functional forms to producing posterior samples. It is an extremely flexible framework for computational inference that carries with it certain well-documented problems, most important amongst them being the assessment of *convergence*. It is not always straightforward to assess when the Markov chain has reached its stationary phase, so certain monitoring steps are usually carried out.

Bayesian Modeling: Examples

Three models that are especially useful in the modeling of systems biological data are *regression models*, *mixture models, and state-space models*. Brief details of each type of model follow.

Regression Models Linear regression models relate an observed response variable Y to a collection of predictor variables $X_1, X_2, \ldots, X_d$ via the model for the ith response

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ij} + \epsilon_i = \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i$$

say, or in vector form, for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathsf{T}}$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^\mathsf{T}$ is a vector of real-valued parameters, and ϵ is a vector random variable with zero-mean and variance-covariance matrix Σ. The objective in the analysis is to make inference about $\boldsymbol{\beta}$, to understand the influence of the predictors on the response, and to perform prediction for Y. The linear regression model (or General Linear Model) is extremely flexible: the design matrix $\boldsymbol{X}$ can be formed from arbitrary, possibly non-linear *basis* functions of the predictor variables. By introducing a covariance structure into Σ, it is possible to allow for dependence amongst the components of $\boldsymbol{Y}$, and allows for the possibility of modeling repeated measures, longitudinal or time-series data that might arise from multiple observation of the same experimental units.

An extension that is often also useful is to *random effect* or *mixed* models that take into account any repeated measures aspect to the recorded data. If data on an individual (person, sample, gene etc) is $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^\mathsf{T}$, then

$$\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U}_i + \epsilon_i \tag{10}$$

where $\boldsymbol{Z}$ is a $d \times p$ constant design matrix, and $\boldsymbol{U}_i$ is a $p \times 1$ vector of *random effects* specific to individual *i*. Typically the random effect vectors are assumed to be drawn from a common population. Similar formulations can be used to construct semi-parametric models that are useful for flexible modeling in regression.

Mixture Models A mixture model presumes that the probability distribution of variable $\boldsymbol{Y}$ can be written

$$f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{y}|\theta_k) \tag{11}$$

where $f_1, f_2, \ldots, f_K$ are distinct component densities indexed by parameters $\theta_1, \ldots, \theta_K$, and for all k, $0 < \omega_k < 1$, with

$$\sum_{k=1}^{K} \omega_k = 1 .$$

The model can be interpreted as one that specifies that with probability ω_k, $\boldsymbol{Y}$ is drawn from density f_k, for $k = 1, \ldots, K$. Hence the model is suitable for modeling in cluster analysis problems.

This model can be extended to an *infinite mixture model*, which has close links with *Bayesian non-parametric* modeling. A simple infinite mixture/Bayesian non-parametric model is the *mixture of Dirichlet processes* (MDP) model [35,36]: for parameter $\alpha > 0$ and distribution function F_0, an MDP model can be specified using the following hierarchical specification: for a sample of size n, we have

$$\begin{aligned} Y_i|\theta_i &\sim f_{Y|\theta}(y|\theta_i) \quad i = 1, \ldots, n \\ \theta_1, \ldots, \theta_n &\sim DP(\alpha, F_0) \end{aligned}$$

where $DP(\alpha, F_0)$ denotes a *Dirichlet process*. The $DP(\alpha, F_0)$ model may be sampled to produce $\theta_1, \theta_2, \ldots, \theta_n$ using the *Polya-Urn* scheme

$$\begin{aligned} \theta_1 &\sim F_0 \\ \theta_k|\theta_1, \ldots, \theta_{k-1} &\sim \frac{\alpha}{\alpha + k - 1} F_0 + \frac{1}{\alpha + k - 1} \sum_{j=1}^{k-1} \delta_{\theta_j} \end{aligned}$$

where δ_x is a point mass at x. For θ_k, conditional on $\theta_1, \ldots, \theta_{k-1}$, the Polya-Urn scheme either samples θ_k from F_0 (with probability $\alpha/(\alpha + k - 1)$), or samples $\theta_k = \theta_j$ for some $j = 1, \ldots, k-1$ (with probability $1/(\alpha + k - 1)$). This model therefore induces *clustering* amongst the θ values, and hence has a structure similar to the finite mixture model – the distinct values of $\theta_1, \ldots, \theta_n$ are identified as the cluster "centers" that index the component densities in the mixture model in Eq. (11). The degree of clustering is determined by α; high values of α encourage large numbers of clusters.

The MDP model is a flexible model for statistical inference, and is used in a wide range of applications such as density estimation, cluster analysis, functional data analysis and survival analysis. The component densities can be univariate or multivariate, and the model itself can be used to represent the variability in observed data or as a prior density. Inference for such models is typically carried out using MCMC or SMC methods [32,33]. For applications in bioinformatics and functional genomic, see [37,38].

State-Space Models A state-space model is specified through a pair of equations that relate a collection of *states*, $\boldsymbol{X}_t$, to *observations* $\boldsymbol{Y}_t$ that represent a system and how that system develops over time. For example, the relationship could be model led as

$$\begin{aligned} \boldsymbol{Y}_t &= f(\boldsymbol{X}_t, \boldsymbol{U}_t) \\ \boldsymbol{X}_{t+1} &= g(\boldsymbol{X}_t, \boldsymbol{V}_t) \end{aligned}$$

where f and g are vector-valued functions, and $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ are random error terms. A *linear state-space model* takes the form

$$\begin{aligned} \boldsymbol{Y}_t &= A_t\boldsymbol{X}_t + \boldsymbol{c}_t + \boldsymbol{U}_t \\ \boldsymbol{X}_{t+1} &= B_t\boldsymbol{X}_t + \boldsymbol{d}_t + \boldsymbol{V}_t \end{aligned}$$

for deterministic matrices A_t and B_t and vectors $\boldsymbol{c}_t$ and $\boldsymbol{d}_t$. The $\boldsymbol{X}_t$ represent the values of unobserved states, and the second equation represents the evolution of these states through time (see [39]).

State-space models can be used as models for scalar, vector and matrix-valued quantities. One application is evolution of a covariance structure, for example, representing dependencies in a biological network. If the network is dynamically changing through time, a model similar to those above is required but where $\boldsymbol{X}_t$ is a square, positive-definite matrix. For such a network, therefore, a probabilistic model for positive-definite matrices can be constructed from the Wishart/Inverse Wishart distributions [40]. For example, we may have for $t = 1, 2, \ldots,$

$$\begin{aligned} \boldsymbol{Y}_t &\sim \text{Normal}(0, \boldsymbol{X}_t) \\ \boldsymbol{X}_{t+1} &\sim \text{InverseWishart}(\nu_t, \boldsymbol{X}_t) \end{aligned}$$

where degrees of freedom parameter ν_t is chosen to induce desirable properties (stationary, constant expectation etc.) in the sequence of $\boldsymbol{X}_t$ matrices.

Transcriptomics and Functional Genomics

A key objective in the study of biological organization is to understand the mechanisms of the transcription of genomic DNA into mRNA that initiates the production of proteins and hence lies at the center of the functioning of the nuclear engine. In a cell in a particular tissue at a particular time, the nucleus contains the entire mRNA profile (*transcriptome*) which, if it could be measured, would provide direct insight into the functioning of the cell. If this profile could be measured in a dynamic fashion, then the patterns of gene regulation for one, several or many genes could be studied. Broadly, if a gene is “active” at any time point, it is producing mRNA transcripts, sometimes at a high rate, sometimes at a lower rate, and understanding the relationships between patterns of up- and down-regulation lies at the heart of uncovering pathways, or networks of interacting genes. *Transcriptomics* is the study of the entirety of recorded transcripts for a given genome in a given condition. *Functional genomics*, broadly, is the study of gene function via measured expression levels and how it relates to genome structure and protein expression.

Microarrays

A common biological problem is to detect *differential expression* levels of a gene in two or more tissue or cell types, as any differences may contribute to the understanding of the cellular organization (pathways, regulatory networks), or may provide a mechanism for discrimination between future unlabeled samples. An important tool for the analysis of these aspects of gene function is the *microarray*, a medium onto which DNA fragments (or *probes*) are placed or etched. Test sample mRNA fragments are tagged with a fluorescent marker, and then allowed to bond or *hybridize* with the matching DNA probes specific to that nucleotide sequence, according to the usual biochemical bonding process. The microarray thus produces a measurement of the mRNA content of the test sample for each of the large number of DNA sequences bound to the microarray as probes. Microarrays typically now contain tens of thousands of probes for simultaneous investigation of gene expression in whole chromosomes, or even whole genomes for simple organisms. The hybridization experiments are carried out under strict protocols, and every effort is made to regularize the production procedures, from the preparation stage through to imaging. Typically, replicate experiments are carried out.

Microarray experiments have made the study of gene expression routine; instantaneous measurements of mRNA levels for large numbers of different genes can be obtained for different tissue or cell types in a matter of hours. The most important aspects of a statistical analysis of gene expression data are, therefore, twofold; the analysis should be readily implementable for large data sets (large numbers of genes, and/or large numbers of samples), and should give representative, robust and reliable results over a wide range of experiments.

Since their initial use as experimental platforms, microarrays have become increasingly sophisticated, allowing measurement of different important functional aspects. Arrays containing whole genomes of organisms can be used for investigation of function, copy-number variation, SNP variation, deletion/insertion sites and other forms of DNA sequence variation (see [41] for a recent summary). High-throughput technologies similar in the form of printed arrays are now at the center of transcriptome investigation in several different organisms, and also widely used for genome-wide investigation of common diseases in humans [42,43]. The statistical analysis of such data represents a major computational challenge. In the list below, a description of details of *first* and *second* generation microarrays is given.

- **First Generation Microarray Studies**
  From the mid 1990s, comparative hybridization experiments using microarrays or gene-chips began to be widely used for the investigation of gene expression. The two principal types of array used were cDNA arrays and oligonucleotide arrays:
  - **cDNA microarrays**: In cDNA microarray competitive hybridization experiments, the mRNA levels of a genes in a target sample are compared to the

mRNA level of a control sample by attaching fluorescent tags (usually red and green respectively for the two samples) and measuring the relative fluorescence in the two channels. Thus, in a test sample (containing equal amounts of target and control material), differential expression **relative** to the control is either in terms of *up-regulation* or *down-regulation* of the genes in the target sample. Any genes that are up-regulated in the target compared to the control and hence that have larger amounts of the relevant mRNA, will fluoresce as predominantly red, and any that are down-regulated will fluoresce green. Absence of differences in regulation will give equal amounts of red and green, giving a yellow fluor. Relative expression is measured on the log scale

$$y = \log \frac{x_{\mathrm{TARGET}}}{x_{\mathrm{CONTROL}}} = \log \frac{x_{\mathrm{R}}}{x_{\mathrm{G}}} \tag{12}$$

where x_{R} and x_{G} are the fluorescence levels in the RED and GREEN channels respectively.

- **Oligonucleotide arrays:** The basic concept oligonucleotide arrays is that the array is produced to interrogate specific target mRNAs or genes by means of a number of oligo probes usually of length no longer than 25 bases; typically 10-15 probes are used to hybridize to a specific mRNA, with each oligo probe designed to target a specific segment of the mRNA sequence. Hybridization occurs between oligos and test DNA in the usual way. The novel aspect of the oligonucleotide array is the means by which the **absolute** level of the target mRNA is determined; each *perfect match* (PM) probe is paired with a *mismatch* (MM) probe that is identical to the prefect match probe **except** for the nucleotide in the center of the probe, for which a mismatch nucleotide is substituted, as indicated in the diagram below.

$$\mathrm{PM:\ ATGTATACTATT}\boxed{\mathrm{A}}\mathrm{TGCCTAGAGTAC}$$
$$\mathrm{MM:\ ATGTATACTATT}\boxed{\mathrm{C}}\mathrm{TGCCTAGAGTAC}$$

The logic is that the target mRNA, which has been fluorescently tagged, will bind perfectly to the *PM* oligo, and not bind at all to the *MM* oligo, and hence the absolute amount of the target mRNA present can be obtained as the difference $x_{PM} - x_{MM}$ where x_{PM} and x_{MM} are the measurements of for the *PM* and *MM* oligos respectively.

- **Second Generation Microarrays**
In the current decade, the number of array platforms has increased greatly. The principle of of hybridization of transcripts to probes on a printed array is often still the fundamental biological component, but the design of the new arrays is often radically different. Some of the new types of array are described below (see [44] for a summary).
  - **ChiP-Chip:** ChIP-chip (*chromatin immunoprecipitation* chip) arrays are *tiling* array with genomic probes systematically covering whole genomes or chromosomes that is used to relate protein expression to DNA sequence by mapping the binding sites of transcription factor and other DNA-binding proteins. See [45] for an application and details of statistical issues.
  - **ArrayCGH:** Array comparative genome hybridization (ArrayCGH) is another form of tiling array that is used to detect *copy number variation* (the variation in the numbers of repeated DNA segments) in subgroups of individuals with the aim of detecting important variations related to common diseases. See [46,47].
  - **SAGE:** Serial Analysis of Gene Expression (SAGE) is a platform for monitoring the patterns of expression of many thousands of transcripts in one sample, which relies on the sequencing of short cDNA tags that correspond to a sequence near one end of every transcript in a tissue sample. See [48,49,50].
  - **Single Molecule Arrays:** Single Molecule Arrays rely on the binding of single mRNA transcripts to the spots on the array surface, and thus allows for extremely precise measurement of transcript levels: see [51]. Similar technology is used for precise protein measurement and antibody detection. See [52].

Statistical Analysis of Microarray Data

In a microarray experiment, the experimenter has access to expression/expression profile data, possibly for a number of replicate experiments, for each of a (usually large) number of genes. Conventional statistical analysis techniques and principles (hypothesis testing, significance testing, estimation, simulation methods/Monte Carlo procedures) are used in the analysis of microarray data. The principal biological objectives of a typical microarray analysis are:

- **Detection of differential expression:** up- or down-regulation of genes in particular experimental contexts, or in particular tissue samples, or cell lines at a given time instant.
- **Understanding of temporal aspects of gene regulation:** the representation and modeling of patterns of changes in gene regulation over time.

- **Discovery of gene clusters:** the partitioning of large sets of genes into smaller sets that have common patterns of regulation.
- **Inference for gene networks/biological pathways:** the analysis of co-regulation of genes, and inference about the biological processes involving many genes concurrently.

There are typically several key issues and models that arise in the analysis of microarray data: such methods are described in detail in [53,54,55,56]. For a Bayesian modeling perspective, see [57].

- **Array normalization:** Arrays are often imaged under slightly different experimental conditions, and therefore the data are often very different even from replicate to replicate. This is a systematic experimental effect, and therefore needs to be adjusted for in the analysis of differential expression. A misdiagnosis of differential expression may be made purely due to this systematic experimental effect.
- **Measurement error**: The reported (relative) gene expression levels models are only in fact proxies for the true level gene expression in the sample. This requires a further level of variability to be incorporated into the model.
- **Random effects modeling**: It may be necessary to use *mixed* regression models, where gene specific *random-effects* terms are incorporated into the model.
- **Multivariate analysis**: The covariability of response measurements, in time course experiments, or between PM and MM measurements for an oligonucleotide array experiment, is best handled using multivariate modeling.
- **Testing:** One- and two-sample hypothesis testing techniques, based on parametric and non-parametric testing procedures can be used in the assessment of the presence of differential expression. For detecting more complex (patterns of) differential expression, in more general structured models, the tools of analysis of variance (ANOVA) can be used to identify the chief sources of variability.
- **Multiple testing/False discovery**: In microarray analysis, a classical statistical analysis using significance testing needs to take into account the fact that a very large number of tests are carried out. Hence significance levels of tests must be chosen to maintain a required *family-wise error rate*, and to control the *false discovery rate***.**
- **Classification:** The genetic information contained in a gene expression profile derived from microarray experiments for, say, an individual tissue or tumor type may be sufficient to enable the construction of a *classification rule* that will enable subsequent classification of new tissue or tumor samples.
- **Cluster analysis:** Discovery of subsets of sets of genes that have common patterns of regulation can be achieved using the statistical techniques of *cluster analysis* (see Sect. "Clustering").
- **Computer-intensive inference**: For many testing and estimation procedures needed for microarray data analysis, simulation-based methods (bootstrap estimation, Monte Carlo and permutation tests, Monte Carlo and MCMC) are often necessary, especially when complex Bayesian models are used.
- **Data compression/feature extraction:** The methods of principal components analysis and extended linear modeling via *basis functions* can be used to extract the most pertinent features of the large microarray data sets.
- **Experimental design**: Statistical experimental design can assist in determining the number of replicates, the number of samples, the choice of time points at which the array data are collected and many other aspects of microarray experiments. In addition, power and sample size assessments can inform the experimenter as to the statistical worth of the microarray experiments that have been carried out.

Typically, data derived from both types of microarrays are highly noise and artefact corrupted. The statistical analysis of such data is therefore quite a challenging process. In many cases, the replicate experiments are very variable. The other main difficulty that arises in the statistical analysis of microarray data is the dimensionality; a vast number of gene expression measurements are available, usually only on a relatively small number of individual observations or samples, and thus it is hard to establish any general distributional models for the expression of a single gene.

Clustering

Cluster analysis is an unsupervised statistical procedure that aims to establish the presence of identifiable subgroups (or *clusters*) in the data, so that objects belonging to the same cluster resemble each other more closely than objects in different clusters; see [58,59] for comprehensive summaries.

In two or three dimensions, clusters can be visualized by plotting the raw data. With more than three dimensions, or in the case of dissimilarity data (see below), analytical assistance is needed. Broadly, clustering algorithms fall into two categories:

- **Partitioning Algorithms:** A partitioning algorithm divides the data set into K clusters, where and the algorithm is run for a range of K-values. Partitioning methods are based on specifying an initial number of groups, and iteratively reallocating observations between groups until some equilibrium is attained. The most famous algorithm is the *K-Means* algorithm in which the observations are iteratively classified as belonging to one of K groups, with group membership is determined by calculating the centroid for each group (the multidimensional version of the mean) and assigning each observation to the group with the closest centroid. The K-means algorithm alternates between calculating the centroids based on the current group memberships, and reassigning observations to groups based on the new centroids. A more robust method uses mediods rather than centroids (that is, medians rather than means in each dimension, and more generally, any distance-based allocation algorithm could be used.
- **Hierarchical Algorithms:** A hierarchical algorithm yields an entire hierarchy of clustering for the given data set. *Agglomerative methods* start with each object in the data set in its own cluster, and then successively merges clusters until only one large cluster remains. *Divisive methods* start by considering the whole data set as one cluster, and then splits up clusters until each object is separated. Hierarchical algorithms are discussed in detail in Sect. "Hierarchical Clustering".

Data sets for clustering of N observations can either take the form of an $N \times p$ data matrix, where rows contain the different observations, and columns contain the different variables, or an $N \times N$ dissimilarity matrix, whose (i,j)th element is d_{ij}, the distance or dissimilarity between observations i and j that obeys the usual properties of a metric. Typical data distance measures between two data points i and j with measurement vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are the L_1 and L_2 Euclidean distances, and the grid-based *Manhattan distance* for discrete variables, or the *Hamming distance* for binary variables. For *ordinal* (ordered categorical) or *nominal* (label) data, other dissimilarities can be defined.

Hierarchical Clustering Agglomerative hierarchical clustering initially places each of the N items in its own cluster. At the first level, two objects are to be clustered together, and the pair is selected such that the potential function increases by the largest amount, leaving $N-1$ clusters, one with two members, the remaining $N-2$ each with one. At the next level, the optimal configuration of $N-2$ clusters is found, by joining two of the existing clusters. This process continuous until a single cluster remains containing all N items. At each level of the hierarchy, the merger chosen is the one that leads to the smallest increase in some objective function.

Classical versions of the hierarchical agglomeration algorithm are typically used with *average*, *single* or *complete* linkage methods, depending on the nature of the merging mechanism. Such criteria are inherently heuristic, and more formal *model-based* criteria can also be used. Model-based clustering is based on the assumption that the data are generated by a mixture of underlying probability distributions. Specifically, it is assumed that the population of interest consists of K different sub populations, and that the density of an observation from the the sub population is for some unknown vector of parameters. Model-based clustering is described in more detail in Sect. "Model–Based Hierarchical Clustering".

The principal display plot for a clustering analysis is the *dendrogram* which plots all of the individual data objects linked by means of a binary "tree". The dendrogram represents the structure inferred from a hierarchical clustering procedure which can be used to partition the data into subgroups as required if it is cut at a certain "height" up the tree structure. As with many of the aspects of the clustering procedures described above, it is more of a heuristic graphical representation rather than a formal inferential summary. However, the dendrogram is readily interpretable, and favored by biologists.

Model-Based Hierarchical Clustering Another approach to hierarchical clustering is *model-based clustering* (see for example [60,61]), which is based on the assumption that the data are generated by a mixture of K underlying probability distributions as in Eq. (11). Given data matrix $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^{\mathsf{T}}$, let $\gamma = (\gamma_1, \ldots, \gamma_N)$ denote the cluster labels, where $\gamma_i = k$ if the ith data point comes from the kth sub-population. In the classification procedure, the maximum likelihood procedure is used to choose the parameters in the model.

Commonly, the assumption is made that the data in the different sub-populations follow multivariate normal distributions, with mean u_k and covariance matrix Σ_k for cluster k, so that

$$\begin{aligned} f_{Y|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) &= \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{y}|u_k, \Sigma_k) \\ &= \sum_{k=1}^{K} \omega_k \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_k|^{1/2}} \\ &\qquad \exp\left\{-\frac{1}{2}(\boldsymbol{y}-u_k)^{\mathsf{T}} \Sigma_k^{-1} (\boldsymbol{y}-u_k)\right\} \end{aligned}$$

where $\Pr[\gamma_i = k] = \omega_k$. If $\Sigma_k = \sigma^2 \boldsymbol{I}_p$ is a $p \times p$ matrix, then maximizing the likelihood is the same as minimizing the sum of within-group sums of squares and corresponds to the case of hyper-spherical clusters with the same variance. Other forms of Σ_k yield clustering methods that are appropriate in different situations. The key to specifying this is the singular value or eigen decomposition of Σ_k, given by eigenvalues $\lambda_1, \ldots, \lambda_p$ and eigen vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$, as in Principal Components Analysis [62]. The eigen vectors of Σ_k, specify the orientation of the kth cluster, the largest eigenvalue λ_1 specifies its variance or size, and the ratios of the other eigenvalues to the largest one specify its shape. Further, if $\Sigma_k = \sigma_k^2 \boldsymbol{I}_p$, the criterion corresponds to hyper spherical clusters of different sizes, and by fixing the eigenvalue ratios $\alpha_j = \lambda_j/\lambda_1$ for $j = 2, 3, \ldots, p$ across clusters, other cluster shapes are encouraged.

Model-Based Analysis of Gene Expression Profiles

The clustering problem for vector-valued observations can be formulated using models used to represent the gene expression patterns via the *extended linear model*, that is, a linear model in non-linear basis functions; see, for example, [63,64] for details.

Generically, the aim of the statistical model is to capture the behavior of the gene expression ratio y_t as a function of time t. The basis of the modeling strategy would be to use models that capture the characteristic behavior of expression profiles likely to be observed due to different forms of regulation. A regression framework and model can be adopted. Suppose that Y_t is model led using a linear model

$$Y_t = \boldsymbol{X}_t \boldsymbol{\beta} + \varepsilon_t$$

where $\boldsymbol{X}_t$ is (in general) a $1 \times p$ vector of specified functions of t, and $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector. In vector representation, the gene expression profile over times $t_1, \ldots, t_T$ can be written $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon\,. \tag{13}$$

The precise form of design matrix $\boldsymbol{X}$ will be specified to model the time-variation in signal. Typically the random error terms $\{\varepsilon_t\}$ are taken as independent and identically distributed Normal random variables with variance σ^2, implying that the conditional distribution of the responses $\boldsymbol{Y}$ is multivariate normal

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_T\right) \tag{14}$$

where now $\boldsymbol{X}$ is $T \times p$ where $\boldsymbol{I}_T$ is the $T \times T$ identity matrix.

In order to characterize the underlying gene expression profile, the parameter vector $\boldsymbol{\beta}$ must be estimated. For this model, the maximum likelihood/ordinary least squares estimates of $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}}_{\text{ML}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \qquad \widehat{\sigma}^2 = \frac{1}{T-p}\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}\right)^{\mathsf{T}}\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}\right)$$

for fitted values $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{ML}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.

Bayesian Analysis in Model–Based Clustering In a Bayesian analysis of the model in (13) a joint prior distribution $\pi\left(\boldsymbol{\beta}, \sigma^2\right)$ is specified for $\left(\boldsymbol{\beta}, \sigma^2\right)$, and a posterior distribution conditional on the observed data is computed for the parameters. The calculation proceeds using Eq. (5) (essentially with $\mathcal{G}$ fixed).

$$\pi\left(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}\right) = \frac{L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}, \sigma^2\right)}{\int L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}, \sigma^2\right) \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2}$$

where $L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)$ is the likelihood function. In the linear model context, a conjugate prior specification is used where

$$\begin{aligned} \pi\left(\boldsymbol{\beta}|\sigma^2\right) &\equiv \text{Normal}\left(\boldsymbol{v}, \sigma^2 \boldsymbol{V}\right) \\ \pi\left(\sigma^2\right) &\equiv \text{IGamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right) \end{aligned} \tag{15}$$

($\boldsymbol{v}$ is $p \times 1$, $\boldsymbol{V}$ is $p \times p$ positive-definite and symmetric, all other parameters are scalars) and IGamma denotes the inverse Gamma distribution. Using this prior, standard Bayesian calculations show that conditional on the data

$$\begin{aligned} \pi\left(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2\right) &\equiv \text{Normal}\left(\boldsymbol{v}^*, \sigma^2 \boldsymbol{V}^*\right) \\ \pi\left(\sigma^2|\boldsymbol{y}\right) &\equiv \text{IGamma}\left(\frac{T+\alpha}{2}, \frac{c+\gamma}{2}\right) \end{aligned} \tag{16}$$

where

$$\begin{aligned} \boldsymbol{V}^* &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1} \\ \boldsymbol{v}^* &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right) \\ c &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{v} - \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right)^{\mathsf{T}} \\ &\qquad \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right) \end{aligned} \tag{17}$$

In regression modeling, it is usual to consider a centered parametrization for $\boldsymbol{\beta}$ so that $\boldsymbol{v} = 0$, giving

$$\begin{aligned} \boldsymbol{v}^* &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\ c &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\ &= \boldsymbol{y}^{\mathsf{T}}\left(\boldsymbol{I}_T - \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\right)\boldsymbol{y} \end{aligned}$$

A critical quantity in a Bayesian clustering procedure is the marginal likelihood, as in Eq. (8), for the data in light of the model:

$$f_Y(\boldsymbol{y}) = \int f_{Y|\boldsymbol{\beta},\sigma^2}\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2\right)\pi\left(\boldsymbol{\beta}|\sigma^2\right)\pi\left(\sigma^2\right)\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2 . \tag{18}$$

Combining terms above gives that

$$f_Y(\boldsymbol{y}) = \left(\frac{1}{\pi}\right)^{T/2}\frac{\gamma^{\alpha/2}\Gamma\left(\frac{T+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\frac{|\boldsymbol{V}^*|^{1/2}}{|\boldsymbol{V}|^{1/2}}\frac{1}{\{c+\gamma\}^{(T+\alpha)/2}} \tag{19}$$

This expression is the marginal likelihood for a single gene expression profile. For a collection of profiles belonging to a single cluster, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$, Eq. (19) can again be evaluated and used as the basis of a dissimilarity measure as an input into a hierarchical clustering procedure. The marginal likelihood in Eq. (19) can easily be re-expressed for clustered data. The basis of the hierarchical clustering method outlined in [64] proceeds by agglomeration of clusters from N to 1, with the two clusters that lead to the **greatest increase** marginal likelihood score at each stage of the hierarchy. This method works for profiles of arbitrary length, potentially with different observation time points, however it is computationally most efficient when the time points are the same for each profile.

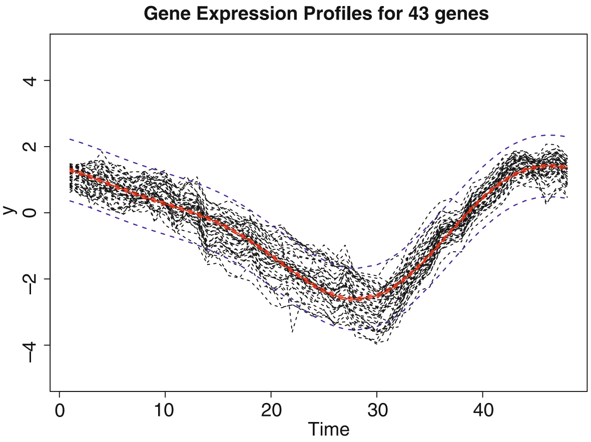
The design matrix $\boldsymbol{X}$ is typically expressed via non-linear basis functions, for example truncated polynomial splines, Fourier bases or wavelets. For T large, it is usually necessary to use a projection through a lower number of bases; for example, for a single profile, $\boldsymbol{X}$ becomes $T \times p$ and $\boldsymbol{\beta}$ becomes $p \times 1$, for $T > p$. Using different designs, many flexible models for the expression profiles can be fitted. In some cases, the linear mixed effect formulation in Eq. (10) can be used to construct the spline-based models; in such models, some of the β parameters are themselves assumed to be random effects (see [65]).

For example, in *harmonic regression*, regression in the Fourier bases is carried out. Consider the extended linear model

$$Y_t = \sum_{j=0}^{p} \beta_j g_j(t) + \varepsilon_t$$

where $g_0(t) = 1$ and

$$g_j(t) = \begin{cases}\cos(\phi_j t) & j \text{ odd}\\ \sin(\phi_j t) & j \text{ even}\end{cases}$$



Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 1

Cluster of gene expression profiles obtained using Bayesian hierarchical model-based clustering: data from the intraerythrocytic developmental cycle of protozoa *Plasmodium falciparum*. Clustering achieved using harmonic regression model with $k = 2$. Solid red line is posterior mean for this cluster, dotted red lines are point wise 95% credible intervals for the cluster mean profile, and dotted blue lines are point wise 95% credible intervals for the observations

where p is an even number, $p = 2k$ say, and $\phi_j, j = 1, 2, \ldots, k$ are constants with $\phi_1 < \phi_2 < \cdots < \phi_k$. For fixed t, $\cos(\phi_j t)$ and $\sin(\phi_j x)$ are also fixed and this model is a linear model in parameters

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathsf{T}} .$$

This model can be readily fitted to time-course expression profiles. The plot below is a fit of the model with $k = 2$ to a cluster of profiles extracted using the method described in [64] from the malaria protozoa *Plasmodium falciparum* data set described in [66].

One major advantage of the Bayesian inferential approach is that any biological prior knowledge that is available can be incorporated in a coherent fashion. For example, the data in Figure 1 illustrate periodic behavior related to the cyclical nature of cellular organization, and thus the choice of the Fourier bases is a natural one.

Choosing the Number of Clusters: Bayesian Information Criterion A hierarchical clustering procedure gives the sequence by which the clusters are merged (in agglomerative clustering) or split (in divisive clustering) according the model or distance measure used, but does not give an indication for the number of clusters that are present in the data (under the model specification). This is obviously an important consideration. One advantage of the model-based approach to clustering is that it allows the use of statistical model assessment procedures to assist in the choice of the number of clusters. A common method is to use approximate *Bayes factors* to compare models of different orders (i. e. models with different numbers of clusters), and gives a systematic means of selecting the parametrization of the model, the clustering method, and also the number of clusters (see [67]).

The Bayes factor is the posterior odds for one model against the other assuming neither is favored a priori. A reliable approximation to twice the log Bayes factor called the *Bayesian Information Criterion* (BIC), which, for model M fitted to n data points is given by

$$\mathrm{BIC}_M = -2 \log L_M(\widehat{\boldsymbol{\theta}}) + d_M \log n$$

where L_M is the Bayesian marginal likelihood from Eq. (18), $L_M(\widehat{\boldsymbol{\theta}})$ is the maximized log likelihood of the data for the model M, and d_M is the number of parameters estimated in the model. The number of clusters is not considered a parameter for the purposes of computing the BIC. The smaller (more negative) the value of the BIC, the stronger the evidence for the model.

Classification via Model–Based Clustering Any clustering procedure can be used as the first step in the construction of *classification* rules. Suppose that it, on the basis of an appropriate decision procedure, it is known that there are C clusters, and that a set of existing expression profiles $y_1, \ldots, y_N$ have been allocated in turn to the clusters. Let $z_1, \ldots, z_N$ be the cluster allocation labels for the profiles. Now, suppose further that the C clusters can be decomposed further into two subsets of sizes C_0 and C_1, where the subsets represent perhaps clusters having some common, known biological function or genomic origin. For example, in a cDNA microarray, it might be known that the clones are distinguishable in terms of the organism from which they were derived. A new objective could be to allocate a novel gene and expression profile to one of the subsets, and one of the clusters within that subset.

Let y_{ijk}, for $i = 0, 1,\ j = 1, 2, \ldots, C_i,\ k = 1, 2, \ldots, N_{ij}$ denote the kth profile in cluster j in subset i. Let y^* denote a new profile to be classified, and ξ^* be the binary classification-to-subset, and z^* the classification-to-cluster variable for y^*. Then, by Bayes Rule, for $i = 1, 2$,

$$P\left[\xi^* = i | y^*, y, z\right] \propto p\left(y^* | \xi^* = i, y, z\right) P\left[\xi^* = i | y, z\right] \tag{20}$$

The two terms in Eq. (20) can be determined on the basis of the clustering output.

Metabolomics

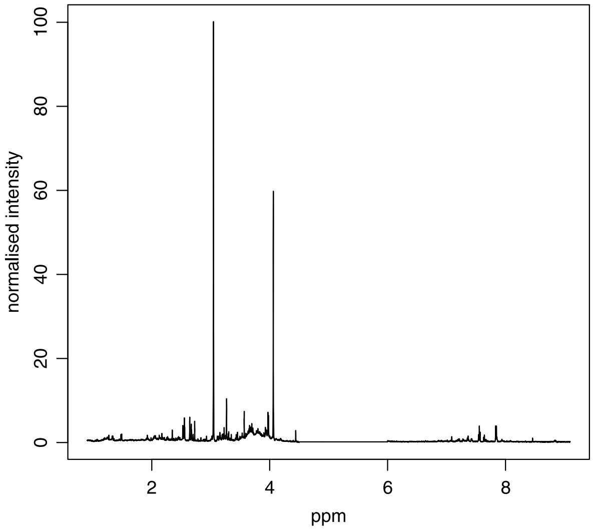
The term *metabolome* refers to the total metabolite content of an organic sample (tissue, blood, urine etc) obtained from a living organism which represents the products of a higher level of biological interaction than that which occurs within the cell. *Metabolomics* and *metabonomics* are the fields in biomedical investigation that combines the application of nuclear magnetic resonance (NMR) *spectroscopy* with multivariate statistical analysis in studies of the composition of the samples. Metabonomics is often used in reference to the static chemical content of the sample, whereas metabolomics is used to refer to the dynamic evolution of the metabolome. Both involve the measurement of the metabolic response to interventions – see for example [68] – and applications of metabolomics include several in public health and medicine [69,70].

Statistical Methods for Spectral Data

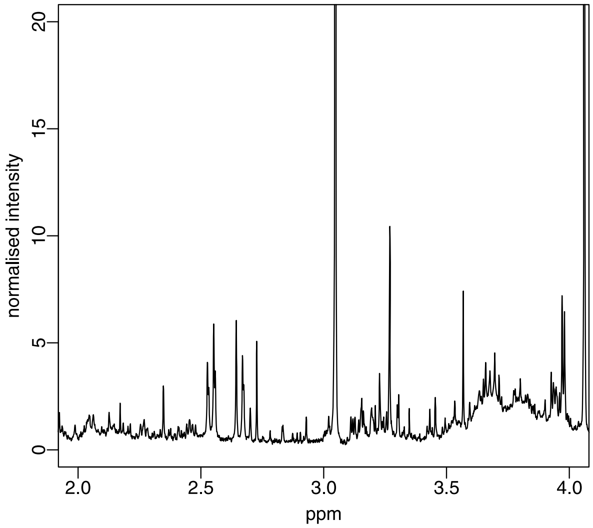
The two principal spectroscopic measurement platforms, NMR and Mass Spectrometry (MS) yield alternative representations of the metabolic spectrum. They produce spectra (or profiles) that consist of several thousands of individual measurements at different resonances or masses. There are several phases of processing of such data; preprocessing using smoothing, alignment and de-noising, peak separation, registration and signal extraction. For an extensive discussion, see [62].

An NMR spectrum consists of measurements of the intensity or frequency of different biochemical compounds (metabolites) represented by a set of *resonances* dependent upon the chemical structure, and can be regarded as a linear combination of peaks (nominally of various widths) that correspond to singletons or multiple peaks according to the neighboring chemical environment. A typical spectrum extracted from rat urine is depicted in Fig. 2 see [71]. Two dominant sharp peaks are visible.

Features of the spectra that require specific statistical modeling include multiple peaks for a single compound, variation in peak shape, and chemical shifts induced by variation in experimental pH Signals from different metabolites can be highly overlapped and subject to peak position variation due primarily to pH variations in the samples, and there are many small scale features (see Fig. 3). Statistical methods of pre-processing NMR spectra for statistical analysis which address the problems outlined above, using, for example, dynamic time warping to achieve alignment of resonance peaks across replicate spectra as a form or spectral registration form part of the necessary holistic Bayesian framework.



Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 2
A normalized rat urine spectrum. The ordinate is parts per million, the abscissa is intensity after standardization



Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 3
Magnified portion of the spectrum showing small scale features

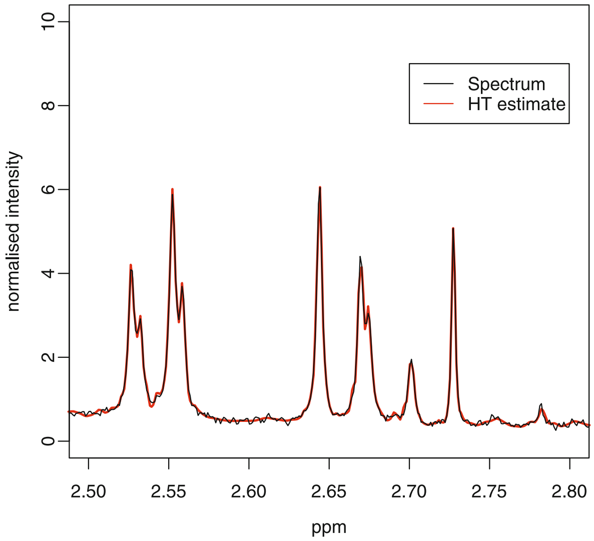
Classical statistical methods for metabolic spectra include the following:

- **Principal Components Analysis (PCA) and Regression:** a linear data projection method for dimension reduction, feature extraction, and classification of samples in an unsupervised fashion, that is, without reference to labeled cases.
- **Partial Least Squares (PLS):** a non-linear projection method similar to PCA, but implemented in a supervised setting for sample discrimination.
- **Clustering:** Clusters of spectra, or peaks within spectra, can be discovered using similar techniques to those described in Sect. "Clustering".
- **Neural Networks:** Flexible non-linear regression models constructed from simple mathematical functions that are learned from the observation of cases, that are ideal models for classification. The formulation of an neural network involves three levels of interlinked variables; *outputs*, *inputs*, and *hidden variables*, interpreted as a collection of unobserved random variables that form the hidden link between inputs and outputs.

Bayesian Approaches

The Bayesian framework is a natural one for incorporating genuine biological prior knowledge into the signal reconstruction, and typically useful prior information (about fluid composition, peak location, peak multiplicity) is available. In addition, a hierarchical Bayesian model structure naturally allows construction of plausible models for the spectra across experiments or individuals.

- **Flexible Bayesian Models:** The NMR spectrum can be represented as a noisy signal derived from some underlying and biologically important mechanism. Basis-function approaches (specifically, wavelets) have been much used to represent non-stationary time-varying signals [65,71,72,73]. The sparse representation of the NMR spectrum in terms of wavelet coefficients makes them an excellent tool in data compression, yet these coefficients can still be easily transformed back to the spectral domain to give a natural interpretation in terms of the underlying metabolites.
  Figure 4 depicts the reconstruction of the rat urine spectrum in the region between 2.5 and 2.8 ppm using wavelet methods; see [71].
- **Bayesian Time Series Models for Complex Non-stationary Signals:** See for example [74]. The duality between semi-parametric modeling of functions and latent time series models allows a view of the analysis of the underlying NMR spectrum not as a set of point wise evaluations of a function, but rather as a (time-ordered) series of correlated observations with some identifiable latent structure. Time series models, computed using dynamic calculation (filtering), provide a method for representing the NMR spectra parsimoniously.
- **Bayesian Mixture Models:** A reasonable generative model for the spectra is one that constructs the



Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 4

Wavelet reconstruction of a region of the spectrum results under the "Least Asymmetric wavelet" with four vanishing moments using the hard thresholding (HT)

> spectra from a large number of symmetric peaks of varying size, corresponding to the contributions of different biochemical compounds. This can be approximated using a finite mixture model, where the number, magnitudes and locations, of the spectral contributions are unknown. Much recent research has focused on the implementation of computational strategies for Bayesian mixtures, in particular Markov chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) have proved vital. The reconstruction of NMR spectra is a considerably more challenging area than those for which mixture modeling is conventionally used, as many more individual components are required. Flexible semi-parametric mixture models have been utilized in [75,76], whilst fully non-parametric mixture models similar to those described in Sect. "Mixture Models" can also be used [73].

A major advantage of using the fully Bayesian framework is that, once again, all relevant information (the spectral data itself, knowledge of the measurement processes for different experimental platforms, the mechanisms via which multiple peaks and shifts are introduced) can be integrated in a coherent fashion. In addition, prior knowledge about the chemical composition of the samples can be integrated via a prior distribution constructed by inspection of the profiles for training samples. At a higher level of synthesis, the Bayesian paradigm offers a method for integrating metabolomic data with other functional or structural data, such as gene expression or protein expression data. Finally, the metabolic content of tissue changes temporally, so dynamic modeling of the spectra could also be attempted.

Future Directions

Biological data relating to structure and function of genes, proteins and other biological substances are now available from a wide variety of platforms. Researchers are beginning to develop methods for coherent combination of data from different experimental processes to get an entire picture of biological cause and effect. For example, the effective combination of gene expression and metabonomic data will be of tremendous utility. A principal challenge is therefore the fusion of expression data derived from different experimental platforms, and seeking links with sequence and ontological information available. Such fusion will be critical in the future of statistical analysis of large scale systems biology and bioinformatics data sets.

In terms of public health impact of systems biology and statistical genomics, perhaps the most prominent is the study of common diseases through high-throughput genotyping of single nucleotide polymorphisms (SNPs). In genome wide association studies, SNP locations that correlate with disease status or quantitative trait value are sought. In such studies, the key statistical step involves the selection of informative predictors (SNP or genomic loci) from a large collection of candidates. Many such genome wide studies have been completed or are ongoing (see [42,43,77,78]). Such studies represent huge challenges for statisticians and mathematical modelers, as the data contain many subtle structures but also as the amount of information is much greater than that available for typical statistical analysis.

Another major challenge to the quantitative analysis of biological data comes in the form of image analysis and extraction. Many high throughput technologies rely on the extraction of information from images, either in static form, or dynamically from a series of images. For example it is now possible to track the expression level of mRNA transcripts in real-time ([79,80,81]), and to observe mRNA transcripts moving from transcription sites to translation sites (see for example [82]). Imaging techniques can also offer insights into aspects of the dynamic organization of nuclear function by studying the positioning of nuclear compartments and how those compartments reposition themselves in relation to each other through time. The challenges for the statistician are to develop real-time analysis methods for tracking and quantifying the nature and content of such images, and tools from spatial modeling and time series analysis will be required.

Finally, *flow cytometry* can measure characteristics of millions of cells simultaneously, and is a technology that offers many promises for insights into biological organization and public health implications. However, quantitative measurement and analysis methods are only yet in the early stages of development, but offer much promise (see [83,84]).

Bibliography

1. Kitano H (ed) (2001) Foundations of Systems Biology. MIT Press, Cambridge
2. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
3. Alon U (2006) An Introduction to Systems Biology. Chapman and Hall, Boca Raton
4. Edwards AWF (2000) Foundations of mathematical genetics, 2nd edn. Cambridge University Press, Cambridge
5. Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. Philos Trans R Soc Lond Ser B 213:21–87
6. Fisher RA (1922) On the dominance ratio. Proc R Soc Edinburgh 42:321–341
7. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
8. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159
9. Murray JD (2002) Mathematical Biology: I An Introduction. Springer, New York
10. Murray JD (2003) Mathematical Biology: II Spatial Models and Biomedical Applications. Springer, New York
11. Lewin B (2007) Genes, 9th edn. Jones & Bartlett Publishers, Boston
12. Spector DL (2001) Nuclear domains. J Cell Sci 114(16):2891–3
13. Bernardo JM, Smith AFM (1994) Bayesian Theory. Wiley, New York
14. Haefner JW (ed) (2005) Modeling Biological Systems: Principles and Applications, 2nd edn. Springer, New York
15. Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach. J Royal Stat Soc: Series B (Methodology) 69(5):741–796
16. Donnet S, Samson A (2007) Estimation of parameters in incomplete data models defined by dynamical systems. J Stat Plan Inference 137(9):2815–2831
17. Rogers S, Khanin R, Girolami M (2007) Bayesian model-based inference of transcription factor activity. BMC Bioinformatics 8(Suppl 2) doi:10.1186/1471-2105-8-S2-S2
18. Wilkinson DJ (2006) Stochastic Modelling for Systems Biology. Chapman & Hall (CRC), Boca Raton
19. Heron EA, Finkenstädt B, Rand DA (2007) Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. Bioinformatics 23(19):2596–2603
20. Airoldi EM (2007) Getting started in probabilistic graphical models. PLoS Comput Biol 3(12):e252
21. Husmeier D, Dybowski R, Roberts S (eds) (2005) Probabilistic Modelling in Bioinformatics and Medical Informatics. Springer, Ney York
22. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799–805
23. Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol 1:37:1–10
24. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 97:12182–12186
25. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using bayesian networks to analyze expression data. J Comput Biol 7:601–620
26. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98(11):462–467
27. Dobra A, Hans C, Jones B, Nevins J, Yao G, West M (2004) Sparse graphical models for exploring gene expression data. J Multivar Anal 90:196–212
28. Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M (2005) Experiments in stochastic computation for high dimensional graphical models. Stat Sci 20:388–400
29. Markowetz F, Bloch J, Spang R (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics 21:4026–4032
30. Eaton D, Murphy KP (2007) Exact Bayesian structure learning from uncertain interventions. Artificial Intelligence & Statistics 2:107–114
31. Robert CP (2007) The Bayesian Choice: From Decision–Theoretic Foundations to Computational Implementation. Texts in Statistics, 2nd edn. Springer, New York
32. Doucet A, de Freitas N, Gordon NJ (eds) (2001) Sequential Monte Carlo Methods in Practice, Statistics for Engineering and Information Science. Springer, New York
33. Robert CP, Casella G (2005) Monte Carlo Statistical Methods. Texts in Statistics, 2nd edn. Springer, New York
34. Gamerman D, Lopes HF (2006) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Texts in Statistical Science, 2nd edn. Chapman and Hall (CRC), Boca Raton
35. Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174
36. Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588
37. Dahl DB (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In: Do KA, Müller P, Vannucci M (eds) Bayesian Inference for Gene Expression and Proteomics. University Press, Cambridge, Chap 10
38. Kim S, Tadesse MG, Vannucci M (2006) Variable selection in clustering via Dirichlet process mixture models. Biometrika 93(4):877–893
39. West M, Harrison J (1999) Bayesian Forecasting and Dynamic models, 2nd edn. Springer, New York
40. Philipov A, Glickman ME (2006) Multivariate stochastic volatility via Wishart processes. J Bus Econ Stat 24(3):313–328
41. Gresham D, Dunham MJ, Botstein D (2008) Comparing whole genomes using DNA microarrays. Nat Rev Genet 9:291–302
42. The Wellcome Trust Case Control Consortium (2007) Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. Nat Genet 39:1329–1337

43. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
44. Liu XS (2007) Getting started in tiling microarray analysis. PloS Comput Biol 3(10):1842–1844
45. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci USA 103(33):12457–62 (2006)
46. Freeman JL et al (2006) Copy number variation: New insights in genome diversity. Genome Res 16:949–961
47. Urban AE et al (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci USA 103(12):4534–4539
48. Saha S et al (2002) Using the transcriptome to annotate the genome. Nat Biotech 20:508–512
49. Shadeo A et al (2007) Comprehensive serial analysis of gene expression of the cervical transcriptome. BMC Genomics 8:142
50. Robinson SJ, Guenther JD, Lewis CT, Links MG, Parkin IA (2007) Reaping the benefits of SAGE. Methods Mol Biol 406:365–386
51. Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ (2005) Elucidation of the Small RNA Component of the Transcriptome. Science 309(5740):1567–1569
52. Weiner H, Glökler J, Hultschig C, Büssow K, Walter G (2006) Protein, antibody and small molecule microarrays. In: Müller UR, Nicolau DV (eds) Microarray Technology and Its Applications. Biological and Medical Physics. Biomedical Engineering. Springer, Berlin, pp 279–295
53. Speed TP (ed) (2003) Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC, Bacon Raton
54. Parmigiani G, Garett ES, Irizarry RA, Zeger SL (eds) (2003) The Analysis of Gene Expression Data. Statistics for Biology and Health. Springer, New York
55. Wit E, McClure J (2004) Statistics for Microarrays: Design, Analysis and Inference. Wiley, New York
56. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds) (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health. Springer, New York
57. Do KA, Müller P, Vannucci M (2006) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge
58. Everitt BS, Landau S, Leese M (2001) Cluster Analysis, 4th edn. Hodder Arnold, London
59. Kaufman L, Rousseeuw PJ (2005) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics, 2nd edn. Wiley, Ney York
60. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL (2001) Model-based clustering and data transformation for gene expression data. Bioinformatics 17:977–987
61. McLachlan GJ, Bean RW, Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18:413–422
62. De Iorio M, Ebbels TMD, Stephens DA (2007) Statistical techniques in metabolic profiling. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of Statistical Genetics, 3rd edn. Wiley, Chichester, Chap 11
63. Heard NA, Holmes CC, Stephens DA, Hand DJ, Dimopoulos G (2005) Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. Proc Natl Acad Sci USA 102(47):16939–16944
64. Heard NA, Holmes CC, Stephens DA (2006) A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. J Am Stat Assoc 101(473):18–29
65. Morris JS, Brown PJ, Baggerly KA, Coombes KR (2006) Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. In: Do KA, Müller P, Vannucci M (eds) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge, pp 269–292
66. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 1(1):E5
67. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795
68. Nicholson JK, Connelly J, Lindon JC, Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1:153–161
69. Lindon JC, Nicholson JK, Holmes E, Antti H, Bollard ME, Keun H, Beckonert O, Ebbels TM, Reily MD, Robertson D (2003) Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxic Appl Pharmacol 187:137
70. Brindle JT, Annti H, Holmes E, Tranter G, Nicholson JK, Bethell HWL, Clarke S, Schofield SM, McKilligin E, Mosedale DE, Graingerand DJ (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. Nat Med 8:143
71. Yen TJ, Ebbels TMD, De Iorio M, Stephens DA, Richardson S (2008) Analysing real urine spectra with wavelet methods. (in preparation)
72. Brown PJ, Fearn T, Vannucci M (2001) Bayesian wavelet regression on curves with applications to a spectroscopic calibration problem. J Am Stat Soc 96:398–408
73. Clyde MA, House LL, Wolpert RL (2006) Nonparametric models for proteomic peak identification and quantification. In: Do KA, Müller P, Vannucci M (eds) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge, pp 293–308
74. West M, Prado R, Krystal A (1999) Evaluation and comparison of EEG traces: Latent structure in non-stationary time series. J Am Stat Assoc 94:1083–1095
75. Ghosh S, Grant DF, Dey DK, Hill DW (2008) A semiparametric modeling approach for the development of metabonomic profile and bio-marker discovery. BMC Bioinformatics 9:38
76. Ghosh S, Dey DK (2008) A unified modeling framework for metabonomic profile development and covariate selection for acute trauma subjects. Stat Med 30;27(29):3776–88
77. Duerr RH et al (2006) A Genome–Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science 314(5804):1461–1463
78. Sladek R et al (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885
79. Longo D, Hasty J (2006) Imaging gene expression: tiny signals make a big noise. Nat Chem Biol 2:181–182
80. Longo D, Hasty J (2006) Dynamics of single-cell gene expression. Mol Syst Biol 2:64
81. Wells AL, Condeelis JS, Singer RH, Zenklusen D (2007) Imaging real-time gene expression in living systems with single-tran-

script resolution: Image analysis of single mRNA transcripts. CSH Protocols, Cold Springer Habor

82. Rodriguez AJ, Condeelis JS, Singer RH, Dictenberg JB (2007) Imaging mRNA movement from transcription sites to translation sites. Semin Cell Dev Biol 18(2):202–208
83. Lizard G (2007) Flow cytometry analyses and bioinformatics: Interest in new softwares to optimize novel technologies and to favor the emergence of innovative concepts in cell research. Cytom A 71A:646–647
84. Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. Cytom Part A 73A(4):321–332

Complex Networks and Graph Theory

GEOFFREY CANRIGHT
Telenor R&I, Fornebu, Norway

Article Outline

Glossary

Directed/Undirected graph A set of vertices connected by directed or undirected edges. A directed edge is one-way ($A \to B$), while an undirected edge is two-way or symmetric: $A - B$.

Network For our purposes, a network is defined identically to a graph: it is an abstract object composed of vertices (nodes) joined by (directed or undirected) edges (links). Hence we will use the terms 'graph' and 'network' interchangeably.

Graph topology The list of nodes i and edges (i, j) or $(i \to j)$ defines the topology of the graph.

Graph structure There is no single agreed definition for what constitutes the "structure" of a graph. To the contrary: this question has been the object of a great deal of research—research which is still ongoing.

Node degree distribution One crude measure of a graph's structure. If n_k is the number of nodes having degree k in a graph with N nodes, then the set of n_k is the node degree distribution—which is also often expressed in terms of the frequencies $p_k = n_k/N$.

Small-worlds graph A "small-worlds graph" has two properties: it has short path lengths (as is typical of random graphs)—so that the "world" of the network is truly "small", in that every node is within a few (or not too many) hops of every other; and secondly, it has (like real social networks, and *unlike* random graphs) a significant degree of *clustering*—meaning that two neighbors of a node have a higher-than-random probability of also being linked to one another.

Graph visualization The problem of displaying a graph's topology (or part of it) in a 2D image, so as to give the viewer insight into the structure of the graph. We see that this is a hard problem, as it involves both the unsolved problem of what we mean by the structure of the graph, and also the combined technological/psychological problem of conveying useful information about a (possibly large) graph via a 2D (or quasi-3D) layout. Clearly, the notion of a good graph visualization is dependent on the use to which the visualization is to be put—in other words, on the information which is to be conveyed.

Section Here, a 'bookkeeping' definition. This article introduces the reader to all of the other articles in the Section of the Encyclopedia which is titled "Complex Networks and Graph Theory". Therefore, whenever the word 'Section' (with a large 'S') is used in this 'roadmap' article, the word refers to that Section of the Encyclopedia. To avoid confusion, the various subdivisions of this roadmap article will be called 'parts'.

Definition of the Subject

The basic network concept is very simple: objects, connected by relationships. Because of this simplicity, the concept turns up almost everywhere one looks. The study of networks (or equivalently, graphs), both theoretically and empirically, has taken off over the last ten years, and shows no sign of slowing down. The field is highly interdisciplinary, having important applications to the Internet and the World Wide Web, to social networks, to epidemiology, to biology, and in many other areas. This introductory article serves as a reader's guide to the 13 articles in the Section of the Encyclopedia which is titled "Complex Networks and Graph Theory". These articles will be discussed in the context of three broad themes: network structure; dynamics of network structure; and dynamical processes running over networks.

Introduction

In the past ten years or so, the study of graphs has exploded, leaving forever the peaceful sanctum of pure mathematics to become a fundamental concept in a vigorous, ill-defined, interdisciplinary, and important field of study. The most common descriptive term for this field is the "study of complex networks". This term is distinguished from the older, more mathematically-bound term "graph theory" in two ways. First – and perhaps most important – this new field is not just theoretical; to the contrary, the curious researcher finds that there is an enormous variety of empirically obtained graphs/networks (we use the terms interchangeably here) available as datasets on the Web. That is, one studies real, measured graphs; and not surprisingly, this empirical connection gives the endeavor much more of an applied flavor also.

The second distinction is perhaps not well motivated by the words used; but the "study of complex networks" typically means studying networks whose structure deviates in important ways from the "classical random graphs" of Erdős and Rényi [10,11]. We note that these two points are related: as one turned to studying real networks [18], one found that they were not described by classical random graphs [24,25]; hence one was forced to look at other kinds of structures.

This article is a reader's guide to the other articles in the Section entitled "Complex Networks and Graph Theory". The inclusion of both terms was very deliberate: graph theory gives the mathematical foundation for the more messy endeavor called "complex networks"; and the two fields have a strong and fruitful interaction. In this article I will describe the 13 other articles of this Section.

These 13 articles amply document the interdisciplinary nature of this exciting field: we find represented mathematics, biology, the Web and the Internet, software, epidemiology, and social networks (with the latter field also having its own Section in the Encyclopedia). For this reason, I will not define the parts of this article by field of study, but rather by general *themes* which run through essentially all studies of networks (at least in principle, and often in fact). In each part of this article I will point out those articles which represent that part's theme to a significant degree. Hence these themes are meant to tie together all of the articles into a simple framework.

In the next part "Graphs, Networks, and Complex Networks", I will concisely present the basic terminology which is in use. Then in entitled part the "Structure of Networks" I discuss the knotty question of the *structure* of a graph; this problem is the first theme, and it is very much unfinished business. In part "Dynamical Network Structures" we look at network topologies (and structures) which are *dynamic* rather than static. This is clearly an important theme, since (i) real empirical networks are necessarily dynamic (on some – often short – time scale), and (ii) the study of how networks grow and evolve can be highly useful for understanding how they have the structure we observe today. Then in part "Dynamical Processes on Networks" we look at the very large question of dynamical processes which take place *over* networks. Examples which illustrate the importance of this topic are epidemic spreading over social and computer networks, and the activation/inhibition processes going on over gene (or neural) networks. Clearly the progress of such dynamical processes may be strongly dependent on the underlying network structure; hence we see that all of these themes (parts "Structure of Networks"–"Dynamical Processes on Networks") are tightly related.

In part "Graph Visualization" we look briefly at an important but somewhat 'orthogonal' theme, namely, the problem of *graph visualization:* given a graph's topology, how can one best present this information visually to a human viewer? Finally, part "Future Directions" offers a very brief, modest, and personal take on the very large question of "future directions" in research on complex networks.

Graphs, Networks, and Complex Networks

Graphs

One of the earliest applications of graph theory may be found in Euler's solution, in 1736, of the problem called 'The seven bridges of Königsberg'. Euler considered the problem of finding a continuous path crossing each of the seven bridges of Königsberg exactly once, and solved the problem by representing each connected piece of land as a vertex (node) of a graph, and each bridge as an undirected (two-way) link. (For a nice presentation of this problem, see http://mathforum.org/isaac/problems/bridges2.html.) This little problem is an excellent example of the power of mathematics to extract understanding via abstraction: the lay person may stare at the bridges, islands etc, and try various ideas—but reducing the entire problem to an abstract graph, composed only of nodes and links, aids the application of pure reason, leading to a final and utterly convincing solution.

To study graphs is to study discrete objects and the relationships between them. Hence graph theory may be regarded as a branch of combinatorics. Erdős and Rényi [10,11] founded the study of "classical random graphs". These graphs are specified by the node number N (which typically is assumed to grow large), and by links laid down at random between these nodes, subject to various constraints. One such constraint is that every node have exactly k links—giving a (k—) *regular random graph*. A more relaxed constraint is simply to specify m links in total (so that the average node degree is $\langle k \rangle = m/N$). Further relaxing of this constraint gives that every possible link, out of the set of

$$\begin{bmatrix} N \\ 2 \end{bmatrix} = N(N-1)/2$$

possible links, is included with probability p. This gives the average node degree (now averaged over many graphs) as $\langle k \rangle = \langle m \rangle / N = p(N-1)/2$. While all of these types of classical random graphs are similar "in spirit", those with the fewest constraints are those for which it is easiest to prove things.

We are fortunate to have in our Section the article ► Random Graphs, A Whirlwind Tour of by Fan Chung. In this article, the reader is given a good introduction to classical random graphs, along with a thorough presentation of the more modern theory of random graphs with a new, and more realistic, type of constraint—namely that they should, on average, have a given *node degree distribution*. This new theory makes the study of random graphs extremely relevant for today's empirically-anchored research: many empirical graphs are characterized by their node degree distribution, and many of these in fact have a *power-law degree distribution*, such that the number n_k of nodes having degree k varies with k by a power law: $n_k \sim k^{-\beta}$. This work is useful, precisely because a random graph with a given node degree distribution is the "most typical" graph of the set of graphs with that degree distribution. Hence, statements about such random graphs are statements about typical graphs with the same degree distribution—unless and until we know more about the empirical graphs.

Networks

As noted earlier in this introduction, we consider the terms 'network' and 'graph' to be interchangeable. Nevertheless there is a bit more to be said about the term. The 'network' concept motivates and infuses a vigorous and lively research activity that has more or less exploded since (roughly) the work of Watts and Strogatz [24,25]. Much of their work was motivated by the 'small worlds problem'. The latter dates back to the work of Milgram [18] (and even earlier). Milgram posed the question: how far is it from Kansas (or Nebraska) to Boston, Massachusetts—when the distance is measured in 'hops' between people who know one another? Modern language would rephrase this question as follows: is the US acquaintanceship network a 'small world'? Milgram's answer was 'yes': after disregarding the chains (of letters—that was the mechanism for a 'hop') that never reached the target, the average path length was roughly 5–6 hops—a 'small world'.

The explosion of interest in the last 10 years is well documented in the set of references in "Books and Reviews" (below).

We also include in this part an introductory discussion of directed graphs. Directed graphs have directed links: it no longer suffices to say "i and j are linked", because there is a *directionality* in the linking: $(i \to j)$ or $(j \to i)$ (or both). Some of the mathematical background for understanding directed graphs is provided in the article in this Section by Berman and Shaked-Monderer (► Non-negative Matrices and Digraphs). One quickly finds, upon coming in contact with directed graphs, that one is in a rather different world—the mathematics has changed, the structures are different, and one's intuition often fails.

Directed graphs are however here to stay, and well worth study. We cite two classic examples to demonstrate the extreme relevance of directed graphs. First, there is early and pioneering work by Stuart Kauffman [14,15] on genetic regulatory networks. These form directed graphs, because the links express the fact that gene $G1$ regulates gene $G2$ $(G1 \to G2)$—a relationship which is by no means symmetric. Understanding gene regulation and expression is a fundamental problem in biology—and we see that the problem may be usefully expressed as one of understanding *dynamics on a directed graph*. The article in this Section by Huang and Kauffman (► Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination) brings us up to date on this exciting problem.

A more well known directed graph, which plays a role for many of us in our daily lives, is the *Web graph* [1,2,8]. We navigate among Web pages using hyperlinks—one-way pointers taking us (e.g.) from page A to page B $(A \to B)$. The utility of the Web as a source of information—and as a platform for interaction—is enormous. Hence the Web is well worth intense study, both intellectually and practically.

Hyperlinks are useful, not only for navigation, but also to aid in *ranking* Web pages. This utility was clearly pointed out in two early papers on Web link analysis [7,17]. "Web link analysis" may be viewed most simply as a process which takes (all or part of) the Web graph as input, and which gives 'importance scores' for each Web page as output. The PageRank approach to link analysis of Brin and Page [7] has almost certainly played a significant role in the meteoric rise of the Web search company Google. Hence there is both practical and commercial value in understanding the Web graph. We have two complementary papers in this Section which treat this important area: that by Adamic on ► World Wide Web, Graph Structure, and that by Bjelland et al. on ▶ Link Analysis and Web Search.

Complex Networks

Now we come to the last substantial word in the title of this Section: 'complex'. This is a word that (as yet) lacks a precise scientific meaning. There are in fact too many definitions; in other words, there is no generally-agreed precise definition. For a good overview of work on this thorny problem we refer the reader to [4].

Fortunately, for the purposes of this Section, we need only a very simple definition: 'complex networks' are those which are *not* well modeled by classical random graphs (described above, and in the article by Fan Chung). The use of the term 'complex networks' is very widespread—while examples which give a clear definition are less common. We note that the definition given here is also cited by Dorogovtsev in his article in this Section (▶ Growth Models for Networks).

Structure of Networks

We have noted already that there is no single, simple answer to the question "what is the structure of this graph"? With this caveat in mind, we offer the reader a guide to the articles in this Section which address the structure of networks.

Undirected Graphs

We have already noted the article by Fan Chung (▶ Random Graphs, A Whirlwind Tour of), giving an up-to-date overview of the properties of random graphs with a given node degree distribution—with the well-studied case being power-law graphs. She cites a number of experimental results which indicate that the experimental exponents β (taken from the power-law degree distributions) fall in one range ($2 < \beta < 3$) for social and technological networks, and another, rather distinct range ($\beta < 2$) for biological networks. This may be regarded as a (quantitative) structural difference between these two types of network; Chung offers an explanation based on qualitatively distinct growth mechanisms (see the next part of this article).

Fortunato and Castellano (▶ Community Structure in Graphs) offer an excellent overview of another broad approach to network structure. Here the idea is to understand structure in terms of substructure (and sub-substructure, etc). That is: here is a network. Can we identify subgraphs—possibly overlapping, possibly disjoint—of this network that in some sense "belong together"? In other words: how can one identify the *community structure* of a graph? The review of Fortunato and Castellano is on the one hand very thorough—and on the other hand makes clear that there is no one agreed answer to this question. There are indeed almost as many answers as there are theoretical approaches; and this problem has received a lot of attention. Fortunato and Castellano note that there is currently a "favorite" approach, defined by finding subgraphs with high *modularity* [21]. Roughly speaking, a subgraph with high modularity has a higher density of internal links than that found for the same subgraph in a randomized 'null model' for the same graph. Fortunato and Castellano give a careful discussion of the strengths and weaknesses of this approach to community detection, as well as of many others.

Since the work of Watts and Strogatz [24,25], the definition of a 'small-world graph' has included two criteria. First, one must have the short average path length which Milgram found in his experiments, and which gives rise to the term 'small worlds'. However, one also insists that a 'true' small-world graph should locally resemble a real social network, in that there is a significant degree of *clustering*. Roughly speaking, this means that the probability that two of my acquaintances know each other is higher than random. More mathematically, to say that a graph G has high clustering means that the incidence of *closed triangles* of links in G is higher than expected in a randomized 'null model' for G.

A closed triangle is a small, simply defined subgraph, and in studying clustering one studies the statistics of the occurrence of this small subgraph. A more general term for this type of small subgraph is a *motif* [19]. Much as with clustering and triangles, one defines a set of motifs, and then generates a *significance profile* for any given network, comparing the frequency of each motif in the profile to that of a corresponding random graph. Valverde and Solé (▶ Motifs in Graphs) offer a stimulating overview of the study of motifs in networks of many types—both directed and undirected. They point out a remarkable consistency of motif significance profiles across networks with very different origins—for example, a software graph and the gene network of a bacterium—and argue that this consistency is best understood in terms of historical accident, rather than in terms of functionality.

The article by Liljeros (▶ Human Sexual Networks) looks at empirical human sexual networks. He addresses the evidence for and against the notion that such networks are power-law (also known as "scale free"). This question is important for understanding epidemic spreading over networks—especially in the light of the results of Pastor-Satorras and Vespignani [23], which showed that epidemic spreading on power-law networks is more difficult to stop than was predicted by earlier models using the well-mixed (all-to-all) approximation. Liljeros examines carefully what is known about the structure of human sexual networks, noting the great difficulty inherent in gathering extensive and/or reliable data.

The article by He, Siganos, and Faloutsos (▶ Internet Topology) looks at a very different empirical network, namely the physical Internet, composed (at the lowest level) of routers and cables. The naïve newcomer might assume that the Internet, being an engineered system, is fully mapped out, and hence its topology should be readily "understood". The article by He et al. presents the reality of the Internet: it is largely self-organized (especially at the higher level of organization, the 'Autonomous System' or AS level); it is far from trivial to experimentally map out this network—even at the AS level; and there is not even agreement on whether or not the AS-graph is a power-law graph—which is after all a rather crude measure of structure of the network. He et al. describe recent work which offers a neat resolution of the conflicting and ambiguous data on this question. They then go on to describe more imaginative models for the structure of the Internet, going beyond simply the degree distribution, and having names like 'Jellyfish' and 'Medusa'.

Directed Graphs

A generic directed graph is immediately distinguished from its undirected counterparts in that a natural unit of substructure is obvious, and virtually always present. That is the 'strongly connected component' or SCC (termed 'class' by Berman and Shaked-Monderer). That is, even when a directed graph is connected, there are node pairs which cannot reach one another by following directed paths. An SCC C is then a maximal set of nodes satisfying the constraint that every node in C is reachable, via a directed path, from every other node in C. The SCCs then form disjoint sets (equivalence classes), and every node is in one SCC. In short, the very notion of 'reachability' is more problematic in directed graphs: all nodes in the same SCC can reach one another, but otherwise, all bets are off!

Lada Adamic (▶ World Wide Web, Graph Structure) gives a good overview of what is known empirically about the structure of the *Web graph*—that abstract and yet very real object in which a node is a Web page, and a link is a (one-way) hyperlink. The Web graph is highly dynamic, in at least two ways: pages have a finite lifetime, with new ones appearing while old ones disappear; also, many Web pages are dynamic, in that they generate new content when accessed—and can in principle represent an infinite amount of content. Also, of course, the Web is huge. Lada Adamic presents the problems associated with '*crawling*' the Web to map out its topology. The reader will perhaps not be surprised to learn that the Web graph obeys a power law—both for the indegree distribution and for the outdegree distribution. Adamic discusses several other measures for the structure of the Web graph, including its gross SCC structure (the 'bow tie'), its diameter, reciprocity, clustering, and motifs. The problem of path lengths and diameter is less straightforward for directed graphs, since the unreachability problem produces infinite or undefined path length for many pairs.

A nice bonus in the article by Adamic is the discussion of some popular and well studied *subgraphs* of the Web graph—query connection graphs, Weblogs, and Wikipedia. Query connection graphs are subgraphs built from a hit list, and can give useful information about the query itself, and the likelihood that the hit list will be satisfactory. Weblogs and Wikipedia are well known to most Web-literate people; and it is fascinating to see these daily phenomena subjected to careful scientific analysis.

The article by Bjelland et al. (▶ Link Analysis and Web Search) has perhaps the closest ties to the mathematical presentation of Berman and Shaked-Monderer (▶ Non-negative Matrices and Digraphs). This is because Web link analysis tends to focus on the *principal eigenvector* of the graph's adjacency matrix (or of some modification of the adjacency matrix), while Berman and Shaked-Monderer discuss in some detail the spectral properties of this matrix, and give some results about the principal eigenvector. The latter yields *importance* or *authority scores* for each page. These scores are the principal output of Web link analysis; and in fact Berman and Shaked-Monderer cite PageRank as an outstanding example of an application of the theory [22]. Bjelland et al. explain the logic leading to the choice of the principal eigenvector, as a way of 'harvesting' information from the huge set of 'collective recommendations' that the hyperlinks constitute. They also present the principal approaches to link analysis (the 'big three'), and place them in a simple logical framework which is completed by the arrival of a new, fourth approach. In addition, Bjelland et al. discuss a number of technical issues related to Web link analysis—which is, after all, a practical and commercial field of endeavor, as well as an object for research and understanding.

Jennifer Dunne (▶ Food Webs) presents a rather special type of directed graph taken from biology: the food web. She offers a very concise definition of this concept in her glossary, which we reproduce here: "the network of feeding interactions among diverse co-occurring species in a particular habitat". We note that most feeding interactions are one-way: bird B eats insect I, but insect I does not eat bird B. However, food webs are rather special among empirical directed graphs, in that they have a lower incidence of loops than that found in typical directed graphs. Early work indicated (or even assumed) that a food web is loop-free; but this has been shown not to be strictly true. (A simple, but real example of a loop is cannibalism: A eats A; but much longer loops also exist.) The field faces (as many do) a real problem in getting good empirical data, and empirically obtained food webs tend to be small. For example, Table 1 of Dunne presents data for 16 empirical food webs, ranging in size from 25 nodes to 172.

Our second biological application of directed graphs is presented by Huang and Kauffman (▶ Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination). This article presents gene regulatory networks (GRN). The directed link ($G1 \rightarrow G2$) in a GRN expresses the fact that gene $G1$ regulates (by inhibition or activation) the expression of gene $G2$, via intermediate proteins (typically transcription factors). The links are thus inherently one way, although reciprocal links ($G1 \leftrightarrow G2$) do occur. The article of Huang and Kauffman is very comprehensive: the structure of a GRN is only their starting point, as they seek to understand and model the *dynamical development process*, in which the set of on/off states of the cell genome moves towards an *attractor* (a steady or quasi-steady state), which represents a mature and stable cell type.

Returning to structure (of GRNs), we again face a severe data-extraction problem, which is compounded by the fact that the 'interaction modality' (activation or inhibition, plus or minus) of each link must also be known before one can study dynamical development of gene expression over the GRN. In short: one needs to know more than just the presence and direction of the links; one needs their type. Huang and Kauffman give a thorough discussion of these problems, and argue for studies using an "ensemble approach". This is much like the random graph approach, in that it takes a set of statistical constraints which are empirically determined, and then studies (often via simulation) a randomly generated ensemble of graphs which satisfy the statistical constraints. In short, the ensemble approach takes the structure as determined by these constraints (node degree distribution, etc), and then studies random graphs with this structure. Note that both the graph topology and the interaction modalities of the links are randomized in this ensemble approach.

Dynamical Network Structures

We have already had a lot to say about the structure of networks—and yet we have left out one important dimension, namely *time*. Empirical networks change over time, so that most measurements that map out such a network are taking a snapshot. Now we will look explicitly at studies addressing the dynamical evolution of networks.

One classic study is the paper by Barabasi and Albert [5]. Here the preferential-attachment (or "rich get richer") model was introduced as an explanation for the ubiquitous power-law degree distributions. In other words, a *growth* (developmental) model was used to explain properties of *snapshots* of "mature" networks. In the preferential-attachment model, new nodes which join a network link to existing nodes with a biased probability distribution—so that the probability of linking to an existing node of degree k is proportional to k.

This simple model indeed gives (after long time) a power-law distribution; and the ideas in [5] stimulated a great interest in various growth models for networks. We are fortunate to have, in this Section of the Encyclopedia, the article by Sergey Dorogovtsev entitled ▶ Growth Models for Networks. This article is of course very short compared to the volume [9]—but it offers a good overview of the field, and a thorough updating of that volume, covering a broad range of questions and ideas, including the simple linear preferential attachment model, and numerous variations of it. Also, a distinctly different class of growth models, termed "optimization based models", is presented, and compared to the class of models involving forms of preferential attachment. In optimization based models, new nodes place links so as to optimize some function of the resulting network properties.

We also mention, in the context of growth models, the article by Fan Chung (▶ Random Graphs, A Whirlwind Tour of). As noted above, she has pointed out a tendency for biological networks to have significantly smaller exponents than technological networks; and she includes a good discussion of both preferential attachment (which tends to give the larger, technological, exponents) and duplication models. The latter involve new nodes "copying" or duplicating (probabilistically) the links of existing nodes. Chung observes that such duplication mechanisms do exist in biology, and also shows that they can give exponents in the appropriate range.

Huang and Kauffman (▶ Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination) also offer a limited discussion of evolution of the genome—not so much as a topic in itself, but because (again) understanding genome evolution can help in understanding—and even determining—the structure of today's genome. They mention both duplication and preferential attachment. Also, it is quite interesting to note the parallels between their discussion and that of Valverde and Solé (▶ Motifs in Graphs)—in that Huang and Kauffman also argue that many observed structures may be due, not to selection pressure or enhanced functionality, but simply to physical constraints on the evolutionary process, and/or to historical accident. The duplication mechanism in fact turns up again in the article by Valverde and Solé, who argue that this growth mechanism can largely account for the high frequency of occurrence of some motifs.

Finally, we note that growth is not the only dynamical process operative in networks. Just as the mature body is constantly shedding and regenerating cells, many mature networks are subject to constant small topology changes. One interesting class of non-growth dynamics is the *attack*. That is: how robust is a given network against a systematic attack which deletes nodes and/or links? A classic study in this direction is [3], which studied the attack tolerance of scale-free networks. There are many reasons to believe that such networks are very well connected, and this study added more: attacking random nodes had little effect on the functionality of the network (as crudely measured by the network diameter, and by the size of the largest surviving component). Thus such networks may be termed "robust"—but only with regard to this kind of "uninformed" attack. The same study showed that a "smart" attack, removing the nodes in descending order of node degree (i. e., highest degree first), caused the network to break down much more rapidly—thus highlighting once again the crucial role played by the 'hubs' in power-law networks.

Jennifer Dunne (▶ Food Webs) reports some studies of the robustness of food webs to attack—where here the 'attack' is extinction of a node (species). The dynamics is slightly different from that in the previous paragraph, however, because of the phenomenon of secondary extinction: removal of one species can cause another (or several others) to go extinct as well, if they are dependent on the first for their food supply. Also, food webs are not well modeled by power-law degree distributions. Nevertheless, the cited studies indicate that removing high-degree species again causes considerably more damage (as measured by secondary extinctions and web fragmentation) than removing low-degree species. Dunne also reports studies seeking to simulate the effects of "ecologically plausible" extinction scenarios; here we find that the studied food webs are in fact very robust to such extinction scenarios—a result which (perhaps) confirms our prejudice, that today's ecosystems are here because they are robust.

Dynamical Processes on Networks

Our next theme is also about dynamics—not *of* the network topology, however, but *over* it. That is, it is often of interest to study *processes* which occur on the nodes (changing their state), and which transmit something (information) over the links, so that a change in one node's state induces a change in other nodes' states.

We can render these ideas less abstract by considering the concrete example of epidemic spreading. The nodes can be individuals (or computers, or mobile phones), and the network is then a network of social contacts (or a computer or phone network). The elementary point that the disease is spread via contact is readily captured by the network model. A classic study in this regard, which strongly underscored the value of network models, was that (mentioned earlier) of Pastor-Satorras and Vespignani [23]. Here it was shown that classical threshold thinking about epidemic spreading fails when the network is scale-free: the effective threshold is *zero*. This study, in yet another way, revealed that such networks are extremely well connected—and it stimulated a great deal of thought about prevention strategies.

We have already mentioned the article by Liljeros (▶ Human Sexual Networks) in this Section, with its evidence for power-law, or nearly power-law, degree distributions in human sexual networks. Liljeros offers a sober and careful discussion of the implications of the theoretical result just cited, for understanding and preventing the spread of sexually transmitted diseases on finite networks. One consequence of a power-law node degree distribution—that a few individuals will have an *extremely* high number of contacts—may seem at first glance implausible or even impossible; and yet, as Liljeros points out, careful empirical studies have tended to support this prediction.

The human sexual network is of course not static; people change partners, and in fact people with many partners tend to change more often. Hence, the dynamics of the network topology must be considered, including how it affects the dynamics of the epidemic spreading process going on over the topology. A term which captures this interplay is "concurrence"—which tells, not how many partners you have, but how many partners you have "simultaneously"—i. e., within a time window which allows the passing of the disease from one contact to another. Considering concurrence leads to another type of graph, termed a 'line graph'—a structure in which contact relationships become nodes, which are linked when they are concurrent.

The reader will not be surprised to hear that epidemic spreading over directed graphs is qualitatively different from the same process over undirected graphs. The topic has real practical interest, because—as pointed out by Kephart and White [16] and by Newman, Forrest and Balthrop [20]—the effective network over which computer viruses propagate is typically directed. We invite the interested reader to consult these sources—and to note the striking similarities between the picture of the 'email address graph' in [20] and the gross structure of the Web graph [8] (and Figure 1 of Adamic).

Another classic study of dynamical processes on networks is that by Duncan Watts [24,25] on synchronization phenomena over networks. Our Section includes a thorough and up-to-date survey of this problem, primarily from the theoretical side, by Chen et al. (▶ Synchronization Phenomena on Networks). The field is large, as there exists a wide variety of dynamical processes on nodes, and types of inter-node coupling, which may (or may not) lead to synchronization over the network. Chen et al. focus on three main themes which neatly summarize the field. First, there is the synchronization process itself, and the theory which allows one to predict when synchronization will occur. Next comes the question of how the network structure affects the tendency to synchronization. The third theme is a logical followup to the second: if we know something about how structure affects synchronization, can we not find design methods which enhance the tendency to synchronize? This latter theme includes a number of ingenious methods for 'rewiring' the network in order to enhance its synchronizability. For example: it is found that a very strong community structure (recall the article by Castellano and Fortunato) will inhibit synchronization; and so one has studied networks called *'entangled networks'*, which are systematically rewired so as to have essentially no community structure, and so are optimally synchronizable.

Dunne (▶ Food Webs) discusses dynamics of species number over food webs. That is, the links telling "who eats who" also mediate transfers of biomass—a simple dynamical process. And yet—as we know from the dynamics of the simple two-species Lotka–Volterra equations [13]—simple nonlinear rules can give rise to complex behavior. Dunne reports that the behavior does not get more simple as one studies larger networks; one finds the same types of asymptotic behavior—equilibrium, limit cycle, and chaotic dynamics. It is a nontrivial task to study nonlinear dynamical models over tens or hundreds of nodes, and at the same time anchor the theory and simulation in reality. One approach to doing this has been to insist that the resulting model satisfy certain stability criteria (hence conforming to reality!); these criteria are framed in terms of 'species persistence' (not too many extinctions) and/or 'population stability' (limited fluctuations in species mass for all species).

A dynamical process also plays a central role in the discussion of Huang and Kauffman (▶ Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination). In a simplified but highly nontrivial model, the state of the gene network is modeled by a vector $\boldsymbol{S}(t)$, which takes binary values (0 or 1, i. e., 'off' or 'on') at each node in the network, at each time t. Also, the interactions between genes are modeled by Boolean truth tables—i. e., a given gene's (binary) output state is some Boolean function of all of its input states. The resulting 'Boolean network model' is at once very simple (discrete time, discrete binary states), and yet very complex, in the sense that it is impossible to predict the behavior of such models without simulating them. Huang and Kauffman describe the three regimes of dynamical behavior found: an 'ordered' regime, a 'chaotic' regime, and an intermediate 'critical' regime. While the ordered regime gives stable behavior, Huang and Kauffman argue that biology favors the critical regime. One gets the flavor of their argument if one recalls that the same genome must be able to converge to many different cell types—so its dynamics cannot be *too* stable—and yet those same cell types must be 'stable enough'. The discussion of the latter point includes a remarkable recent experimental study which graphically shows the return of two cell populations to the same 'preferred' stable attractor, after two distinct perturbations, via two distinct paths in state space.

Valverde and Solé (▶ Motifs in Graphs) also discuss dynamical processes, in terms of motifs. They give a clear picture for the simple test case of the three-node motif called 'Feed Forward Loop' or FFL. Here again—as for the GRNs of Huang and Kauffman—the modality (activation/inhibition) of the three links must be considered, and the resulting possibilities are either 'coherent' (non-conflicting) or 'incoherent'. It is interesting to note that the FFL motif has been studied both via simulation, and experimentally, in real gene networks. Valverde and Solé also make contact with the article of Chen et al. (▶ Synchronization Phenomena on Networks) by briefly discussing the connection between network synchronizability and the distribution of motif types.

Graph Visualization

The problem of graph visualization is a marvelous mixture of art and science. To produce a good graph visualization is to translate the science—what is known factually and analytically about the graph—into a 2D (or quasi-3D) image that the human brain can appreciate. Here the word 'appreciate' can include both understanding and the experience of beauty. Both of these aspects are present in great quantities in the many figures of this Section. We invite the reader interested in visualization to visit and compare the following figures: Figure 2 in Fan Chung's article; Figures 1 and 3 in Dunne; Figures 2 and 5 in Liljeros; Figure 17 in Chen et al.; and Figures 1, 2, and 6b in Valverde and Solé.

Graph visualization is included in the articles of this Section because it is a vital tool in the study of networks, and also an unfinished research challenge. The article by Vladimir Batagelj (▶ Complex Networks, Visualization of) offers a fine overview of this large field. This article in fact has a very broad scope, touching upon many visualization problems (molecules, Google maps) which may be called "information visualization". The connection to networks (no pun intended) is that information is readily understood in terms of connections between units—i. e., as a network.

We note that one natural way of 'understanding' a graph is in terms of its *communities*; and since community substructure offers a way (perhaps even hierarchical) of 'coarse-graining' the network, methods for defining communities often lead naturally to methods for graph visualization. One example of this may be found in Figure 10 of Fortunato and Castellano (▶ Community Structure in Graphs); for other examples, see [12] and [6]. Batagelj describes a wide variety of graph visualization methods; but a careful reading of his article shows that many methods which are useful for large graphs are dependent on finding and exploiting some kind of community structure. The terms used for approaches in this vein include 'multilevel algorithms', 'clustering', 'block modeling', coarse graining, partitions, and hierarchies.

The lessons learned from Fortunato and Castellano are brought home again in the article by Batagelj: there is no 'magic bullet' that gives a universally satisfying answer to the problem of visualizing large networks. To quote Batagelj in regard to an early graph visualization: "Nice as a piece of art, but with an important message: there are big problems with visualization of dense and/or large networks." A more recent example is a 2008 visualization of the Internet: here there is a 'natural' unit of coarse graining, namely the autonomous system or AS level (recall the article by He et al. on Internet Topology), so that the network of almost 5 million nodes reduces to 'only' about 18 000 ASes. Yet the resulting visualization (Figure 4 of Batagelj) clearly reveals that 18 000 nodes is still 'large' relative to human visual processing capacity.

For other beautiful and mystifying examples of this same point, I recommend figures 5 and 7 in the same article. The difficulty of visualizing large networks is perhaps most succinctly captured in the outline of Batagelj's stimulating article. Besides the normal introductory parts, we find only two others: 'Attempts' and 'Perspectives'. Yet the article is by no means discouraging—it is rather fascinating and inspiring, and I encourage the reader to read and enjoy it.

Future Directions

The many articles related to "Complex Networks and Graph Theory" have each offered their own view of 'Future directions' for the corresponding field of study. Therefore it would be both redundant and presumptuous for me to attempt the same task for all of these fields. Instead I will offer a very short and entirely personal assessment of the 'Future' for the broad field of complex networks and graph theory.

I view the field somewhat as a living organism: it is highly dynamic, and still growing vigorously. New ideas continue to pop up—many of which could not be covered in this Section, simply due to practical limitations. Also, there is a fairly free flow of ideas across disciplines. The reader has perhaps already gotten a feeling for this cross-boundary flow, by seeing the same basic ideas crop up in many articles, on problems coming for distinctly different traditional disciplines. In short, I feel that the interdisciplinarity of this field is real, it is vigorous and healthy, and it is exciting.

Another aspect contributing to the excellent health of the field is its strong connection to empiricism. Network studies thrive on getting access to real, empirically-obtained graphs—we have seen this over and over again in discussion of the articles in this Section. From the Web graph with its tens of billions of nodes, to food webs with perhaps a hundred nodes, the science of networks is stimulated, challenged, and enriched by a steady influx of new data.

Finally, the study of complex networks is eminently practical. The path to direct application can be very short. Again we cite the Web graph and Google's PageRank algorithm as an example of this. The study of gene networks is somewhat farther from immediate application; but the possible benefits from a real understanding of cell and organism development can be enormous. The same holds for the problem of epidemic spreading. These examples are only picked out to illustrate the point; all of the articles, and topics represented by them, are not far removed from practical application.

In short: the field is exciting, vigorous, and interdisciplinary, and offers great practical benefits to society. The study of graphs and networks shows no signs of becoming moribund. Hence I will hazard a guess about the future: that the field will continue to grow and inspire excitement for many years to come. It is my hope that many readers of this Section will be infected by this excitement, and will choose to join in the fun.

Bibliography

Primary Literature

1. Adamic LA (1999) The Small World Web. In: Proc 3rd European Conf Research and Advanced Technology for Digital Libraries, ECDL, London, pp 443–452
2. Albert R, Jeong H, Barabasi AL (1999) Diameter of World-Wide Web. Nature 410:130–131
3. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
4. Badii R, Politi A (1997) Complexity: Hierarchical Structures and Scaling in Physics. Cambridge Nonlinear Science Series, vol 6. Cambridge University Press, Cambridge
5. Barabasi AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286:509–512
6. Bjelland J, Canright G, Engø-Monsen K, Remple VP (2008) Topographic Spreading Analysis of an Empirical Sex Workers' Network. In: Ganguly N, Mukherjee A, Deutsch A (eds) Dynamics on and of Complex Networks. Birkhauser, Basel,also in http://delis.upb.de/paper/DELIS-TR-0634.pdf
7. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the seventh international conference on World Wide Web, pp 107–117
8. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata S, Tomkins A, Wiener J (2000) Graph structure in the web. Comput Netw 33:309–320
9. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford
10. Erdős P (1947) Some remarks on the theory of graphs. Bull Amer Math Soc 53:292–294
11. Erdős P, Rényi A (1959) On random graphs, I. Publ Math Debrecen 6:290–297
12. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:8271–8276
13. Kaplan D, Glass L (1995) Understanding Nonlinear Dynamics. Springer, New York
14. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224:177–8
15. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467
16. Kephart JO, White SR (1991) Directed-graph epidemiological models of computer viruses. In: Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy, pp 343–359
17. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
18. Milgram S (1967) The Small World Problem. Psychol Today 2:60–67
19. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298:824–827
20. Newman MEJ, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66:35–101
21. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
22. Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: Bringing order to the web. Technical report. Stanford University, Stanford
23. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203
24. Watts DJ (1999) Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton Studies in Complexity. Princeton University Press, Princeton
25. Watts D, Strogatz S (1998) Collective dynamics of 'small world' networks. Nature 393:440–442

Books and Reviews

Albert R, Barabasi AL (2002) Statistical Mechanics of Complex Networks. Rev Mod Phys 74:47–97

Bornholdt S, Schuster HG (2003) Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH, Berlin

Caldarelli G, Vespignani A (2007) Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. Cambridge University Press, Cambridge

Chung F, Lu L (2006) Complex Graphs and Networks, CBMS Regional Conference. Series in Mathematics, vol 107. AMS, Providence

da F Costa L, Osvaldo N, Oliveira J, Travierso G, Rodrigues FA, Paulino R, Boas V, Antiqueira L, Viana MP, da Rocha LEC Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Working Paper: http://arxiv.org/abs/0711.3199

Kauffman SA (1993) The origins of order. Oxford University Press, Oxford

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Newman M, Barabasi A, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton Studies in Complexity. Princeton University Press, Princeton

Complex Networks, Visualization of

VLADIMIR BATAGELJ
University of Ljubljana, Ljubljana, Slovenia

Article Outline

Glossary

For basic notions on graphs and networks, see the articles by Wouter de Nooy: ► Social Network Analysis, Graph Theoretical Approaches to and by Vladimir

Batagelj: ► Social Network Analysis, Large-Scale in the Social Networks Section. For complementary information on graph drawing in social network analysis, see the article by Linton Freeman: ► Social Network Visualization, Methods of.

***k*-core** A set of vertices in a graph is a k-core if each vertex from the set has an internal (restricted to the set) degree of at least k and the set is maximal – no such vertex can be added to it.

Network A network consists of vertices linked by lines and additional data about vertices and/or lines. A network is large if it has at least some hundreds of vertices. Large networks can be stored in computer memory.

Partition A partition of a set is a family of its nonempty subsets such that each element of the set belongs to exactly one of the subsets. The subsets are also called classes or groups.

Spring embedder is another name for the energy minimization graph drawing method. The vertices are considered as particles with repulsive force between them, and lines as springs that attract or repel the vertices if they are too far or too close, respectively. The algorithm is a means of determining an embedding of vertices in two or three dimensional space that minimizes the 'energy' of the system.

Definition of the Subject

The earliest pictures containing graphs were magic figures, connections between different concepts (for example the Sephirot in Jewish Kabbalah), game boards (nine men's morris, pachisi, patolli, go, xiangqi, and others) road maps (for example Roman roads in Tabula Peutingeriana), and genealogical trees of important families [33].

The notion of the graph was introduced by Euler. In the eighteenth and nineteenth centuries, graphs were used mainly for solving recreational problems (Knight's tour, Eulerian and Hamiltonian problems, map coloring). At the end of the nineteenth century, some applications of graphs to real life problems appeared (electric circuits, Kirchhoff; molecular graphs, Kekulé). In the twentieth century, graph theory evolved into its own field of discrete mathematics with applications to transportation networks (road and railway systems, metro lines, bus lines), project diagrams, flowcharts of computer programs, electronic circuits, molecular graphs, etc.

In social science the use of graphs was introduced by Jacob Moreno around 1930 as a basis of his sociometric approach. In his book *Who shall survive?* [36], a relatively large network *Sociometric geography of community – map III* (435 individuals, 4350 lines) is presented. Linton Freeman wrote a detailed account of the development of social network analysis [20] and the visualization of social networks [19]. The networks studied in social network analysis until the 1990s were mostly small – some tens of vertices.

Introduction

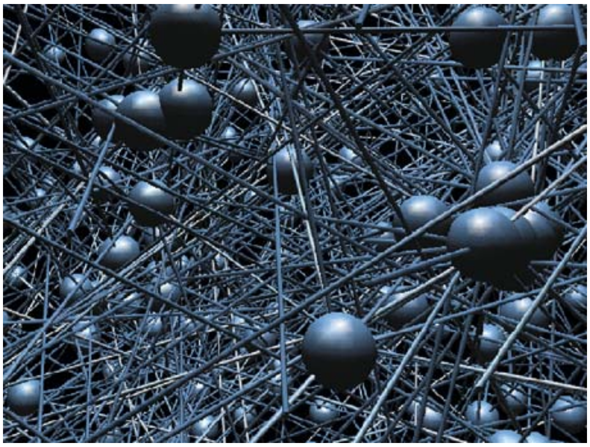
Through the 1980s, the development of information technology (IT) laid the groundwork for the emerging field of computer graphics. During this time, the first algorithms for graph drawing appeared:

- Trees: Wetherell and Shannon [45].
- Acyclic graphs: Sugiyama [42].
- Energy minimization methods (spring embedders) for general graphs: Eades [17], Kamada and Kawai [30], Fruchterman and Reingold [21].

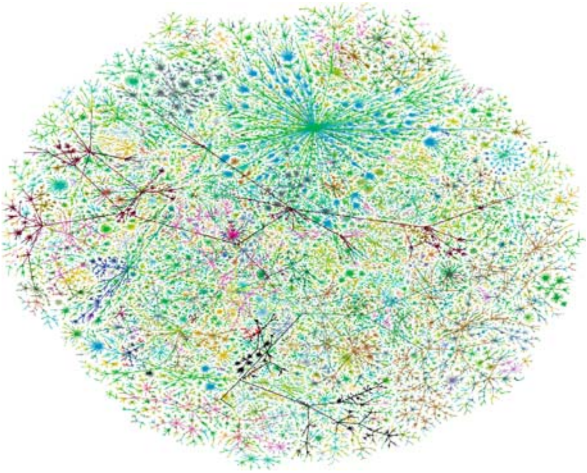
In energy minimization methods, vertices are considered as particles with repulsive force between them and lines as springs that attract or repel the vertices if they are too far or too close, respectively. The algorithms provide a means of determining an embedding of vertices in two or three dimensional space that minimizes the 'energy' of the system.

As early as 1963, William Tutte proposed an algorithm for drawing planar graphs [43] and Donald E. Knuth put forth an algorithm for drawing flowcharts [31].

A well known example of an early graph visualization was produced by Alden Klovdahl using his program View_Net – see Fig. 1. As nice a piece of art as it was, it held an important message: there are big problems with the visualization of dense, large graphs.



Complex Networks, Visualization of, Figure 1
Klovdahl: Social links in Canberra, Australia



Complex Networks, Visualization of, Figure 2
Network of traceroute paths for 29 June 1999

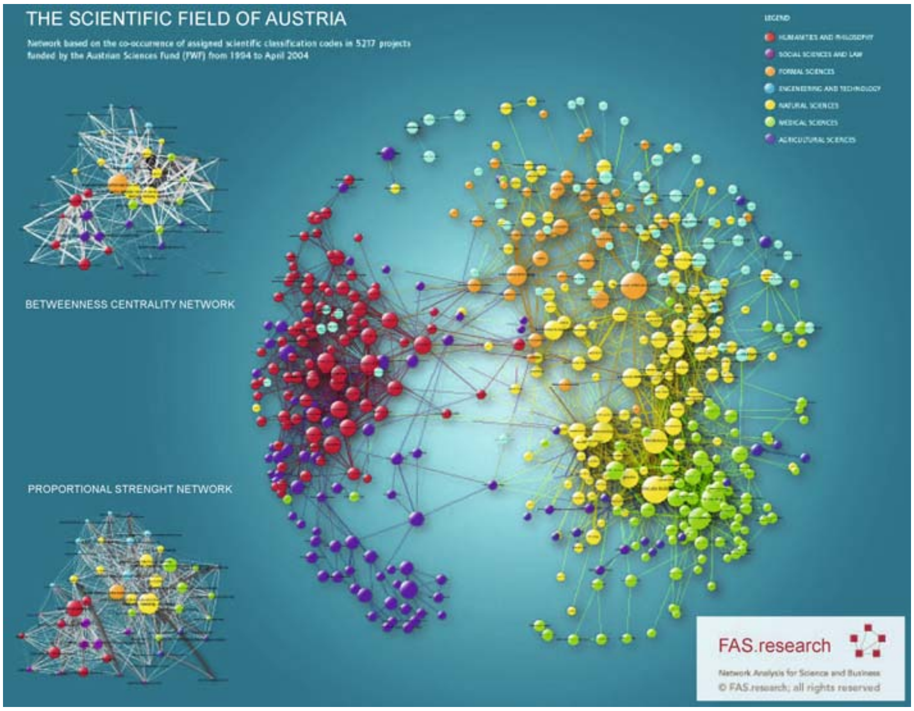
These developments led to the emergence of a new field: graph drawing. In 1992 a group of computer scientists and mathematicians (Giuseppe Di Battista, Peter Eades, Michael Kaufmann, Pierre Rosenstiehl, Kozo Sugiyama, Roberto Tamassia, Ioannis Tollis, and others) started the conference *International Symposium on Graph Drawing* which takes place each year. The proceedings of the conference are published in the *Lecture Notes in Computer Science* series by Springer [24]. To stimulate new approaches to graph drawing, a *graph drawing contest* accompanies each conference. Many papers on graph drawing are published in the *Journal of Graph Algorithms and Applications* [79].

Most of the efforts of the graph drawing community were spent on problems of drawing special types of graphs (trees, acyclic graphs, planar graphs) or using special styles (straight lines, orthogonal, grid-based, circular, hierarchical) and deriving bounds on required space (area) for selected types of drawing.

In the 1990s, further development of IT (GUI, multimedia, WWW) made large graph analysis a reality. For example, see the studies of large organic molecules (PDB [51]), the Internet (Caida [52]), and genealogies (White and Jorion [46], FamilySearch [60]). In chemistry, several tools for dynamic three-dimensional visualization and inspection of molecules were developed (Kinemage [48], Rasmol [87], MDL Chime [82]).

One of the earliest systems for large networks was SemNet (see Fairchild, Poltrock, Furnas [18]), used to explore knowledge bases represented as directed graphs.

In 1991 Tom Sawyer Software [91] was founded – the premier provider of high performance graph visualization,


THE SCIENTIFIC FIELD OF AUSTRIA
Network based on the co-occurrence of assigned scientific classification codes in 5217 projects funded by the Austrian Sciences Fund (FWF) from 1994 to April 2004
LEGEND
HUMANITIES AND PHILOSOPHY
SOCIAL SCIENCES AND LAW
FORMAL SCIENCES
ENGINEERING AND TECHNOLOGY
NATURAL SCIENCES
MEDICAL SCIENCES
AGRICULTURAL SCIENCES
BETWEENNESS CENTRALITY NETWORK
PROPORTIONAL STRENGHT NETWORK
FAS.research
Network Analysis for Science and Business
© FAS.research; all rights reserved

Complex Networks, Visualization of, Figure 3
FAS: The scientific field of Austria

layout, and analysis systems that enable the user to see and interpret complex information to make better decisions.

In 1993 at AT&T, the development of GraphViz tools for graph visualization began (dot, neato, dotty, tcldot, libgraph) [69]. Becker, Eick and Wilks developed a SeeXYZ family of network visualization programs [8].

In 1996 Vladimir Batagelj and Andrej Mrvar started the development of *Pajek* – a program for large network analysis and visualization [5].

In 1997 at La Sepienza, Rome, the development of GDToolkit [63] started as an extension of LEDA (Library of Efficient Data types and Algorithms) to provide implementations of several classical graph-drawing algorithms. The new version, GDT 4.0 (2007), produced in collaboration with University of Perugia, is LEDA independent.

Graham Wills, a principal investigator at Bell Labs (1992–2001), built the Nicheworks system for the visual analysis of very large weighted network graphs (up to a million vertices).

In the summer of 1998 Bill Cheswick and Hal Burch started work on the *Internet Mapping Project* at Bell Labs [53]. Its goal was to acquire and save Internet topological data over a long period of time. This data has been used in the study of routing problems and changes, distributed denial of service (DDoS) attacks, and graph theory. In the fall of 2000 Cheswick and Burch moved to a spin-off from Lucent/Bell Labs named Lumeta Corporation. Bill Cheswick is now back at AT&T Labs. Figure 2 shows a network obtained from traceroute paths for 29 June 1999 with nearly 100 000 vertices.

In the years 1997–2004, Martin Dodge maintained his *Cybergeography Research* web pages [58]. The results were published in the book *The Atlas of Cyberspace* [14]. A newer, very rich, site on information visualization is *Visual complexity* [95], where many interesting ideas on network visualizations can be found. These examples, and many others, can also be accessed from the Infovis *1100+ examples of information visualization* site [77]. An interesting collection of graph/network visualizations can be found also on the CDs of Gerhard Dirmoser. The collection also contains many artistic examples and other pictures not produced by computers.

In 1997 Harald Katzmair founded FAS research in Vienna, Austria [61], a company providing network analysis services. FAS emphasizes the importance of nice-looking final products (pictures) for customers by using graphical tools to enhance the visual quality of results obtained from network analysis tools. In Fig. 3, a network of Austrian research projects is presented. A similar company, Aguidel [49], was founded in France by Andrei Mogoutov, author of the program Réseau-Lu.

Every year (from 2002) at the INSNA Sunbelt conference [78], the *Viszards group* has a special session in which they present their solutions – analysis and visualizations of selected networks or types of networks. Most of the selected networks are large (KEDS, Internet Movie Data Base, Wikipedia, Web of Science).

Attempts

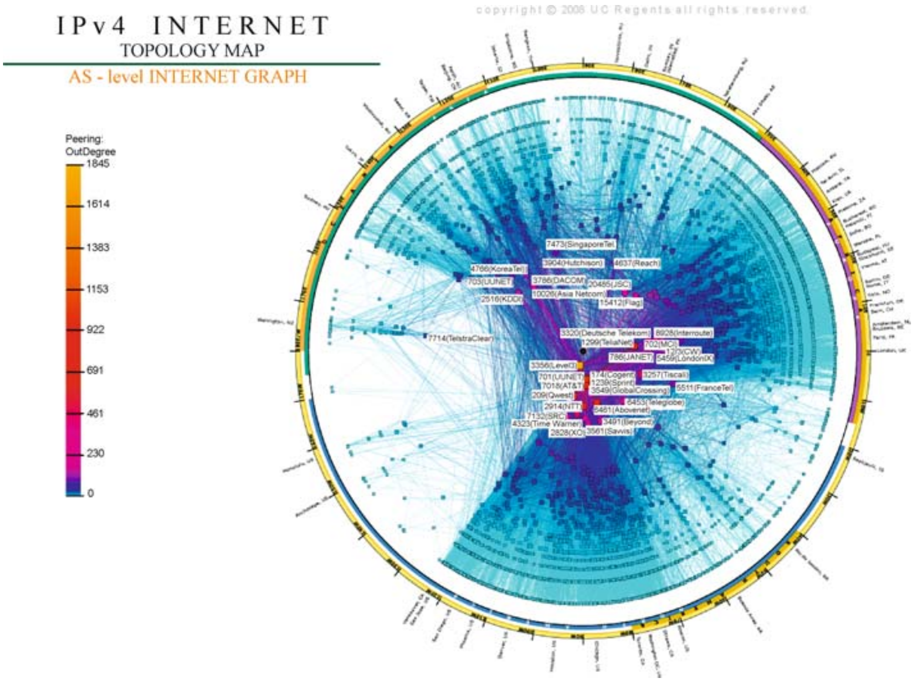
The new millennium has seen several attempts to develop programs for drawing large graphs and networks. Most of the following descriptions are taken verbatim from the programs' web pages.

Stephen Kobourov, with his collaborators from the University of Arizona, developed two graph drawing systems: GRIP (2000) [22,70] and Graphael (2003) [66]. GRIP – *Graph dRawing with Intelligent Placement* was designed for drawing large graphs and uses a multi-dimensional force-directed method together with fast energy function minimization. It employs a simple recursive coarsening scheme – rather than being placed at random, vertices are placed intelligently, several at a time, at locations close to their final positions.

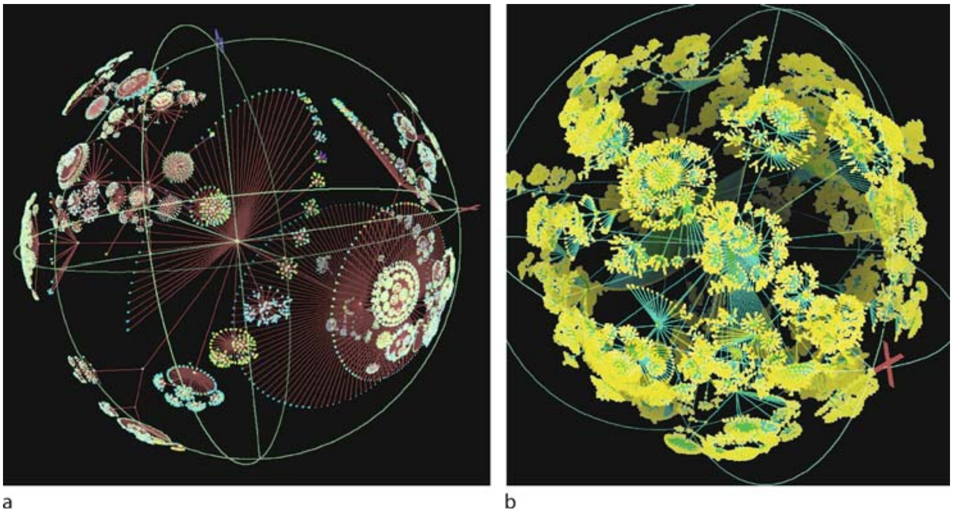
The Cooperative Association for Internet Data Analysis (CAIDA) [52], co-founded in 1998 by kc claffy, is an independent research group dedicated to investigating both the practical and theoretical aspects of the Internet to promote the engineering and maintenance of a robust, scalable, global Internet infrastructure. They have been focusing primarily on understanding how the Internet is evolving, and on developing a state-of-the-art infrastructure for data measurement that can be shared with the entire research community.

Figure 4 represents a macroscopic snapshot of the Internet for two weeks: 1–17 January 2008. The graph reflects 4 853 991 observed IPv4 addresses and 5 682 419 IP links. The network is aggregated into a topology of Autonomous Systems (ASes). The abstracted graph consists of 17 791 ASes (vertices) and 50 333 peering sessions (lines).

Walrus is a tool for interactively visualizing large directed graphs in three-dimensional space. It is best suited to visualizing moderately sized graphs (a few hundred thousand vertices) that are nearly trees. Walrus uses three-dimensional hyperbolic geometry to display graphs under a fisheye-like magnifier. By bringing different parts of a graph to the magnified central region, the user can examine every part of the graph in detail. Walrus was developed by Young Hyun at CAIDA based on research by Tamara Munzner. Figure 5 presents two examples of visualizations produced with Walrus.



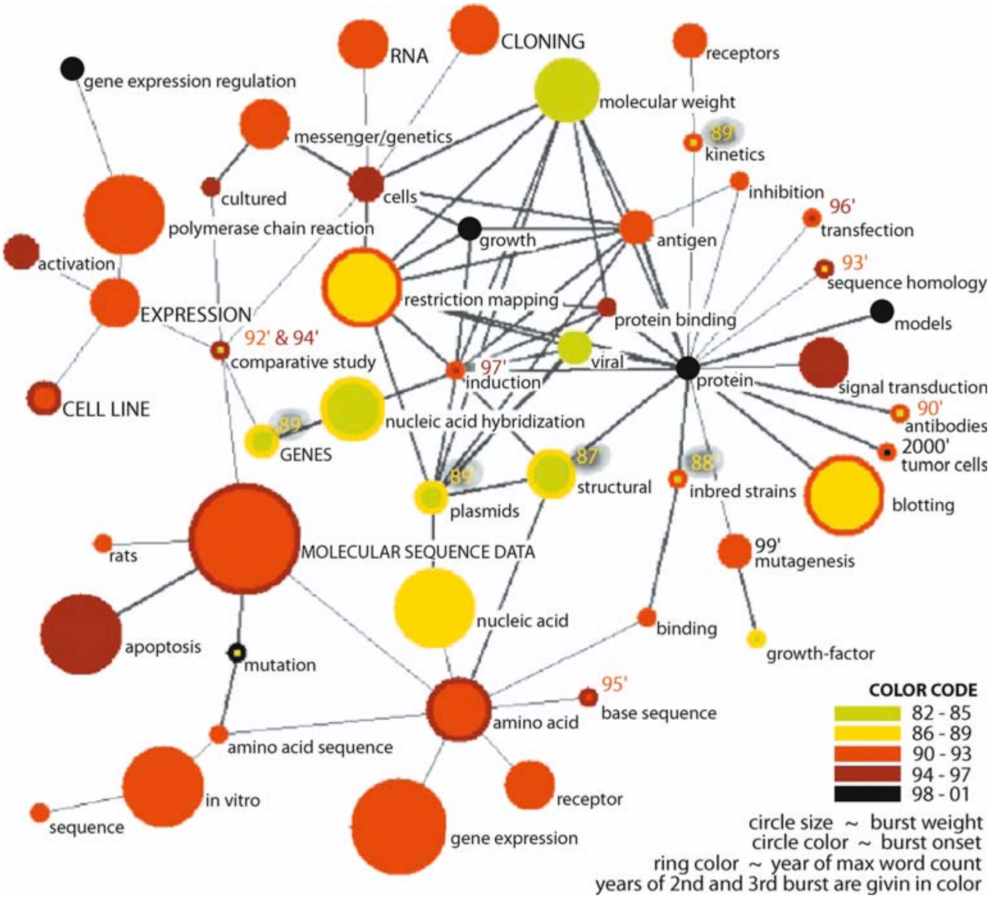
Complex Networks, Visualization of, Figure 4
CAIDA: AS core 2008



Complex Networks, Visualization of, Figure 5
Walrus

Some promising algorithms for drawing large graphs have been proposed by Ulrik Brandes, Tim Dwyer, Emden Gansner, Stefan Hachul, David Harel, Michael Jünger, Yehuda Koren, Andreas Noack, Stephen North, Christian Pich, and Chris Walshaw [11,16,23,25,26,32,39,44]. They are based either on a multilevel energy minimization approach or on an algebraic or spectral approach that reduces to some application of eigenvectors.

The multilevel approach speeds-up the algorithms. *Multilevel algorithms* are based on two phases: a *coarsening phase*, in which a sequence of coarse graphs with decreasing sizes is computed, and a *refinement phase*, in which successively finer drawings of graphs are computed, using the drawings of the next coarser graphs and a variant of a suitable force-directed single-level algorithm [25]. The fastest algorithms combine the multilevel approach



Complex Networks, Visualization of, Figure 6
Katy Börner: Text analysis

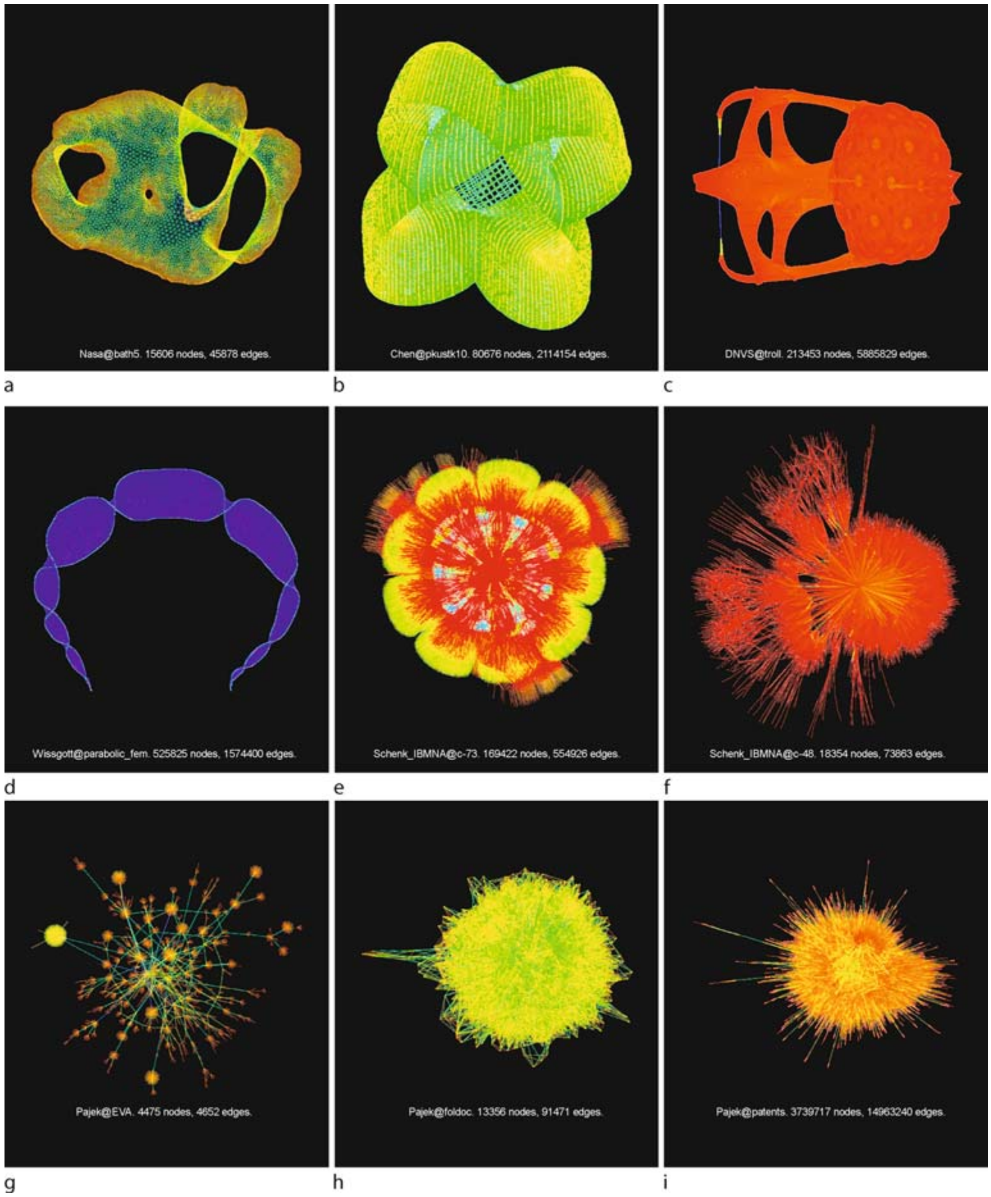
with fast approximation of long range repulsive force using nested data structures, such as quadtree or kd-tree.

Katy Börner from Indiana University, with her collaborators, produced several visualizations of scientometric networks such as Backbone of Science [9] and Wikipedia [73]. They use different visual cues to produce information-rich visualizations – see Fig. 6. She also commissioned the *Map of Science* based on data (800 000 published papers) from Thomson ISI and produced by Kevin Boyack, Richard Klavans and Bradford Paley [9].

Yifan Hu from AT&T Labs Information Visualization Group developed a multilevel graph drawing algorithm for visualization of large graphs [29]. The algorithm was first implemented in 2004 in Mathematica and released in 2005. For demonstration he applied it to the *University of Florida Sparse Matrix collection* [56] that contains over 1500 square matrices. The results are available in the *Gallery of Large Graphs* [74]. The largest graph (vanHeukelum/cage15) has 5 154 859 vertices and 47 022 346 edges. In Fig. 7, selected pictures from the gallery are presented.

From the examples that we have given, we can see that, in some cases, graph drawing algorithms can reveal symmetries in a given graph and also a 'structure' ((sub)trees, clusters, planarity, etc.). Challenges remain in devising ways to represent graphs with dense parts.

For dense parts, a better approach is to display them using matrix representation. This representation was used in 1999 by Vladimir Batagelj, Andrej Mrvar and Matjaž Zaveršnik in their partitioning approach to visualization of large graphs [7] and is a basis of systems such as *Matrix Zoom* by James Abello and Frank van Ham, 2004 [1,2], and *MatrixExplorer* by Nathalie Henry and Jean-Daniel Fekete, 2006 [27]. A matrix representation is determined by an ordering of vertices. Several algorithms exist that can produce such orderings. A comparative study of them was published by Chris Mueller [37,84]. In Fig. 8 three orderings of the same matrix are presented. Most ordering al-


Nasa@bath5. 15606 nodes, 45878 edges.
a
Chen@pkustk10. 80676 nodes, 2114154 edges.
b
DNVS@troll. 213453 nodes, 5885829 edges.
c
Wissgott@parabolic_fem. 525825 nodes, 1574400 edges.
d
Schenk_IBMNA@c-73. 169422 nodes, 554926 edges.
e
Schenk_IBMNA@c-48. 18354 nodes, 73863 edges.
f
Pajek@EVA. 4475 nodes, 4652 edges.
g
Pajek@foldoc. 13356 nodes, 91471 edges.
h
Pajek@patents. 3739717 nodes, 14963240 edges.
i

Complex Networks, Visualization of, Figure 7
Examples from the Gallery of Large Graphs

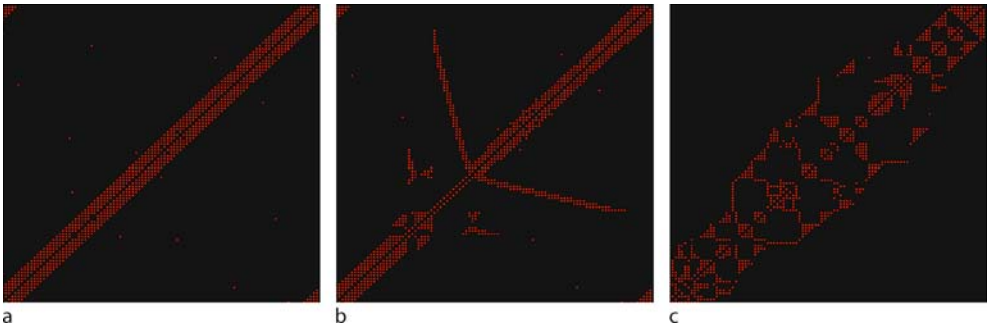
gorithms were originally designed for applications in numerical, rather than data, analysis. The orderings can also be determined using clustering or blockmodeling methods [15].

An important type of networks are *temporal networks*, where the presence of vertices and lines changes through time. Visualization of such networks requires special approaches (Sonia [88], SVGanim [90], TecFlow [75]). An interesting approach to visualization of temporal networks was developed by Ulrik Brandes and his group [12].

Perspectives

In this section, we present a collection of ideas on how to approach visualization of large networks. These ideas are only partially implemented in different visualization solutions.

While the technical problems of graph drawing strive for a single 'best' picture, network analysis is also a part of data analysis. Its goal is to gain insight not only into the structure and characteristics of a given network, but also



Complex Networks, Visualization of, Figure 8
Matrix representations

Complex Networks, Visualization of, Figure 9
Big picture, V. Batagelj, AE'04

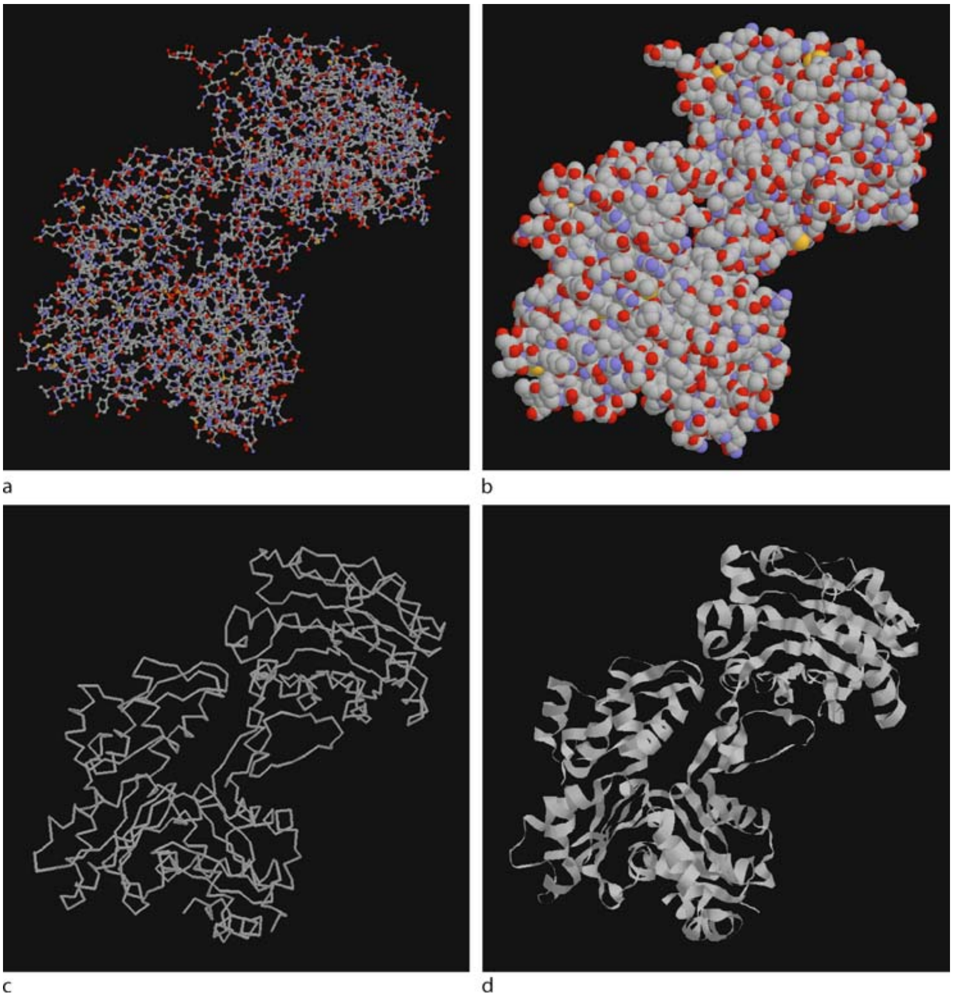
into how this structure influences processes going on over the network. We usually need several pictures to present the obtained results.

Small graphs can be presented in their totality and in detail within a single view. In a comprehensive view of large graphs, details become lost – conversely a detailed view can encompass only a part of a large graph.

The literature on graph drawing is dominated by the 'sheet of paper' paradigm – the solutions and techniques are mainly based on the assumption that the final result is a static picture on a sheet of paper. In this model, to present a large data set we need a large 'sheet of paper' – but this has a limit. Figure 9 presents a visualization of a symmetrized subnetwork of 5952 words and 18 008 associations from the *Edinburgh Associative Thesaurus* [59] prepared by Vladimir Batagelj on a 3 m × 5 m 'sheet of paper' for Ars Electronica, Linz 2004 *Language of networks* exhibition.

The main tool for dealing with large objects is abstraction. In graphs, abstraction is usually realized using a hierarchy of partitions. By shrinking selected classes of a partition we obtain a smaller reduced graph. The main operations related to abstraction are:

- Cut-out: Display of only selected parts (classes of partition) of a graph;



Complex Networks, Visualization of, Figure 10
Glasses: Rasmol displays – BallStick, SpaceFill, Backbone, Ribbons

- Context: Display details of selected parts (classes) of a graph and display the rest of the graph in some reduced form;
- Model: Display the reduced graph with respect to a given partition;
- Hierarchy: Display the tree representing the nesting of graph partitions.

In larger, denser networks there is often too much information to be presented at once. A possible answer is an interactive layout on a computer screen where the user controls what (s)he wants to see.

The computer screen is a medium which offers many new possibilities: parallel views (global and local); brushing and linking; zooming and panning; temporary elements (additional information about the selected elements, labels, legends, markers, etc.); highlighted selections; and others. These features can and should be maximally leveraged to support data analytic tasks; or repeating Shneiderman's mantra: overview first, zoom and filter, then details on-demand (extended with: Relate, history and extract) [40].

When interactively inspecting very large graphs, a serious problem appears: how does one avoid the "lost within the forest" effect? There are several solutions that can help the user maintain orientation:

- Restart option: Returns the user to the starting point;
- Introduction of additional orientation elements: Allows elements to be switched on and off.
- Multiview: Presents at least two views (windows):

- Map view: Shows an overall global view which contains the current position and allows 'long' moves (jumps). For very large graphs, a map view can be combined with zooming or fish-eye views.
- Local view: Displays a selected portion of the graph.

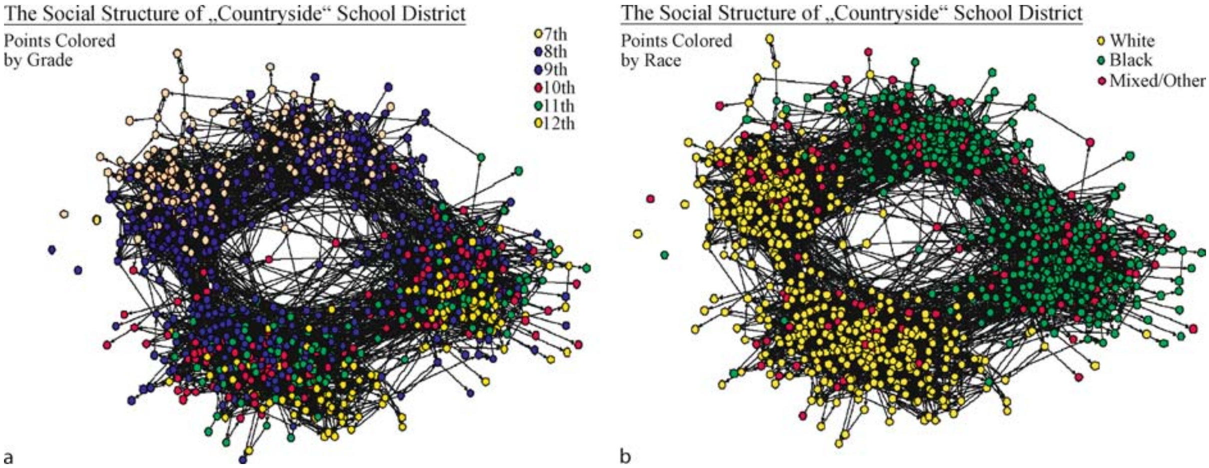
Additional support can be achieved by implementing trace, backtrack, and replay mechanisms and guided tours.

An interactive dynamic visualization of a graph on the computer screen need not be displayed in its totality. Inspecting a visualization, the user can select which parts and elements will be displayed and in what way. See, for example, TouchGraph [92].

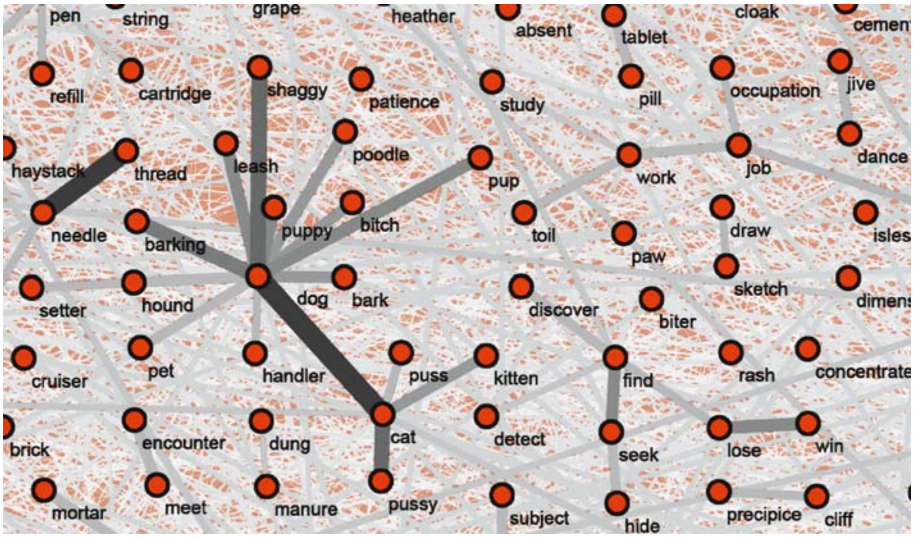
Closely related to the multiview concept are the associated concepts of glasses, lenses and zooming. Glasses affect the entire window, while lenses affect only selected region or elements.

By selecting different glasses, we can obtain different views on the same data supporting different visualization aims. For example, in Fig. 10 four different glasses (ball and stick, space-fill, backbone, ribbons) were applied in the program Rasmol to the molecule 1atn.pdb (deoxyribonuclease I complex with actin).

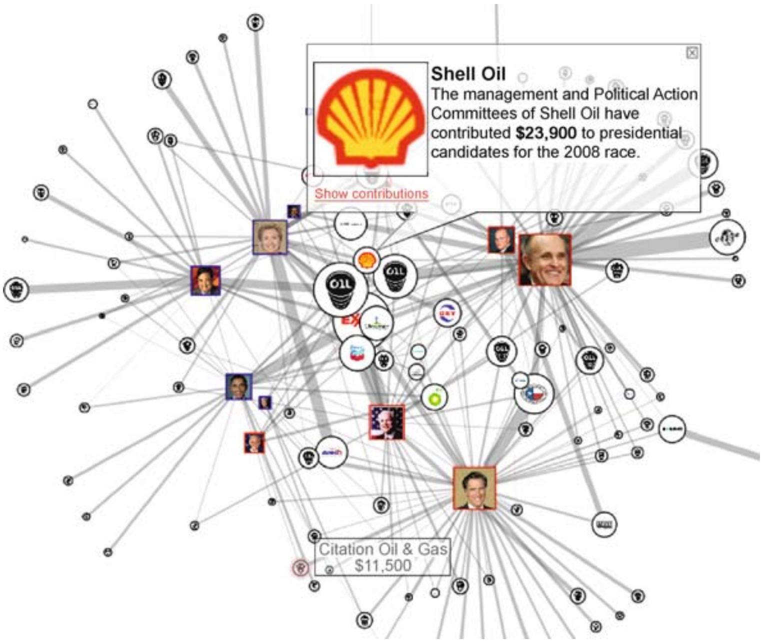
Another example of glasses is presented in Fig. 11. The two pictures were produced by James Moody [35]. The graph pictured was obtained by applying spring em-


The Social Structure of „Countryside" School District
Points Colored by Grade
○7th
●8th
●9th
●10th
●11th
○12th
a
The Social Structure of „Countryside" School District
Points Colored by Race
○ White
● Black
● Mixed/Other
b

Complex Networks, Visualization of, Figure 11
Glasses: Display of properties – school

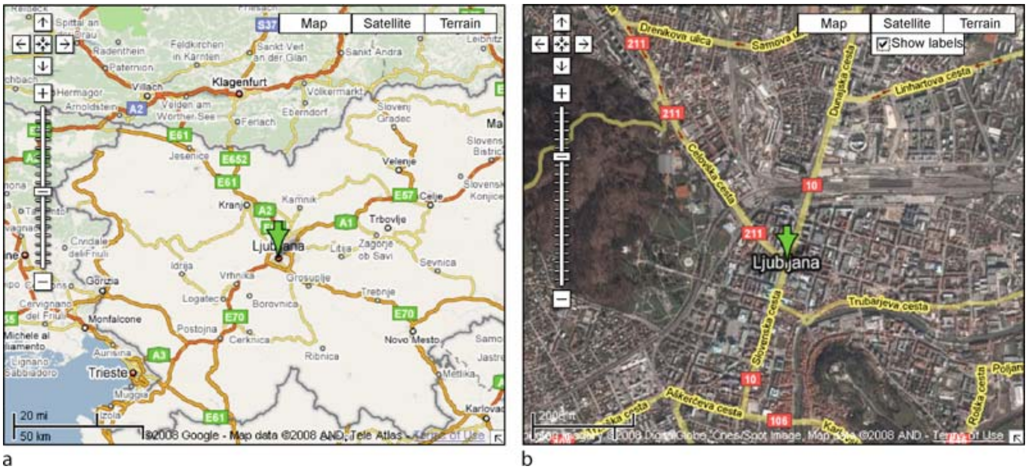


Complex Networks, Visualization of, Figure 12
Part of the big picture



Shell Oil
The management and Political Action Committees of Shell Oil have contributed **$23,900** to presidential candidates for the 2008 race.
Show contributions

Complex Networks, Visualization of, Figure 13
Lenses: Temporary info about the selected vertex

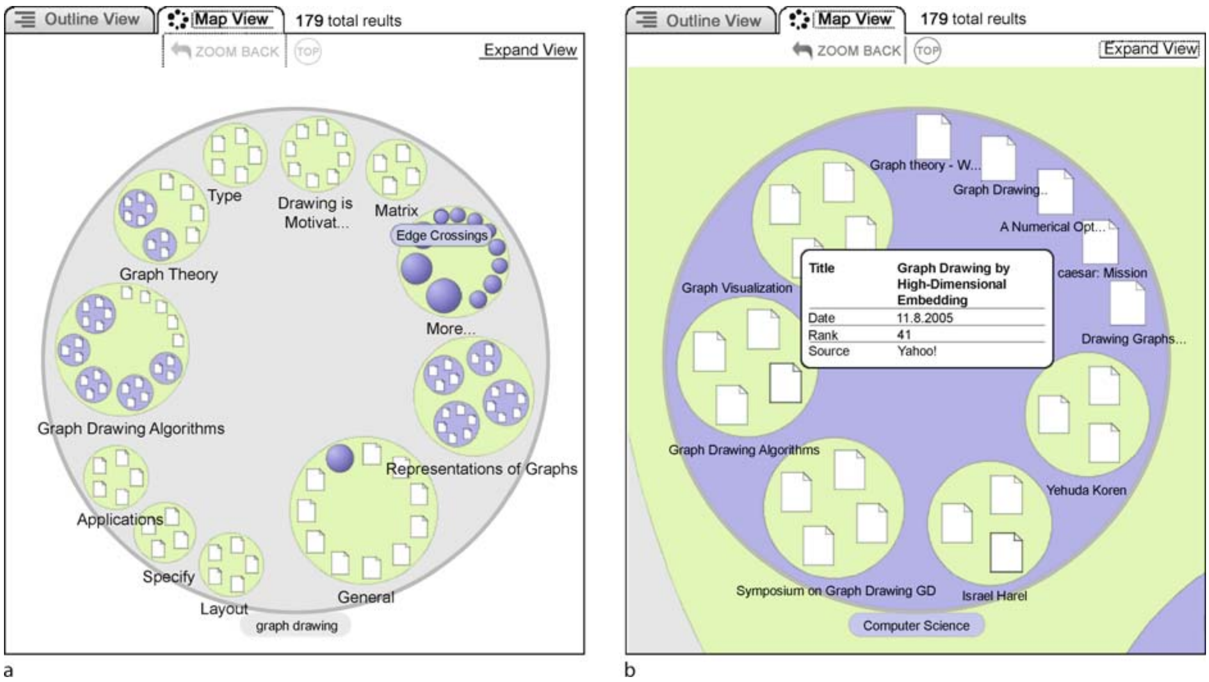


Complex Networks, Visualization of, Figure 14
Zoom, glasses, lenses, navigation: Google Maps

bedders. It represents the friendships among students in a school. The glasses are the coloring of its vertices by different partitions: an age partition (left picture) and a race partition (right picture). This gives us an explanation of the four groups in the resulting graph picture, characterized by younger/older and white/black students.

Figure 12 shows a part of the big picture presented in Fig. 9. The glasses in this case are based on ordering the edges in increasing order of their values and drawing them in this order – stronger edges cover the weaker. The picture emphasizes the strongest substructures; the remaining elements form a background.

There are many kinds of glasses in representation of graphs, for example, fish-eye views, matrix representation, using application field conventions (genealogies, molecules, electric circuits, SBGN), displaying vertices only, selecting the type of labels (long/short name, value), displaying only the important vertices and/or lines, size of vertices determined by core number or "betweenness".



Title Graph Drawing by High-Dimensional Embedding
Date 11.8.2005
Rank 41
Source Yahoo!

Complex Networks, Visualization of, Figure 15
Zoom, glasses, lenses, navigation: Grokker

An example of lens is presented in Fig. 13 – contributions of companies to various presidential candidates from *Follow the Oil Money* by Greg Michalec and Skye Bender-deMoll [62]. When a vertex is selected, information about that vertex is displayed. Another possible use of a lens would be to temporarily enhance the display of neighbors of a selected vertex [94] or to display their labels. The "shaking" option used in *Pajek* to visually identify all vertices from a selected cluster is also a kind of lens; so are the matrix representations of selected clusters in NodeTrix [72].

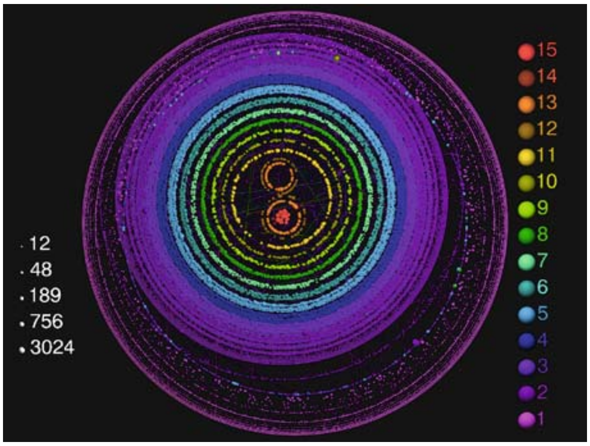
Additional enhancement of a presentation can be achieved by the use of support elements such as labels, grids, legends, and various forms of help facilities.

An important concept connected with zooming is the level of detail, or LOD – subobjects are displayed differently depending on the zoom depth.

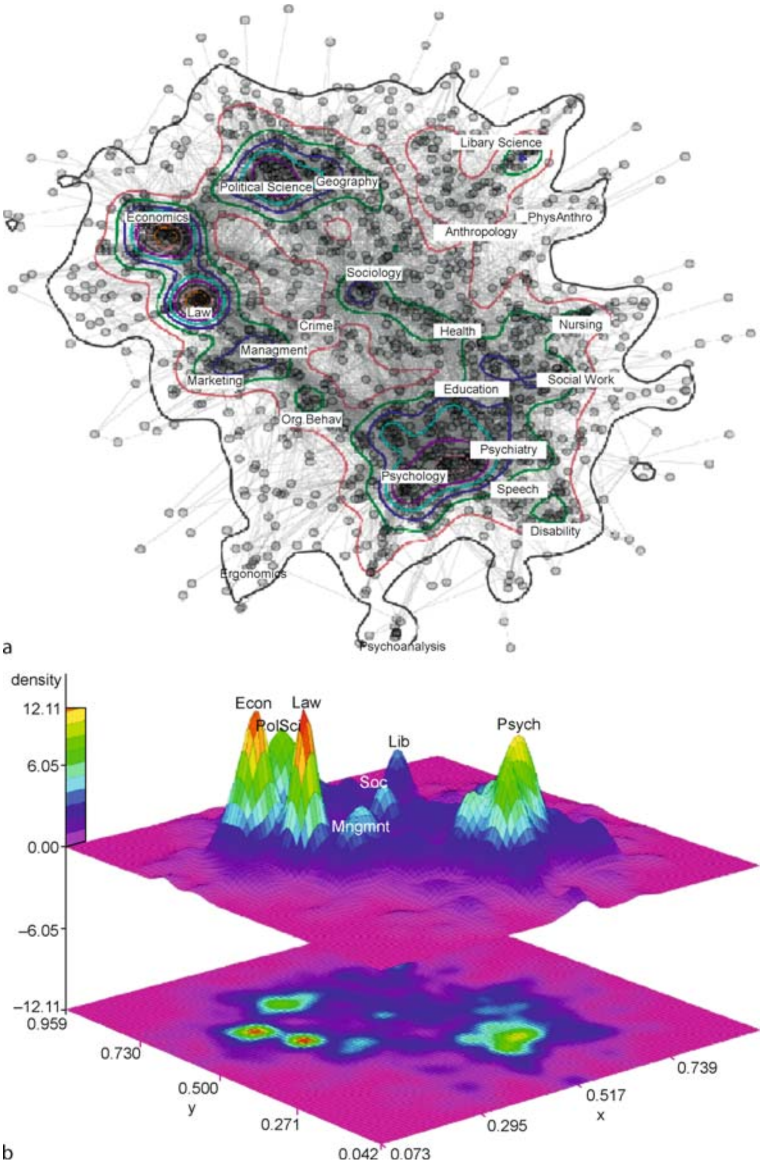
A nice example of a combination of these techniques is the Google Maps service [65] – see Fig. 14. It combines zooming, glasses (Map, Satellite, Terrain), navigation (left, right, up, down) and lenses (info about points). The maps at different zoom levels provide information at different levels of detail and in different forms.

A similar approach could be used for inspection of a large graph or network by examining selected hierarchical clusterings of its vertices. To produce higher level 'maps,' different methods can be used: *k*-core representation [4], density contours [83], generalized blockmodeling [15], clustering [71] (Fig. 15), preserving only important vertices and lines, etc. In visualizing 'maps,' new graphical elements (many of them still to be invented) can be used (see [13,80], p. 223 in [15]) to preserve or to indicate information about structures at lower levels.

The *k*-core representation [4,81] is based on *k*-core decomposition of a network [6,7] and was developed by Alessandro Vespignani and his collaborators [54]. Fig-



Complex Networks, Visualization of, Figure 16
***k*-core structure of a portion of the web at the .fr domain**



Complex Networks, Visualization of, Figure 17
Density structure

ure 16 shows a portion of the web at the .fr domain with 1 million pages. Each node represents a web page and each edge is a hyperlink between two pages.

Density contours were introduced by James Moody in 2006. First, a spring embedder layout of a (valued) network is determined. Next, vertices and lines are removed and replaced by density contours. Figure 17 shows this process applied to the case of a social science co-citation network. The left side shows the network layout and the right bottom part presents the corresponding density contours.

The basic steps in graph/network visualization are:

$$\text{graph/network} \to \boxed{\text{analysis}} \to \text{layouts}$$
$$\to \boxed{\text{viewer}} \to \text{pictures}\,.$$

Development of different tools can be based on this scheme, depending on the kind of users (simple, advanced) and the tasks they address (reporting, learning, monitoring, exploration, analysis). In some cases, a simple viewer will be sufficient (for example SVG viewer, X3D viewer, or a special graph layout viewer), in others a complete network analysis system is needed (such as Geomi [3,64], ILOG [76], Pajek [50], Tulip [93], yFiles [96]).

To visualize a network, layouts are obtained by augmenting network data with results of analysis and users' decisions. In *Pajek*'s input format, there are several layout elements from *Pajek*'s predecessors (see *Pajek*'s manual, pp. 69–73 in [50]). As in typesetting

$$\text{text} \ + \ \text{formatting} \ = \ \text{formatted text}$$

so in network visualization

$$\text{network} \ + \ \text{layout} \ = \ \text{picture}\,.$$

It would be useful to define a common layout format (an extension of GraphML [68]?) so that independent viewer modules can be developed and combined with different layout algorithms. Some useful ideas can be found in the nViZn ("envision") system [89]. To specify layouts we can borrow from typesetting the notion of *style*.

Bibliography

Primary Literature

1. Abello J, van Ham F (2004) Matrix zoom: A visual interface to semi-external graphs. IEEE Symposium on Information Visualization, October 10–12 2004, Austin, Texas, USA, pp 183–190
2. Abello J, van Ham F, Krishnan N (2006) ASK-GraphView: A large scale graph visualization system. IEEE Trans Vis Comput Graph 12(5):669–676
3. Ahmed A, Dwyer T, Forster M, Fu X, Ho J, Hong S, Koschützki D, Murray C, Nikolov N, Taib R, Tarassov A, Xu K (2006) GEOMI: GEOmetry for maximum insight. In: Healy P, Eades P (eds) Proc 13th Int Symp Graph Drawing (GD2005). Lecture Notes in Computer Science, vol 3843. Springer, Berlin, pp 468–479
4. Alvarez-Hamelin JI, DallAsta L, Barrat A, Vespignani A (2005) Large scale networks fingerprinting and visualization using the k-core decomposition. In: Advances in neural information processing systems 18, Neural Information Processing Systems, NIPS 2005, December 5–8, 2005, Vancouver, British Columbia, Canada
5. Batagelj V, Mrvar A (2003) Pajek – analysis and visualization of large networks. In: Jünger M, Mutzel P (eds) Graph drawing software. Springer, Berlin, pp 77–103
6. Batagelj V, Zaveršnik M (2002) Generalized cores. arxiv cs.DS/0202039
7. Batagelj V, Mrvar A, Zaveršnik M (1999) Partitioning approach to visualization of large graphs. In: Kratochvíl J (ed) Lecture notes in computer science, vol 1731. Springer, Berlin, pp 90–97
8. Becker RA, Eick SG, Wilks AR (1995) Visualizing network data. IEEE Trans Vis Comput Graph 1(1):16–28
9. Boyack KW, Klavans R, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3):351–374
10. Boyack KW, Klavans R, Paley WB (2006) Map of science. Nature 444:985
11. Brandes U, Pich C (2007) Eigensolver methods for progressive multidimensional scaling of large data. In: Proc 14th Intl Symp Graph Drawing (GD '06). Lecture notes in computer science, vol 4372. Springer, Berlin, pp 42–53
12. Brandes U, Fleischer D, Lerner J (2006) Summarizing dynamic bipolar conflict structures. IEEE Trans Vis Comput Graph (special issue on Visual Analytics) 12(6):1486–1499
13. Dickerson M, Eppstein D, Goodrich MT, Meng J (2005) Confluent drawings: Visualizing non-planar diagrams in a planar way. J Graph Algorithms Appl (special issue for GD'03) 9(1):31–52
14. Dodge M, Kitchin R (2001) The atlas of cyberspace. Pearson Education, Addison Wesley, New York
15. Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodeling. Cambridge University Press, Cambridge
16. Dwyer T, Koren Y (2005) DIG-COLA: Directed graph layout through constrained energy minimization. INFOVIS 2005:9
17. Eades P (1984) A heuristic for graph drawing. Congressus Numerantium 42:149–160
18. Fairchild KM, Poltrock SE, Furnas GW (1988) SemNet: Three-dimensional representations of large knowledge bases. In: Guindon R (ed) Cognitive science and its applications for human-computer interaction. Lawrence Erlbaum, Hillsdale, pp 201–233
19. Freeman LC (2000) Visualizing social networks. J Soc Struct 1(1). http://www.cmu.edu/joss/content/articles/volume1/Freeman/
20. Freeman LC (2004) The development of social network analysis: A study in the sociology of science. Empirical, Vancouver
21. Fruchterman T, Reingold E (1991) Graph drawing by force directed placement. Softw Pract Exp 21(11):1129–1164
22. Gajer P, Kobourov S (2001) GRIP: Graph drawing with intelligent placement. Graph Drawing 2000 LNCS, vol 1984:222–228
23. Gansner ER, Koren Y, North SC (2005) Topological fisheye views for visualizing large graphs. IEEE Trans Vis Comput Graph 11(4):457–468
24. Graph Drawing. Lecture Notes in Computer Science, vol 894 (1994), 1027 (1995), 1190 (1996), 1353 (1997), 1547 (1998), 1731 (1999), 1984 (2000), 2265 (2001), 2528 (2002), 2912 (2003), 3383 (2004), 3843 (2005), 4372 (2006), 4875 (2007). Springer, Berlin
25. Hachul S, Jünger M (2007) Large-graph layout algorithms at work: An experimental study. JGAA 11(2):345–369
26. Harel D, Koren Y (2004) Graph drawing by high-dimensional embedding. J Graph Algorithms Appl 8(2):195–214
27. Henry N, Fekete J-D (2006) MatrixExplorer: A dual-representation system to explore social networks. IEEE Trans Vis Comput Graph 12(5):677–684
28. Herman I, Melancon G, Marshall MS (2000) Graph visualization and navigation in information visualization: A survey. IEEE Trans Vis Comput Graph 6(1):24–43
29. Hu YF (2005) Efficient and high quality force-directed graph drawing. Math J 10:37–71
30. Kamada T, Kawai S (1988) An algorithm for drawing general undirected graphs. Inf Proc Lett 31:7–15
31. Knuth DE (1963) Computer-drawn flowcharts. Commun ACM 6(9):555–563
32. Koren Y (2003) On spectral graph drawing. COCOON 2003: 496–508
33. Kruja E, Marks J, Blair A, Waters R (2001) A short note on the history of graph drawing. In: Proc Graph Drawing 2001. Lecture notes in computer science, vol 2265. Springer, Berlin, pp 272–286
34. Lamping J, Rao R, Pirolli P (1995) A focus+context technique

based on hyperbolic geometry for visualizing large hierarchies. CHI 95:401–408

35. Moody J (2001) Race, school integration, and friendship segregation in America. Am J Soc 107(3):679–716
36. Moreno JL (1953) Who shall survive? Beacon, New York
37. Mueller C, Martin B, Lumsdaine A (2007) A comparison of vertex ordering algorithms for large graph visualization. APVIS 2007, pp 141–148
38. Munzner T (1997) H3: Laying out large directed graphs in 3D hyperbolic space. In: Proceedings of the 1997 IEEE Symposium on Information Visualization, 20–21 October 1997, Phoenix, AZ, pp 2–10
39. Noack A (2007) Energy models for graph clustering. J Graph Algorithms Appl 11(2):453–480
40. Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualization. In: IEEE Conference on Visual Languages (VL'96). IEEE CS Press, Boulder
41. Shneiderman B, Aris A (2006) Network visualization by semantic substrates. IEEE Trans Vis Comput Graph 12(5):733–740
42. Sugiyama K, Tagawa S, Toda M (1981) Methods for visual understanding of hierarchical systems. IEEE Trans Syst, Man, Cybern 11(2):109–125
43. Tutte WT (1963) How to draw a graph. Proc London Math Soc s3-13(1):743–767
44. Walshaw C (2003) A multilevel algorithm for force-directed graph-drawing. J Graph Algorithms Appl 7(3):253–285
45. Wetherell C, Shannon A (1979) Tidy drawing of trees. IEEE Trans Softw Engin 5:514–520
46. White DR, Jorion P (1992) Representing and computing kinship: A new approach. Curr Anthr 33(4):454–463
47. Wills GJ (1999) NicheWorks-interactive visualization of very large graphs. J Comput Graph Stat 8(2):190–212

Web Resources

48. 3D Macromolecule analysis and Kinemage home page: http://kinemage.biochem.duke.edu/. Accessed March 2008
49. Aguidel: http://www.aguidel.com/en/. Accessed March 2008
50. Batagelj V, Mrvar A (1996) Pajek – program for analysis and visualization of large network: http://pajek.imfm.si. Accessed March 2008. Data sets: http://vlado.fmf.uni-lj.si/pub/networks/data/. Accessed March 2008
51. Brookhaven Protein Data Bank: http://www.rcsb.org/pdb/. Accessed March 2008
52. Caida: http://www.caida.org/home/. Accessed March 2008. Walrus gallery: http://www.caida.org/tools/visualization/walrus/gallery1/. Accessed March 2008
53. Cheswick B: Internet mapping project – map gallery: http://www.cheswick.com/ches/map/gallery/. Accessed March 2008
54. Complex Networks Collaboratory: http://cxnets.googlepages.com/. Accessed March 2008
55. Cruz I, Tamassia R (1994) Tutorial on graph drawing. http://graphdrawing.org/literature/gd-constraints.pdf. Accessed March 2008
56. Davis T: University of Florida Sparse Matrix Collection: http://www.cise.ufl.edu/research/sparse/matrices. Accessed March 2008
57. Di Battista G, Eades P, Tamassia R, Tollis IG (1994) Algorithms for drawing graphs: An annotated bibliography. Comput Geom: Theory Appl 4:235–282. http://graphdrawing.org/literature/gdbiblio.pdf. Accessed March 2008
58. Dodge M: Cyber-Geography Research: http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/. Accessed March 2008
59. Edinburgh Associative Thesaurus (EAT): http://www.eat.rl.ac.uk/. Accessed March 2008
60. FamilySearch: http://www.familysearch.org/. Accessed March 2008
61. FASresearch, Vienna, Austria: http://www.fas.at/. Accessed March 2008
62. Follow the Oil Money: http://oilmoney.priceofoil.org/. Accessed March 2008
63. GDToolkit – Graph Drawing Toolkit: http://www.dia.uniroma3.it/~gdt/gdt4/index.php. Accessed March 2008
64. GEOMI (Geometry for Maximum Insight): http://www.cs.usyd.edu.au/~visual/valacon/geomi/. Accessed March 2008
65. Google Maps: http://maps.google.com/. Accessed March 2008
66. Graphael: http://graphael.cs.arizona.edu/. Accessed March 2008
67. Graphdrawing home page: http://graphdrawing.org/. Accessed March 2008
68. GraphML File Format: http://graphml.graphdrawing.org/. Accessed March 2008
69. Graphviz: http://graphviz.org/. Accessed March 2008
70. GRIP: http://www.cs.arizona.edu/~kobourov/GRIP/. Accessed March 2008
71. Grokker – Enterprise Search Management and Content Integration: http://www.grokker.com/. Accessed March 2008
72. Henry N, Fekete J-D, Mcguffin M (2007) NodeTrix: Hybrid representation for analyzing social networks: https://hal.inria.fr/inria-00144496. Accessed March 2008
73. Herr BW, Holloway T, Börner K (2007) Emergent mosaic of wikipedian activity: http://www.scimaps.org/dev/big_thumb.php?map_id=158. Accessed March 2008
74. Hu YF: Gallery of Large Graphs: http://www.research.att.com/~yifanhu/GALLERY/GRAPHS/index1.html. Accessed March 2008
75. iCKN: TeCFlow – a temporal communication flow visualizer for social network analysis: http://www.ickn.org/. Accessed March 2008
76. ILOG Diagrams: http://www.ilog.com/. Accessed March 2008
77. Infovis – 1100+ examples of information visualization: http://www.infovis.info/index.php?cmd=search&words=graph&mode=normal. Accessed March 2008
78. INSNA – International Network for Social Network Analysis: http://www.insna.org/. Accessed March 2008
79. Journal of Graph Algorithms and Applications: http://jgaa.info/. Accessed March 2008
80. KartOO visual meta search engine: http://www.kartoo.com/. Accessed March 2008
81. LaNet-vi – Large Network visualization tool: http://xavier.informatics.indiana.edu/lanet-vi/. Accessed March 2008
82. MDL Chime: http://www.mdli.com/. Accessed March 2008
83. Moody J (2007) The network structure of sociological production II: http://www.soc.duke.edu/~jmoody77/presentations/soc_Struc_II.ppt. Accessed March 2008
84. Mueller C: Matrix visualizations: http://www.osl.iu.edu/~chemuell/data/ordering/sparse.html. Accessed March 2008
85. OLIVE, On-line library of information visualization environments: http://otal.umd.edu/Olive/. Accessed March 2008

86. Pad++: Zoomable user interfaces: Portal filtering and 'magic lenses': http://www.cs.umd.edu/projects/hcil/pad++/tour/lenses.html. Accessed March 2008
87. RasMol Home Page: http://www.umass.edu/microbio/rasmol/index2.htm. Accessed March 2008
88. Sonia – Social Network Image Animator: http://www.stanford.edu/group/sonia/. Accessed March 2008
89. SPSS nViZn: http://www.spss.com/research/wilkinson/nViZn/nvizn.html. Accessed March 2008
90. SVGanim: http://vlado.fmf.uni-lj.si/pub/networks/pajek/SVGanim. Accessed March 2008
91. Tom Sawyer Software: http://www.tomsawyer.com/home/index.php. Accessed March 2008
92. TouchGraph: http://www.touchgraph.com/. Accessed March 2008
93. Tulip: http://www.labri.fr/perso/auber/projects/tulip/. Accessed March 2008
94. Viégas FB, Wattenberg M (2007) Many Eyes: http://services.alphaworks.ibm.com/manyeyes/page/Network_Diagram.html. Accessed March 2008
95. Visual complexity: http://www.visualcomplexity.com/vc/. Accessed March 2008
96. yWorks/yFiles: http://www.yworks.com/en/products_yfiles_about.htm. Accessed March 2008

Books and Reviews

Bertin J (1967) Sémiologie graphique. Les diagrammes, les réseaux, les cartes. Mouton/Gauthier-Villars, Paris/La Haye

Brandes U, Erlebach T (eds) (2005) Network analysis: Methodological foundations. LNCS. Springer, Berlin

Carrington PJ, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge

de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge

di Battista G, Eades P, Tamassia R, Tollis IG (1999) Graph drawing: Algorithms for the visualization of graphs. Prentice Hall, Englewood Cliffs

Jünger M, Mutzel P (eds) (2003) Graph drawing software. Springer, Berlin

Kaufmann M, Wagner D (2001) Drawing graphs, methods and models. Springer, Berlin

Tufte ER (1983) The visual display of quantitative information. Graphics, Cheshire

Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge

Wilkinson L (2000) The grammar of graphics. Statistics and Computing. Springer, Berlin

Complex Systems and Emergent Phenomena

HENRIK JELDTOFT JENSEN
Institute for Mathematical Sciences and Department of Mathematics, Imperial College London, London, UK

Article Outline

Glossary

Correlations is the degree, to which events at different positions and at different times depend on or influence each other, is measured by correlation functions. If two events are statistically independent, the correlation between them is zero. The opposite is not necessarily the case, but one will often expect that if correlations decay, the mutual dependence does likewise.

Correlation function describes correlations between two quantities and depends on their separation in time and space.

Complex systems consist of a large number of interacting components. The interactions give rise to emergent hierarchical structures. The components of the system and properties at systems level typically change with time. A complex system is inherently open and its boundaries often a matter of convention.

Equilibrium In statistical mechanics the prototype equilibrium system consists of a "small" system in thermal contact with another system, the latter being big enough to act as a heat bath. A heat bath is defined as a system so big that when it exchanges energy with the small system the temperature of the heat bath remains the same. The statistical properties of equilibrium systems are independent of time.

Generalized rigidity is a term introduced by P.W. Anderson [1] to describe the situation, when a many component system acts as a globally connected unit, in the sense that if one apply a force at one point, the effect can be transmitted across the system. Ice has rigidity, if we push at one point, the entire piece of ice will start moving. If the ice, on the other hand melts to water, a force applied locally will only have an effect locally.

Hamiltonian expresses the energy of a system as a function of the degrees of freedom, in terms of which the system is defined at the considered level of description. Emergence in physical systems can sometimes be understood in terms of lumping degrees of freedom, in the Hamiltonian, together in sets of effective degrees of freedom, e. g. the center of mass of a solid body.

Non-equilibrium systems is a term used to describe any system that is not in equilibrium. Needless to say this is a characterization of limited value, since there are many very different types of systems included in this category.

Order parameter is a quantity that allows one to discriminate between two phases of a physical system. The order parameter changes from zero to non-zero as one passes from one phase to the other. To identify the relevant order parameter is often non-trivial and, is in itself, a first important step.

Renormalization group analysis is a systematic mathematical procedure that enables a derivation of the emergent behavior at the macroscopic systems level. The behavior at long length and time scales is obtained from the underlying microskopic short length scales and fast dynamics.

Statistical mechanics seeks to understand how properties at systems level emerge from the level of the system-components and their interactions. This often involves the application of probability theory, and a number of mathematical techniques. Throughout, we draw a distinction between statistical mechanics and statistical physics. The latter is mainly concerned with the microscopic foundation of thermodynamics and, e. g., phenomena such as phase transitions and superconductivity.

Definition of the Subject

Matter in the universe is organized in a hierarchical structure. At the bottom (if there is one) we have elementary particles, atoms and molecules from which we get macro molecules like proteins and DNA, these are the building blocks of organelles, which together form the cells. From cells we get organs, which put together form organisms: animals and plants of a great variety of species. One level of structure emerges from the level below. Is it possible to scientifically describe, let alone, predict emergence. Sometimes emergence is described as a phenomena beyond analysis. The perplexity with which this concept is sometimes met, is well illustrated by the following quote from a recent call for participation in a meeting, held by the British research council EPSRC, to look at ways to explore emergence in complex systems. Emergence is described in the following words: "For the first time since the enlightenment in the western tradition we have started to understand that there are non-causal systems in which some things just "are". The concept of emergence by whisch patterns of possibility arise through interactions of agents over time, accepts that even with the same starting conditions the same pattern would not necessarily repeat." Another attitude is represented by Lord Roberts May's statement in a recent lecture that "when people say something is an emergent property, it just means they don't understand the phenomena".

In this brief article we will argue that emergence is neither an empty concept nor a mysterious non-causal enigma. On the contrary, emergence is central to scientific enquiry. Emergence occurs when many components interact and combine to form an identifiable system. In philosophy this is the observation that quantitative changes accumulate and give rise to new qualitative changes. A proposition that can be traced from the ancient Greek philosophers through Hegel to Dialectical Materialism. In physics Phil Anderson famously summed the fact that new levels of organization need new types of description, up in the phrase "More is Different" [2]. By following the tradition of statistical mechanics it is sometimes possible to reduce significantly the confusion surrounding what emergence is, and how it can be investigated and described. Kenneth Wilson got the 1982 Nobel prize for his Renormalization Group theory, which is a particular beautiful method for extracting certain emergent properties with great mathematical detail and precision [3].

Introduction

Statistical mechanics is concerned with the interaction of many components. From interactions between the components at one given level the aim is to understand the collective coherent behavior, which emerges as many components are put together. It is through the interactions of the components at one level that the next level emerges. We consider a collection of *interacting* atoms and the outcome is, say, a transistor. In some cases the microscopic details of the properties of the *individual* building blocks are not so crucial. It happens that the collective behavior is controlled by general properties of the *interactions* between the components more than by the intrinsic properties of the components. A number of methods have been developed to bridge the gap between the individual components comprising a system and the collective whole. This often involves predictions of the asymptotic behavior at long distances and at long times. In particular the philosophy and technique of the Renormalization Group have been successful in a number of cases in doing this. Many other approaches exist, some in which a coarse grained description is sought similar to the one used in fluid dynamics, other times the collective systems level behavior is generated by use of Individual Based Agent models typically simulated on computers. The latter method goes beyond traditional statistical mechanics, nevertheless it is reasonable to consider simulations of agent based models as part of statistical mechanics since they are very similar in spirit to traditional statistical mechanics simulations starting from the microscopic degrees of freedom, as e. g. extensively done for the celebrated Ising model used in studies of magnetic systems, melting and many other phenomena.

In this article we will discuss a number of examples of how statistical mechanics is able to deal with emergent phenomena. Our first example is from the theory of equilibrium properties of magnetic materials. In the model the microscopic magnetic moments – or spin – combine to form macroscopic coherent vortices. The vortices are bound in pairs at low temperature. When the temperature is increased, the biggest pairs are able to unbind, or fall apart. This is a subtle collective effect caused by the smaller pairs weakening the binding force between the vortices in the biggest pairs. We use this case to describe in some detail how the statistical mechanics formalism of Boltzmann and Gibbs allows us to identify the macroscopic excitations. These macroscopic excitations are what constitute the macroscopic "components" of the system, once they are identified we can calculate the unbinding of the vortex pairs. This discussion focuses on spatial aspects. The mathematical description developed from this model has applications to a range of very different phenomena, such as melting, superconductivity, superfluidity, electrical charges in two-dimensional space and crystal growth. This is a good typical example of how the mathematical formalism of statistical mechanics is able to deliver understanding, and a description, transcending the particular and unify apparently disconnected phenomena.

Many-component systems also often exhibit emergent temporal behavior that is caused by the interactions and ensuing collective motion of the components. For example, one typically see very persistent correlations, or long time memory, in the macro-dynamics of many-component systems. A particular version of this phenomena is called $1/f$ fluctuations, and we explain below what this is, and how it is related to long time correlations or long memory effects. To make the discussion concrete we will present the details of a very simple model of diffusing particles. The model might be related to motorway traffic.

To go beyond models taken from physics we will finish by a discussion of models inspired by the observed collective behavior of social insects. Aspects of trail formation and mound building of ants and termites have been reproduced in computer simulations. Often the models consider "agents" with a tendency to perform random walks and picking up and laying down material. The agents deviate from random walking when they come across traces of smell – pheromones – laid down by other ants. This indirect interaction can lead to the formation of surprisingly intricate structures of trails and mounds.

Since essentially any activity within statistical mechanics is concerned with a description of emerging phenomena a very large literature exists and we list here only a few books of particular relevance to the view point of the present article [1,4,5,6,7,8,9,10,11,12,13] and some more specialized papers as we go along.

Equilibrium Averages

To describe how statistical mechanics is able to identify structures emerging at the macroscopic level we briefly recall how macroscopic (or systems level) quantities are obtained through averaging procedures. The reason equilibrium systems can be analyzed in particular detail is that the situation, where the systems of interest can be considered as in thermal equilibrium with a heat bath, allows for the determination of the probability weights of the individual micro-states. One starts out with the following fundamental hypothesis concerning isolated or closed systems:

- **Micro Canonical Ensemble**. For a closed system it is assumed that *all* micro-states, consistent with the macroscopic constraints, occur with *equal* probability.

The macroscopic constraints can, for example, be the total energy E (which is constant for a closed system) and the volume V. Denote by $\Omega(E, V)$ the total number of micro-states possible under these constrains. Meaning the components or particles of the system have to be located within the given volume V, and that when we add all the energies of the particles the sum must equal E. The probability $p(s)$ that the system is in a particular state s is then

$$p(s) = \frac{1}{\Omega(E, V)} \, . \tag{1}$$

Closed systems are not very interesting in the sense that one is unable to interact with them. A much more interesting situation is when the system **S** under consideration is brought in contact with a heat bath **B** or heat reservoir. The heat bath is a system so big that even when it exchanges energy with the small system of experimental interest, the heat bath remains unchanged. Say a cup of tea in contact with the Pacific Ocean. The heat bath is characterized by its temperature T. We can now use the fundamental hypothesis above to determine the probabilistic weights for the states of **S**. Since the combined system $\mathbf{B}+\mathbf{S}$ is closed the weights for the combined system is given by the Micro Canonical Ensemble, i. e. all micro-states of the combined system are equally likely. The number of micro-states for the combined system of total energy $E_{\text{Tot}} = E_{\mathbf{B}} + E_{\mathbf{S}}$ will be a product

$$\Omega_{\text{Tot}}(E_{\text{Tot}}) = \Omega_{\mathbf{B}}(E_{\mathbf{B}})\Omega_{\mathbf{S}}(E_{\mathbf{S}}) \,. \tag{2}$$

Here one neglects the interactions between the heat bath and the system. Now focus on one particular micro-state s of **S** of energy $E_{\mathbf{s}}$. Since we have a particular state, s, in mind we have $\Omega_{\mathbf{S}}(s) = 1$. This state can be combined in many ways with states of the bath **B** as long as those fulfill the constraint $E_{\text{Tot}} = E_{\mathbf{B}} + E_s$. So the probability, $p(s)$, for finding the system **S** in s, when **S** is in equilibrium with the bath, is proportional to $\Omega_{\mathbf{B}}(E_{\text{Tot}} - E_s)$. Namely

$$p(s) = \frac{\Omega_{\mathbf{B}}(E_{\text{Tot}} - E_s)}{\sum_{\text{state}} \Omega_{\mathbf{B}}(E_{\text{Tot}} - E_{\text{state}})}, \tag{3}$$

the denominator ensures normalization. In order to introduce the temperature into the mathematical formalism it turns out that we should consider the logarithm of $p(s)$. We have

$$\ln[p(s)] = \text{constant} + \ln[\Omega_{\mathbf{B}}(E_{\text{Tot}} - E_s)] \tag{4}$$

$$= \text{constant} + \ln[\Omega_{\mathbf{B}}(E_{\text{Tot}})] - \frac{\partial \ln[\Omega_{\mathbf{B}}(E_{\text{Tot}})]}{\partial E_{\text{Tot}}} E_s \tag{5}$$

$$= \text{constant} - \frac{1}{k_{\text{B}} T} E_s \,. \tag{6}$$

Here we Taylor expanded to linear order to obtain the first equality. The second equality follows, because it can be shown by use of the first and second law of thermodynamics that the temperature is given by

$$\frac{1}{k_{\text{B}} T} = \frac{\partial \ln[\Omega_{\mathbf{B}}(E_{\text{Tot}})]}{\partial E_{\text{Tot}}} \,. \tag{7}$$

This is obtained in the following way. The first and second law of thermodynamics lead to the following thermodynamic identity $\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V$ where the entropy $S = k_{\text{B}} \ln[\Omega(E)]$. Since the thermodynamic identity takes the form of an exact differential we conclude that $\partial E/\partial S = T$ from which Eq. (7) follows. We now conclude

$$p(s) = \frac{\mathrm{e}^{-\frac{E_s}{k_{\text{B}} T}}}{Z} \,, \tag{8}$$

where the constant Z is obtained from the normalization condition

$$\sum_{\text{states}} p(s) = 1 \,, \tag{9}$$

to be given by

$$Z = \sum_{\text{states}} = \mathrm{e}^{-\frac{E_s}{k_{\text{B}} T}} \,. \tag{10}$$

We conclude that the probabilistic weights, needed to calculate the average macroscopic behavior of a system in contact with a heat bath at temperature T, is given by the (Boltzmann) weights in Eq. (8). And we mention that a large number of average quantities can be calculated from the sum in Eq. (10). This important sum is called the *partition function* or partition sum. Some states, or configurations of the microscopic degrees of freedom, will contribute more to the partition sum than others, such configurations can sometimes be identified as macroscopic collective excitations. These may possess a degree of robustness and stability and can in such cases be identified as macroscopic emergent objects with specific properties that can be considered essential building blocks. Perhaps it is instructive to have the following picture in mind. Think of a pool table. To describe the motion of the balls we can either follow the trajectories of all the individual molecules making up 15 colored balls or we can notice that some of the molecules move together in a coordinated way and thereby form each of the 15 balls. We can therefore instead simply follow the trajectories of the center of mass (COM) of each of the balls. Obviously we lose a lot of information since we can't go from the COM of the balls to the motion of all the molecules; whereas we can drive the COM motion if we know the motion of all the molecules. Hence we note that emergence involves a loss of information. We will discuss in detail an important and illustrative example in the next section.

The Two-Dimensional *XY*-Model

Here we describe how the averaging procedure described in the previous section can be structured in a way that allows the introduction of new effective collective degrees of freedom. These describe macroscopic excitations created by the coherent motion of a huge number of microscopic variables. When the collective degrees have been identified information concerning the detailed motion of the microscopic variables can be neglected, and one is in this way able to reduce the computational effort needed and at the same time identify the essential emergent structures. A particular clear example of this procedure consists of the physics of systems modeled by the so-called two-dimensional XY-model.

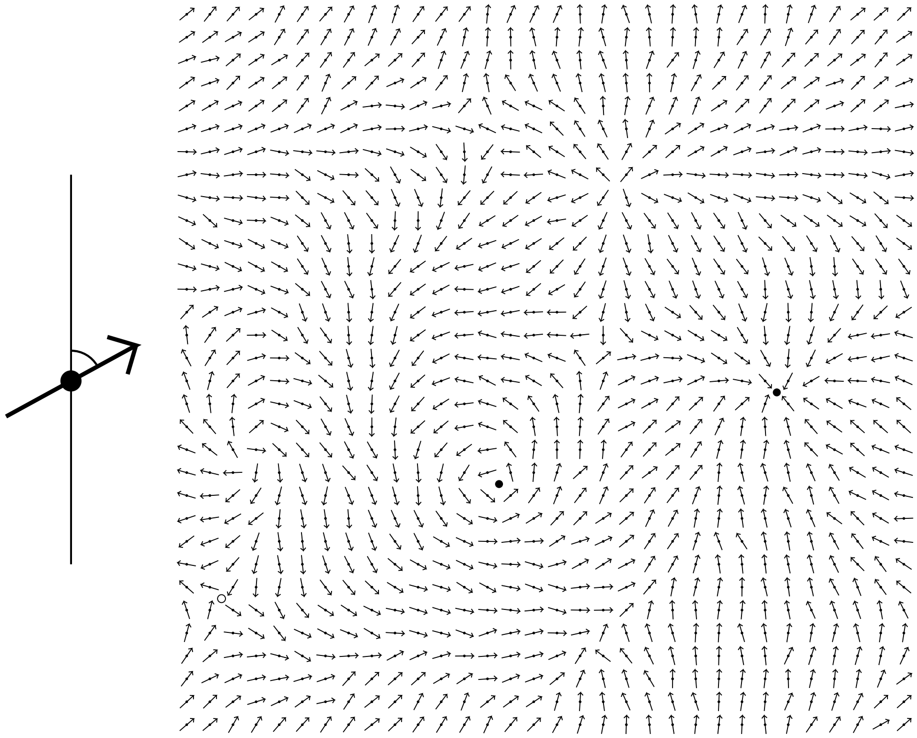
We start our discussion by considering the formation of vortices in the sea of two-dimensional magnetic moments. The individual microscopic magnetic moments sit on the sites of a two-dimensional square lattice and the direction of the moments are confined to two dimensions, see Fig. 1 (A beautiful online interactive simulation can be found on Hans Weber's web page at http://www.mt.luth.se/~weber/). Each magnetic moment can be thought of as a magnetic needle, or an arrow, confined to two dimensions and pointing in a specific direction described by the angle θ. The magnetic moment number i is given by the vector $\mathbf{S}_i = (\cos(\theta_i), \sin(\theta_i))$.

We will use it as our reference model. We think of the model as consisting of planar rotors of unit length arranged on a two-dimensional square lattice. The Hamiltonian of the system is given by

$$H = -J \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j = -J \sum_{\langle i,j \rangle} \cos(\theta_i - \theta_j) \,. \qquad (11)$$

Here J is the coupling constant between the magnetic moments, $\langle i, j \rangle$ denotes summation over all nearest neighbor sites in the lattice, and θ_i denotes the angle of the rotor on site i with respect to some (arbitrary) polar direction in the two-dimensional vector space containing the rotors.

We shall see below how the components, the rotors, work together to form certain collective coherent structures: topological defects or topological charges. In Fig. 1 these excitations are depicted. Each consists of a whirl or vortex in the configuration of the rotors. There are vortices of opposite sign. As one moves around in a positive direction along a contour encircling the center of a vortex (black circle) the rotors perform a full rotation in the positive direction as well. When we move around one of the anti-vortices (white circle) in a similar way, the rotors undergo a full rotation in the negative direction. Although these charges are here seen as arising from rotors or magnetic moments, the impressive fact is that these topological charges also represent Coulomb charges in two dimensions, or dislocations in two-dimensional crystals, or vortices in two-dimensional superconductors or a large number of other collective excitations. The interaction between



Complex Systems and Emergent Phenomena, Figure 1
Rotor configuration of the *XY*-model. Vortices (*black circle*) and anti-vortices (*white circle*) are clearly seen. The configuration is neutral, i. e. there is an equal number of vortices and anti-vortices; but only one white circle is indicated. The reader may find it amusing to try to locate the missing anti-vortex

the topological charges depends in all cases logarithmically on the spatial separation and this leads to some very general collective behavior, most spectacular the logarithmic dependence on separation causes a certain type of phase transition: the Kosterlitz–Thouless transition [14].

If we assume that the direction of the rotors varies smoothly from site to site, we can approximate $\cos(\theta_i - \theta_j)$ by the first two terms $1 - 1/2(\theta_i - \theta_j)^2$ in the Taylor expansion of cos. The sum over the nearest neighbors corresponds to the discrete Laplace operator, which we can express in terms of partial derivatives through $\theta_i - \theta_j = \partial_x \theta$ for two sites i and j which differs by one lattice spacing in the x-direction. This leads to the continuum Hamiltonian

$$H = E_0 + \frac{J}{2} \int \mathrm{d}\mathbf{r} (\nabla\theta)^2 \, . \tag{12}$$

Here $E_0 = 2JN$ is the energy of the completely aligned ground state of N rotors.

The thermodynamics of the system is obtained from the partition function

$$Z = \mathrm{e}^{-\beta E_0} \int D[\theta] \exp\left\{ -\beta \frac{J}{2} \int \mathrm{d}\mathbf{r} (\nabla\theta)^2 \right\} \, , \tag{13}$$

a functional integral over all possible configurations of the director field $\theta(\mathbf{r})$. Not all configurations will be of the same importance. By focusing on the terms in the sum that contribute most, we can identify the configurations that may be used as building blocks at the next level of description. Since the energy appears in the exponential with a negative sign in front, the most significant contributions will be those with the smaller energy – thus we have to pick out the local minima. We therefore divide the integral over $\theta(\mathbf{r})$ into a sum over the local minima θ_{vor} of $H[\theta]$ plus fluctuations θ_{sw} around the minima

$$\begin{aligned} Z = \mathrm{e}^{-\beta E_0} \sum_{\theta_{\mathrm{vor}}} \int D[\theta_{\mathrm{sw}}] \exp \Bigg\{ & -\beta (H[\theta_{\mathrm{vor}}] \\ & + \frac{1}{2} \int \mathrm{d}\mathbf{r}_1 \int \mathrm{d}\mathbf{r}_2 \theta_{\mathrm{sw}}(\mathbf{r}_1) \frac{\delta^2 H}{\delta\theta(\mathbf{r}_1)\delta\theta(\mathbf{r}_2)} \theta_{\mathrm{sw}}(\mathbf{r}_2)) \Bigg\} \, . \end{aligned} \tag{14}$$

The field configurations corresponding to local minima of H are solutions to the extremal condition

$$\frac{\delta H}{\delta\theta(\mathbf{r})} = 0 \;\Rightarrow\; \nabla^2\theta(\mathbf{r}) = 0 \, . \tag{15}$$

There are two types of solutions to this equation. The first consists of the ground state $\theta(\mathbf{r}) =$ constant. The second type of solutions consist of vortices in the director field (see Fig. 1) and are obtained by imposing the following set of boundary conditions on the circulation integral of $\theta(\mathbf{r})$:

1) For all closed curves encircling the position $\mathbf{r}_0$ of the center of the vortex

$$\oint \nabla\theta(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 2\pi n \, , \quad n = 1, 2, \ldots \, . \tag{16}$$

2) For all paths that don't encircle the vortex position $\mathbf{r}_0$

$$\oint \nabla\theta(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 0 \, . \tag{17}$$

Condition 1) imposes a singularity in the director field. Note the circulation integral *must* be equal to an integer times 2π since we circle a closed path and therefore $\theta(\mathbf{r})$ has to point in the same direction after traversing the path as it did when we started.

We can estimate the energy of a vortex in the following way. The problem is spherical symmetric, hence the vortex field θ_{vor} must be of the form $\theta(\mathbf{r}) = \theta(r)$. The dependence on r can be found from Eq. (16). We calculate the circulation integral along a circle of radius r centered at the position $\mathbf{r}_0$ of the vortex

$$2\pi n = \oint \nabla\theta(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 2\pi r |\nabla\theta| \, . \tag{18}$$

We solve and obtain $|\nabla\theta(r)| = n/r$. Substitute this result into the Hamiltonian Eq. (12)

$$E_{\mathrm{vor}} - E_0 = \frac{J}{2} \int \mathrm{d}\mathbf{r} [\nabla\theta(\mathbf{r})]^2 \tag{19}$$

$$= \frac{Jn^2}{2} \int_0^{2\pi} \mathrm{d}\phi \int_a^L r \, \mathrm{d}r \frac{1}{r^2} \tag{20}$$

$$= \pi n^2 J \ln \frac{L}{a} \, . \tag{21}$$

Here a denotes the lattice constant and L is the linear size of the considered lattice. The circulation condition Eq. (16) creates a distortion in the phase field $\theta(\mathbf{r})$ that persists infinitely far from the center of the vortex. $|\nabla\theta|$ decays only as $1/r$ leading to a logarithmic divergence of the energy. Hence we need to take into account that the integral over r in Eq. (20) is cut-off for large r-values by the finite system size L and for small r-values by the lattice spacing a. We recall that our continuum Hamiltonian is an approximation to the lattice Hamiltonian in Eq. (11). A vortex with the factor n in Eq. (16) larger than one is called multiple charged. We notice that the energy of the vortex is quadratic in the charge. In a macroscopically large system even the energy of a single charge vortex will be large, and therefore we do not expect individual vortices to be thermally induced.

Consider now a pair consisting of a single charged vortex and a single charged anti-vortex. When we encircle the vortex, we pick up $\oint \mathrm{d}\mathbf{l} \cdot \nabla\theta = 2\pi$ and when we encircle the anti-vortex, we pick up $\oint \mathrm{d}\mathbf{l}\cdot\nabla\theta = -2\pi$. Hence, if we choose a path large enough to enclose both vortices, we pick up a circulation of the phase equal to $2\pi + (-2\pi) = 0$. I. e. the distortion of the phase field $\theta(\mathbf{r})$ from the vortex-anti-vortex pair is able to cancel out at distances from the center of the two vortices large compared to the separation R between the vortex and the anti-vortex, see Fig. 1. This explains why the energy of the vortex pair is of the form [15,16,17]

$$E_{2\mathrm{vor}}(R) = 2E_c + E_1 \ln(R/a) \,. \tag{22}$$

Where E_c is the energy of a vortex core and E_1 is proportional to J. In detail, the phase field $\theta_{2\mathrm{vor}}(\mathbf{r})$ of a vortex located at $\mathbf{r} = (-a, 0)$ and an anti-vortex located at $\mathbf{r} = (a, 0)$ is given by [15]

$$\theta_{2\mathrm{vor}}(\mathbf{r}) = \mathrm{arctg}\left(\frac{2ay}{a^2 - r^2}\right) \,. \tag{23}$$

Significant aspects of the macroscopic behavior of the XY-model can be understood by treating the vortices as particles characterized by their position and their charge and ignoring the underlying sea of rotors. Indication of this follows from the expression for the energy of a pair of vortices in Eq. 22. This energy is given in terms of the relative position of the two vortices, no reference is needed to the microscopic rotor field given in Eq. 23. The pairs of vortices have dramatic effects on the macroscopic behavior of the XY-model. At low temperature the vortices are organized in fairly small bound pairs, as the temperature is increased and more thermal energy is available the separation between paired up vortices grow and at a certain temperature the pairs break apart with the effect that the individual vortices now can move freely around as they are no longer kept in check by their partner of the opposite charge. The result is the Kosterlitz–Thouless transition which manifests itself in various ways in different realizations of the XY-model. Before we discuss this transition we will look at the average ordering of the rotors. This quantity – the magnetization – is usually able to monitor if a dramatic change in the macroscopic behavior occurs as a function of temperature. But not so in the 2d XY-model. To understand this makes it clearer how important it is to identify correctly the emergent excitations of a many component system.

Lack of Ordering in Two Dimensions

In order to highlight the peculiarity of two dimensions we consider the d-dimensional XY-model. We imagine a d-dimensional cubic lattice. Each lattice site contains a planar rotor or a phase. In the continuum limit the Hamiltonian is still given by Eq. (12) except the integral over $\mathbf{r}$ is now a d-dimensional integral and therefore the factor J is replaced by Ja^{2-d}. The average size of the projection of the rotors along, say, the x-direction in $\mathbf{S}$ space, i. e. the magnetization, is

$$\langle S_x\rangle = \langle\cos\theta(\mathbf{r})\rangle \tag{24}$$

$$= \langle\cos\theta(0)\rangle \,. \tag{25}$$

Note that we might as well have chosen the y-direction. The model is isotropic and the x and the y directions are equivalent. When $\langle S_x\rangle \neq 0$ a preferred direction is singled out in the sense that on average $\mathbf{S}$ points in the direction given by $\langle S_x\rangle$. In this case we say that the rotor field possesses order. In contrast if $\langle S_x\rangle = 0$ we also have $\langle S_y\rangle = 0$, since the model is isotropic. The zero projection comes about because the rotors circulate around and on average point equally much in all directions. So we say that the rotor field is disordered or does not possess any ordering.

First we neglect the singular vortex contributions (which is perfectly safe at low temperature) and Fourier transform the phase field

$$\theta(\mathbf{r}) = \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^d}\hat{\theta}(\mathbf{k})\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}} \tag{26}$$

$$\theta(0) = \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^d}\hat{\theta}(\mathbf{k}) \tag{27}$$

$$\int \mathrm{d}\mathbf{r}(\nabla\theta)^2 = \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^d} k^2\theta(\hat{\mathbf{k}})\hat{\theta}(-\mathbf{k}) \,. \tag{28}$$

These equations are substituted into the expression

$$\begin{aligned}\langle S_x\rangle &= \frac{\int D[\theta]\cos(\theta(0))\mathrm{e}^{-\beta H}}{\int D[\theta]\mathrm{e}^{-\beta H}}\\ &= Re\left(\frac{\int D[\theta]\mathrm{e}^{-\beta H + i\theta(0)}}{Z}\right) \,.\end{aligned} \tag{29}$$

After some algebra one obtains the following expression

$$\langle S_x\rangle = \exp\left(-\frac{T}{2Ja^{2-d}}S_d\int_{\pi/L}^{\pi/a}\mathrm{d}k k^{d-3}\right) \,. \tag{30}$$

The behavior of $\langle S_x\rangle$ is controlled by the integral

$$I(L) = \int_{\frac{\pi}{L}}^{\frac{\pi}{a}} \mathrm{d}k k^{d-3} \,. \tag{31}$$

The behavior of $I(L)$ strongly depends on the dimension d. For $d < 2$ we have $I(L) \sim L^{2-d} \to \infty$ as $L \to \infty$. Hence,

$\langle S_x \rangle = 0$ in the limit of large systems for dimensions less than 2. For $d > 2$ we have that

$$I(L) \to A = \frac{1}{d-2}\left(\frac{\pi}{a}\right)^{d-2} \tag{32}$$

and therefore

$$\langle S_x \rangle = \exp\left(-\frac{S_d}{2Ja^{2-d}} AT\right) > 0\,. \tag{33}$$

Finally for $d = 2$ the integral $I(L)$ is logarithmically divergent $I(L) = \ln(L/a)$ which is sufficient to force $\langle S_x \rangle$ to zero for any non-zero temperature.

We conclude that there is no ordered phase according to the behavior of $\langle S_x \rangle$ at low temperature for $d \le 2$. For $d < 2$ this means that there is no phase transition. The same was for a while thought to be the case for $d = 2$. Since the thermal motion included in the calculation of $\langle S_x \rangle$ is able to prevent a preferred direction and hence ensure $\langle S_x \rangle = 0$ for $T > 0$ including other types of excitations, such as vortices, can surely not make $\langle S_x \rangle$ different from zero. So it is safe to conclude that the rotors are unable to order along a common direction for any non-zero temperature and it was accordingly expected that a phase transition in the 2d YX-model was excluded. This conclusion was reached since in magnetic systems the average of the local magnetic moment, i. e. $\langle S_x \rangle$ is the order parameter and the phase transition takes place at the temperature where the order parameter changes from zero to a non-zero value. It turned out that by identifying the vortices as emergent collective excitations and by understanding their physical effects, a phase transition of a new kind was discovered in the 2d XY-model.

Vortex Unbinding

An indication of the importance of vortices as the temperature is increased can be obtained from the following simple and heuristic argument. We estimate the free energy of a single vortex. The Helmholtz free energy is given by the difference between the energy and the entropy multiplied by the temperature $F = E - TS$. The energy is given by Eq. (21). We estimate the entropy from the number of places where we can position the vortex center, namely on each of the $(L/a)^2$ plaquette of the square lattice, i. e., $S = k_B \ln(L^2/a^2)$. Accordingly, the free energy is given by

$$F = E_0 + (\pi J - 2k_B T)\ln(L/a)\,. \tag{34}$$

For $T < \pi J/2k_B$ the free energy will diverge to plus infinity as $L \to \infty$. At temperatures $T > \pi J/2k_B$ the system can lower its free energy by producing vortices: $F \to -\infty$ as $L \to \infty$. This simple heuristic argument points to the fact that the logarithmic dependence on system size of the energy of the vortex combines with the logarithmic dependence of the entropy to produce the subtleties of the vortex unbinding transition. Assume a different dependence of the energy on system size and one will either have thermal activation of vortices at all temperatures (in case $E_{\text{vor}} \to \text{const.} < \infty$) or vortices will not be activated at any temperature (in case $E_{\text{vor}} \sim (L/a)^b$ with $b > 0$). It is the logarithmic size dependence of the 2d vortex energy that allows the outcome of the competition between the entropy and the energy to change qualitatively at a certain finite temperature T_{KT}.

In reality it is not single vortices of the same sign that proliferate at a certain temperature. What happens is that the larger vortex pairs which are bound together for temperatures below T_{KT} unbind at T_{KT}. This is a collective effect that can be treated quantitatively by use of a special Renormalization Group method design by Kosterlitz [18]. The vortex pairs induced as one approaches T_{KT} disturb the phase field so much that the effective value of the vortex binding term E_1 in the vortex pair free energy, that is Eq. (22) generalized to non-zero temperature, is driven to zero for large vortex separations. In the next section we shall see in detail how this happens.

The Spin Wave Stiffness

As our concern in this article is with emergent entities, we will now briefly discuss how a focus on, and an understanding of, the vortex degrees of freedom makes it possible to identify and describe the previously "hidden" phase transition in the 2d XY-system.

The effect of the thermally activated vortex pairs is describe by the temperature dependent *spin wave stiffness* ρ_s^R. This is an example of what Philip W. Anderson calls a generalized rigidity [1]. The spin wave stiffness describes how much free energy it costs to apply a twist, or gradient, to the rotors (also called spins):

$$\theta(\mathbf{r}) = \theta_0(\mathbf{r}) + \mathbf{v}_{\text{ex}} \cdot \mathbf{r}\,, \tag{35}$$

here $\theta_0(\mathbf{r})$ is allowed to vary according to the canonical ensemble. The increase in the free energy is given by

$$F(\mathbf{v}_{\text{ex}}) - F(0) = \frac{1}{2} V \rho_s^R v_{\text{ex}}^2\,. \tag{36}$$

A number of comments concerning the notation are illuminating. The notation $\mathbf{v}_{\text{ex}}$ for the gradient applied to the phase field $\theta(\mathbf{r})$ has its origin in the fact that the same physics, as we describe here, applies to superfluid films and superconducting films. In these cases the field $\theta(\mathbf{r})$ is the phase of the complex order parameter, the wave function of the super-fluid. Being the phase of a quantum mechanical wave function the gradient of $\theta(\mathbf{r})$ is related to a probability current and thereby to the velocity field of the superfluid. The notation ρ_s^R is meant to remind one that this phase rigidity, is determined by the **density** of superfluid in the case of a superfluid or a superconductor. The superscript R in ρ_s^R indicates that thermal excitations renormalize the quantity. It follows immediately from the Hamiltonian in Eq. (12) that at zero temperature $\rho_s^R = J = \rho_s$. The spin wave stiffness is similar to the shear constant of a material. The shear constant determines how the (free) energy increase when a shear deformation is imposed. As temperature is increased the shear constant decreases and drops abruptly to zero when the solid melts into a liquid.

To obtain ρ_s^R one calculates the left hand side of Eq. (36). Details can be found in the wonderful book by Chaikin and Lubensky [9]. The phase field is split into two parts

$$\theta_0(\mathbf{r}) = \theta_s(\mathbf{r}) + \theta_v(\mathbf{r}) \,, \tag{37}$$

where the first term describes smooth spin waves and the second term contains the singular vortex contribution. The free energy is obtained from $F = k_B T \ln Z$ and the partition function is given by Eq. (13). To calculate Z introduce Fourier transforms of the phase field. After quite a bit of algebra one arrives at the following simple expression

$$\rho_s^R = \rho_s - \frac{1}{2}\frac{\rho_s^2}{T} \lim_{k\to 0} \frac{\langle \hat{n}(\mathbf{k})\hat{n}(-\mathbf{k})\rangle_0}{k^2} \,, \tag{38}$$

which expresses the renormalized stiffness in terms of the correlation function of the Fourier transform of the vortex density function

$$n(\mathbf{r}) = \sum_\alpha n_\alpha \delta(\mathbf{r} - \mathbf{r}_\alpha) \,, \tag{39}$$

for a collection of vortices of charge n_α (see Eq. (16)) with centers located at positions $\mathbf{r}_\alpha$. The vortices are now described entirely by their position exactly like if they were ordinary particles. So what started out as a complex configuration in the field of rotors is now possible to treat as point particles. The effect of the extended disturbance of the rotor field is taken care of by the interaction energy between two vortices. The thermodynamic average in Eq. (38) is over the canonical ensemble with no twist imposed, hence the subscript 0. Eq. (38) can be used to determine how the spin wave stiffness behaves at large distances as a function of temperature. We will discuss how in the next section.

The KT Transition

Let us first summarize the phenomenology of the Kosterlitz–Thouless transition. As the temperature is increased more and more vortex pairs are thermally activated. This makes ρ_s^R decrease, see Eq. (38). This corresponds to a decrease in the increment of the free energy induced by a certain twist $\mathbf{v}_{ex}$. We can understand the effect from the fact that the phase field $\theta(\mathbf{r})$ becomes more and more distorted as the temperature is increased, hence the extra perturbation caused by $\mathbf{v}_{ex}$ becomes relatively less important. Quantitatively one finds

$$\rho_s^R = \begin{cases} \rho_s^R(T_{KT}^-)[1 + \text{const.}(T_{KT} - T)^{1/2}] & \text{for } T < T_{KT} \\ 0 & \text{for } T > T_{KT} \,. \end{cases} \tag{40}$$

Here, T_{KT} is the Kosterlitz–Thouless temperature at which vortex pairs unbind. The value of T_{KT} differs from one system to another. In the 2d XY-model $T_{KT}/J \simeq 0.893 \pm 0.002$ [19]. The remarkable thing is, that the ratio

$$\rho_s^R(T_{KT}^-)/T_{KT} = 2/\pi \tag{41}$$

is universal for all systems that undergoes a KT-transition. Since $\rho_s^R(T_{KT}^+) = 0$ Eq. (41) is referred to as the universal jump. The correlation length $\xi(T)$ behaves in a very unusual way as one approaches T_{KT} from above. We are used to a relatively slow algebraic divergence of the correlation length as the critical temperature is approached. For the KT-transition the divergence is, however, much faster

$$\xi(T) \sim \exp\left(\frac{\text{const.}}{(T - T_{KT})^{1/2}}\right) \quad \text{for} \quad T > T_{KT} \,. \tag{42}$$

Can we in a simple way understand this exponential divergence? Yes, we can. The phase field is significantly distorted by unbound vortices, since these vortices are not screened by a nearby anti-vortex. I. e. the phases $\theta(\mathbf{r})$ can remain correlated over distances shorter than the typical distance $D = 1/\sqrt{n_{ub}}$ between unbound vortices of density n_{ub} [20]. Or in other words, we expect the correlation length $\xi \sim D$. The vortices are thermally induced and therefore their density is expected to depend on the temperature through a Boltzmann factor $\exp(-E_{vor}/T)$. The situation described here is exactly what happens in the one-dimensional so-called ϕ^4 model. This model supports thermally activated solitons. The correlation length is set by the inverse of the soliton density and diverges exponentially as the temperature goes to zero [21]. The same thing, in a slightly simpler version, also happens in the one-dimensional Ising model. This argument can indicate the cause of the exponential dependence of ξ. But it is no more than an indication since the exponential dependence in Eq. (42) is significantly different from a simple Boltzmann factor. This difference is due to corrective renormalization effects.

Continuous phase transitions are accompanied by divergences in thermodynamic quantities caused by the divergence of the correlation length as the critical temperature T_c is approached. The singular part of the free energy density f can be estimated as the amount of thermal energy T_c within a correlated volume ξ^d, which gives $f \sim T_c/\xi^d$. The specific heat c_V is given by the second derivative of the free energy $c_V = -T\partial^2 f/\partial T^2 \sim \partial^2 \xi^{-d} \partial T^2$. For the KT-transition the exponential divergence of $\xi(T)$ in Eq. (42) is so rapid and occur over such a narrow temperature range that the divergence in c_V cannot be resolved in simulations or in experiment. This is another reason why the vortex unbinding transition remained unnoticed for so long. It doesn't leave any dramatic signature in the thermodynamic quantities. However, as mentioned above, the macroscopic rigidity clearly changes at the transition.

The Vortex Unbinding Transition in Other Systems

We have above mentioned that not only the XY-model exhibits the Kosterlitz–Thouless vortex unbinding transition [22]. Any two-dimensional system that supports thermally induced "charges" or topological defects that interact logarithmically will undergo this transition. The $U(1)$ symmetry of the phase field $\theta(\mathbf{r})$ of the XY-model is also present in the Ginzburg–Landau free energy of superfluids and of superconductors. The topological excitations in the case of a superfluid consist of vortices in the flow of the superfluid. Vortices like those observed when one empties a bath tub. In thin superfluid helium film such vortices destruct the superfluid phase with increasing temperature according to the scenario of the KT-transition [23,24].

The situation is slightly more complicated in superconductors. Because the superfluid in this case is charged (the superconducting pairs of electrons), screening effects play a role [9,10,22]. However, for thin superconducting films of thickness δ the effective screening length is given by $\lambda_{\text{eff}} = \lambda^2/\delta$, which can easily become a macroscopic length. In this case the loss of superconductivity is caused by the unbinding of vortex pairs according to the KT-transition. The broken pairs can move freely when they respond to an applied electric current. As they move they cause phase slips in the superconducting order parameter. These phase slips induce a voltage drop according to the Josephson relation. The superconductor is now unable to support an electric current without a voltage drop, i. e. it is not a superconductor any longer [25,26,27,28].

Dislocations in two-dimensional crystals interact through the strain field. Two edge dislocations of opposite sign correspond to an extra row of atoms inserted along the line connecting the location of the two dislocation cores. The extra line of atoms produces strain and leads to an increase in the energy which is logarithmic in the separation between the two dislocations. Thus, the situation is very similar to the one encountered in the XY-model. When the dislocations unbind, free dislocations are produced. A shear applied to the system can now be accommodated by the mobile dislocations without an increase in the (free) energy. I. e., the shear constant has dropped to zero and the system is melted. The 2d melting theory of Kosterlitz–Thouless–Halperin–Nelson–Young predicts that melting occurs in two stages. At the first stage dislocations unbind and make the shear constant drop to zero. The dislocations are topological defects, their effect on the order of the lattice are, however, not very dramatic. Before the unbinding of dislocations, the translational and the orientational order of the lattice are both described by correlation functions that depend algebraically on distance. When the dislocations unbind the translational correlation function becomes exponential but the orientational correlations remain algebraic. At a somewhat higher temperature topological defects called disclinations unbind with the effect that the orientational order becomes exponential. Details can be found in Chaikin and Lubensky [9].

There are many other cases where the logarithmic vortex interaction and the KT-transition play a role. For instance, the shape of surfaces in three dimensions may undergo a transition from smooth to rough [29]. Assume that the surface energy of the two-dimensional surface is proportional to the area of the surface. And assume that the surface is defined in terms of its height $h(x, y)$ above the xy-plane, i. e. no over hangs. In other words the points on the surface have the coordinates $(x, y, h(x, y))$. The Hamiltonian for the surface is then

$$H = \sigma \int \mathrm{d}x \int \mathrm{d}y \sqrt{1 + (\nabla h)^2} \,. \tag{43}$$

Here σ is a measure of the surface tension. If the height only varies slowly as a function of (x, y) we can assume $|\nabla h| \ll 1$ and then expand the square root. In this approximation the Hamiltonian in Eq. (43) can be written as

$$H = \sigma L^2 + \frac{1}{2} \int \mathrm{d}x \int \mathrm{d}y (\nabla h)^2 \,, \tag{44}$$

(L is the linear size of the system in the xy-plane) which is equivalent to Eq. (19), and we expect the same physical phenomenology to apply to the surface as we found for the XY-model.

Non-Equilibrium Systems

The Boltzmann probabilities cannot be used to calculate the macroscopic properties when we deal with situations that have no equivalence among systems in contact with a heat bath. At present there is no general procedure for the determination of the probability weights of the individual microstates. This doesn't mean that statical mechanics can't describe how macroscopic properties emerge in systems out of equilibrium. Indeed approaches of broad interest exist. Here we will briefly introduce the use of Langevin equations and algorithmic models through two concrete simplistic models. The first is inspired by the very long memory or correlation times often observed in complex systems. We will discuss slow flowing motorway traffic. The other is concerned with the emergence of self-organized structures among interacting living organisms. Specifically we will think of the formation of ant trails.

1/f Fluctuations – a Langevin Approach

In this section we describe an example of emergence in time. We will assume that interaction between the components of our system forces these to diffuse around, rather than to move around in a ballistic manner. The result is a time signal that contains very strong correlations and is characterized by what is denoted a $1/f$ power spectrum [30].

We want to study correlations in a time signal $f(t)$. We measure the signal again and again at two times separated by T time units. To make life simple we will *neglect* the normalization factor in the empirical averages, i. e. we don't divide by the number of terms in the sum in Eq. (45) below. A justification for this is that we are interested in the functional dependence of the correlations on the time interval T and not so much interested in the actual specific value of the correlation coefficient. Since the correlation coefficient will depend on T we talk about the correlation function. Moreover, since we are correlating the signal with itself we talk about the autocorrelation function given by:

$$\begin{aligned} C(T) &= \sum_t [f(t) - \langle f(t)\rangle][f(t+T) - \langle f(t+T)\rangle] \\ &= \int \mathrm{d}t [f(t) - \langle f(t)\rangle][f(t+T) - \langle f(t+T)\rangle] \\ &= \int \mathrm{d}t [f(t) - \langle f(t)\rangle][f(t+T) - \langle f(t)\rangle] \,. \end{aligned} \tag{45}$$

In the last equality we made use of the fact that the average of value $f(t)$ and $f(t+T)$ are identical.

The autocorrelation function is an important object for the study of memory effects or causality effects in a signal. The correlation function is equivalent to the *power spectrum*. The power spectrum of the signal $f(t)$ is defined as

$$S_f(\omega) = |\hat{f}(\omega)|^2 \,. \tag{46}$$

That is the absolute value squared of the Fourier transform of the signal and the power spectrum is related to the Fourier transform of the autocorrelation function:

$$S_f(\omega) = \hat{C}(\omega) \,. \tag{47}$$

This relation explains why power spectra that approximately depend inversely proportional on the frequency

$$S_f(\omega) \propto 1/\omega^\beta \,, \tag{48}$$

with $\beta \simeq 1$, are of special interest [31,32,33]. Namely, at a somewhat heuristic level, we can substitute Eq. (48) into Eq. (47) and then into

$$C(T) = \int_{-\infty}^{\infty} \mathrm{d}\omega \hat{C}(\omega) \mathrm{e}^{-i\omega t} \,, \tag{49}$$

to obtain

$$C(T) = \int_{-\infty}^{\infty} \mathrm{d}\omega \omega^{-\beta} \mathrm{e}^{-i\omega T} \tag{50}$$

$$= T^{1-\beta} \int_{-\infty}^{\infty} \mathrm{d}u u^{-\beta} \mathrm{e}^{u} \,. \tag{51}$$

We made the substitution $u = \omega T$ and note that the integral in the above equation now is independent of T. So when $\beta \simeq 1$, the correlation function $C(T)$ depends very weakly on T meaning very slow decay of correlations. This indicates the particular interest in power spectra that approximately decays as one over the frequency – called $1/f$ noise. The way we carried the argument through is slightly dangerous due to possible divergent integrals; the conclusion is, however, sound.

Transport by Diffusion

For concreteness imagine a piece of motorway stretching from $x = -\infty$ to $x = \infty$. At $x = 0$ vehicles can enter or leave at an intersection. We will develop a model for the time evolution of the density of cars $n(x, t)$ at position x at time t. Since the cars – particles – only can leave or enter our system at $x = 0$, at all other positions, $x \neq 0$, changes during a brief time interval δ in the number of particles in a small interval $[x, x+\delta x]$ about x

$$\delta n(x,t) = n(x,t+\delta t)\delta x - n(x,t)\delta x \tag{52}$$

will be caused by a difference during the time δt between the number of particles leaving the section $[x, x+\delta x]$ at $x+\delta x$ and the number of particles entering the section at x. Let $J(x,t)$ denote the particle current (number of particles crossing the position x at time t per time unit). We can then write

$$\delta n(x,t) = J(x+\delta x,t)\delta t - J(x,t)\delta t\,. \tag{53}$$

Substituting Eq. (53) into Eq. (52) we obtain

$$\begin{gathered} n(x,t+\delta t)\delta x - n(x,t)\delta x = -J(x+\delta x,t)\delta t + J(x,t)\delta t \\ \Downarrow \end{gathered} \tag{54}$$

$$\frac{\partial n(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x}\,. \tag{55}$$

The last equation follows in the limit $\delta x \to 0$ and $\delta t \to 0$.

This equation is exact and only assumes conservation of the particles. To obtain a closed equation for $n(x,\ t)$ we need to relate $J(x,\ t)$ to $n(x,\ t)$. And to do so we need to make assumptions concerning the nature of how the particles, or cars, move along the line. Let us imagine that congestion makes it impossible for cars to move freely. On the contrary assume that the drivers are forced to effectively perform random walks; i. e. diffuse along the motorway.

We emphasize that it is through this assumption that the cars, or particles, are made to interact. Here we assume that the diffusion is a result of over-crowding and interaction amongst the cars. Obviously particles might perform diffusive motion as a result of other interactions. Pollen in water diffuses because it is bombarded by large numbers of water molecules. In any case some sort of complex interaction is always responsible for diffusive motion since the particles otherwise would move around according to Newton's laws. The model might in fact be more relevant to small particles (pollen, say) suspended in a long narrow strip of water – or something else.

To model the jamming and resulting diffusive motion, we will assume that the net particle current (at coarse grained level) is from high particle density to low particle density, and linear in the density difference. We express this as

$$J(x,t) = -\gamma\frac{\partial n(x,t)}{\partial x}\,. \tag{56}$$

We combine Eq. (55) and Eq. (56) to obtain a closed form of dynamical equation for $n(x,t)$:

$$\frac{\partial n(x,t)}{\partial t} = \gamma\frac{\partial^2 n(x,t)}{\partial x^2}\,. \tag{57}$$

This is the well-known diffusion equation. It describes how inhomogeneities in the density $n(x,\ t)$ relaxes by diffusion. The equation describes a closed system. Next we include the particles that might be added or removed at a certain rate $g(x,\ t)$ at position x at time t. According to the description above we have in particular in mind that $g(x,\ t)$ must describe cars leaving and entering at $x=0$. We will later return to how this particular requirement can be imposed on $g(x,\ t)$. We now have our final equation of motion for $n(x,\ t)$

$$\frac{\partial n(x,t)}{\partial t} = \gamma\frac{\partial^2 n(x,t)}{\partial x^2} + g(x,t)\,. \tag{58}$$

This is an inhomogeneous partial differential equation and we solve it easily by Fourier transformation

$$n(x,t) = \int_{-\infty}^{\infty}\frac{\mathrm{d}k}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega}{2\pi}\hat{n}(k,\omega)\mathrm{e}^{i(kx+\omega t)}\,. \tag{59}$$

Substitute into Eq. (58) to obtain an expression for $\hat{n}(k,\omega)$ in terms of the Fourier transform of the drive $\hat{g}(k,\omega)$

$$\hat{n}(k,\omega) = \frac{\hat{g}(k,\omega)}{i\omega+\gamma k^2}\,. \tag{60}$$

Now substitute Eq. (60) into Eq. (59):

$$n(x,t) = \int_{-\infty}^{\infty}\frac{\mathrm{d}k}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega}{2\pi}\frac{\hat{g}(k,\omega)}{i\omega+\gamma k^2}\mathrm{e}^{i(kx+\omega t)}\,. \tag{61}$$

Next we want to focus on the density fluctuations at a specific position $x_0 > 0$. Therefore we define

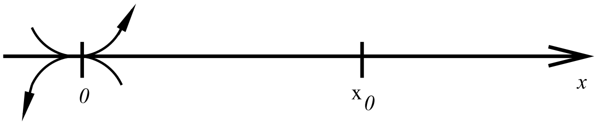
$$N(t) \equiv n(x_0,t) - \langle n(x_0,t)\rangle_t\,, \tag{62}$$

where we have subtracted the temporal averaged density. We will determine the power spectrum of $N(t)$ and for this purpose need the Fourier transform

$$\hat{N}(\omega) = \int_{-\infty}^{\infty}\mathrm{d}tN(t)\mathrm{e}^{-i\omega t} \tag{63}$$

$$= \int_{-\infty}^{\infty}\mathrm{d}t\ n(x_0,t)\mathrm{e}^{-i\omega t} - \langle n(x_0,t)\rangle_t \int_{-\infty}^{\infty}\mathrm{d}t\ \mathrm{e}^{-i\omega t} \tag{64}$$

$$= \int_{-\infty}^{\infty}\frac{\mathrm{d}k}{2\pi}\frac{\hat{g}(k,\omega)}{i\omega+\gamma k^2}\mathrm{e}^{ikx_0} - \langle n(x_0,t)\rangle_t\delta(\omega)\,. \tag{65}$$



Complex Systems and Emergent Phenomena, Figure 2
Cars/particles diffusing up and down a motorway stretching from $x = -\infty$ to $x = \infty$. At $x = 0$ an intersection allows the vehicles to enter or leave the motor way. At $x = x_0$ a traffic warden is monitoring the number of vehicles, $N(t)$, in front of him

That is how far we can go without further assumptions concerning the nature of the drive $g(x, t)$. Since this source term is meant to represent vehicles entering and leaving at position $x = 0$ we will now use

$$g(x,t) = \delta(x)\chi(t) \Rightarrow \hat{g}(\omega) = \hat{\chi}(\omega)\,. \tag{66}$$

We then have that for $x \neq 0$ the source $g(x,t) = 0$ and at $x = 0$ the temporal variation in the flow onto and away from the "motorway" is given by $\chi(t)$. From Eq. (65) we get

$$\hat{N}(\omega) = \hat{\chi}(\omega)\int_{-\infty}^{\infty} \mathrm{d}k \frac{\mathrm{e}^{ikx_0}}{i\omega + \gamma k^2}\,. \tag{67}$$

The power spectrum is finally calculated as the absolute value square of $\hat{N}(\omega)$

$$|\hat{N}(\omega)|^2 = \frac{|\hat{\chi}(\omega)|^2}{4\gamma\omega}\mathrm{e}^{-\sqrt{\frac{2\omega}{\gamma}}x_0}\,. \tag{68}$$

The power spectrum of the density fluctuations is clearly influenced by the power spectrum of $\chi(t)$. Let us assume that vehicles enter and leave at the intersection in a totally uncorrelated manner (perhaps not a totally realistic assumption) which translates into $|\hat{\chi}(\omega)|^2 =$ constant. In this case

$$|\hat{N}(\omega)|^2 \propto \frac{1}{\omega}\mathrm{e}^{-\sqrt{\frac{2\omega}{\gamma}}x_0}\,. \tag{69}$$

For frequencies so small that $\sqrt{(2\omega)/\gamma}x_0 < 1$ we have $\exp(-\sqrt{(2\omega)/\gamma}x_0) \simeq 1$ and therefore

$$|\hat{N}(\omega)|^2 \propto \frac{1}{\omega} \quad \text{for} \quad \omega < \frac{\gamma}{2x_o^2} \equiv \frac{1}{2T_{\text{diff}}}\,. \tag{70}$$

Where we introduced the time scale $T_{\text{diff}} = x_o^2/\gamma$. This is the characteristic time it takes for particles, under going diffusion with a diffusion constant γ, to move from $x = 0$ to $x = x_0$.

Very long temporal correlations, as indicated by the $1/f$ behavior of the power spectrum, is observed in very many and diverse situations: the light intensity from quasars, the ocean current, the pitch or pressure fluctuations in speech and music, the flow of traffic, the fluctuations in the resistivity of a conductor, and many more [31, 32]. Is the model we have sketched above able to explain the observed $1/f$ correlations in all these many different systems? No, probably not. Surface driven diffusion doesn't seem to be central to all these situations. The question whether a general explanation for $1/f$ exists is still an open one.

We have considered $1/f$ fluctuations here for mainly two reasons. It is a fascinating problem which is often encountered in complex systems and our discussion illustrates how one can use stochastic differential equations to go beyond equilibrium statistical mechanics to analyze temporal emergent behavior.

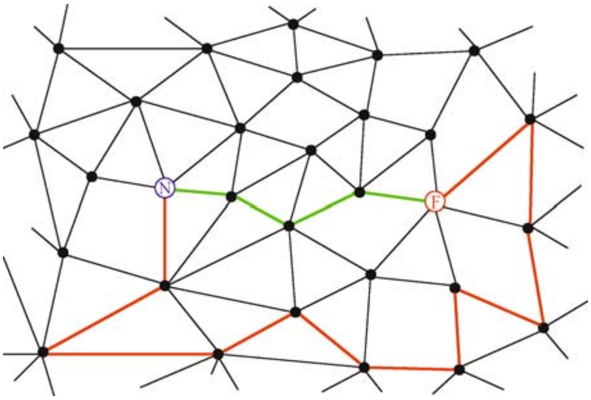
Self-organized Structure Formation – Ant Trails

As our next example of non-equilibrium statistical mechanics and emergence in complex systems we will briefly consider an algorithmic model inspired by ant trail formation. The model is schematic, simplistic and its relevance to the actual mechanism involved when real ants form trails is not known in detail. Nevertheless, the model is an interesting example of how emergent structures can appear from a dynamical algorithm.

The model describes how the path selected by ants gradually converge towards the *shortest* path between the nest and a location of food, as a result of ants' tendency to walk around at random combined with a preference for following the smell (the pheromone trail) left behind by preceding ants [12,13]. The model discussed here is presented by Danilo Benzatti at http://ai-depot.com/CollectiveIntelligence/Ant.html; but a rich literature on individual based models of formation of patterns by social insects exists [12,13,34,35,36].

We imagine a network of possible paths around the ant nest (see Fig. 3).

The nest is located at one node. Around the nest many edges connect a system of nodes. The ants can only move along the edges and in this way travel from node to node. At some of the nodes food is present. Often more than one



Complex Systems and Emergent Phenomena, Figure 3
A network of possible paths the ants can choose between as they travel from the nest at *N* to the food at *F*. The *green* short path will maintain a high level of pheromones, and will continue to attract ants, while the longer *red* path will gradually sustain a high level of pheromones and will therefore cease to attract ants

possible route will lead from the nest to the food. How do ants, with no global overview and no sophisticated means of measuring traveled distance, identify the shortest route from the nest to a food position? The assumption is that all what an individual ant is able to do is:

(1) **Ant Motion:** Perform random walks.
(2) **Effect of ant on path:** Lay down a unit of pheromone when traversing an edge.
(3) **Feedback:** Feel attracted to pheromones deposited on the edges. The more pheromones the more an ant is attracted to an edge; and the more likely it is that the ant chooses to walk along that edge.
(4) **Return to nest:** When food is found the ant follows its own pheromone trail back to the nest.

The reason this algorithm can converge towards the shortest path, is that the shortest path between the nest and the location of the food will have more ants traveling along its edges per time, and therefore will have a higher pheromone concentration than the longer paths.

The details of the algorithm is as follows. Let the nodes be enumerated by $i = 1, 2, \dots, N$. Let edges between node number i and node number j be labeled E_{ij}. Each edge carries a time-dependent pheromone weight $\phi(E_{ij},\ t)$. When an ant traverses an edge it deposits one unit of pheromone leading to

$$\phi(E_{ij}, t) \mapsto \phi(E_{ij}, t+1) = \phi(E_{ij}, t) + 1\ . \qquad (71)$$

At each node the ants chose probabilistically between the edges sprouting from the node. The probability $p(E_{ij},\ t)$ that an ant chooses a certain edge E_{ij} among all the edges connected to a specific node number i is given by

$$p(E_{ij}, t) = \frac{\phi(E_{ij}, t)}{\sum_l \phi(E_{il}, t)}\ . \qquad (72)$$

At the start of the simulation all ants are in the nest. Next they begin their random exploration of the network. When an ant locates the food, it picks up a unit of food and returns to the nest along the path it followed on its way out. On the return journey pheromones are also laid down, reinforcing this specific path. Finally it is assumed that the pheromones evaporate at a constant rate, say,

$$\phi(E_{ij}, t) \mapsto \phi(E_{ij}, t+1) = \phi(E_{ij}, t) - \nu\Theta(E_{ij}, t)\ . \quad (73)$$

(Note the Θ-function [defined as $\Theta(x) = 1$ for $x > 0$ and $\Theta(x) = 0$ for $x \le 0$] ensures that the evaporation stops when there are no more pheromones on an edge, i. e. when $\phi(E_{ij} = 0.)$ Thus, paths rarely used will lose their pheromone signature and will not be attractive to the ant, whereas the pheromone level will be maintained along the more often used paths. Thus, the shorter paths with a more frequent ant-traffic will prevail over the longer paths, since the ants on the latter paths visit the individual edges less frequently.

The collective effect of this algorithm is to find the shortest path between nest and food, although the individual ant doesn't need to know that this is what is going on. For more detail see [37] and for modeling of emergent collective intelligence and pattern formation amongst social insects see [12,13,34,35,36]. It is still an open question how to represent algorithmic models, like the one described in this section, in a precise way.

Summary and Future Directions

We have described a specific example where one can follow in great detail the steps from one level of structure to the next in the hierarchical order of matter. Topological defects arise as coherent structures of the "atoms" at one level and can be considered as (composite) particles at the next level. Their interaction can be derived from the behavior of the constituent "atoms". Many different systems may support composite particles that interact in the same way. We looked in particular at vortex physics, where the Kosterlitz–Thouless transition is caused by the logarithmic interaction between the topological defects. The one most important fact in determining the KT-transition is that both energy and entropy depend logarithmically on length scale for the two-dimensional topological charges. This example is hoped to make clear that within equilibrium statistical mechanics emergent phenomena can be described and analyzed in great quantitative detail.

When we move to systems out of equilibrium, which of course by far constitute the majority, no universally applicable formalism exists so far. We illustrated, however, by two very different examples that emergent collective behavior produced by interactions between components can also in non-equilibrium situations be modeled either by use of various mathematical techniques or through the application of computer simulations.

I will finish by suggesting the following *conclusion*. There is nothing mysterious about emergent phenomena. They are a wonderful thing – but they are not of new character, something science never has seen or dealt with before. On the contrary, I will claim that understanding emergent phenomena is exactly what all science is aimed at. Often this is not so explicitly clear as in the examples discussed above where the emphasis is explicitly on the macroscopic, or systems level, effects of the interactions between the components constituting the system. Nevertheless, even when one studies, say, atomic physics (as in contrast to statistical physics), one is dealing with the effect of interacting components. An atom consists after all of interacting protons, neutrons and electrons and the properties of the atom are the emergent result of the interactions between these particles.

Notwithstanding, the focus of the statistical mechanics approach is towards generality. As illustrated by our discussion of the XY-model and the Kosterlitz–Thouless transition, the same phenomenology can be observed in many very different systems with very different types of components (magnetic moments, atoms in a lattice, superfluids etc.), if the interactions between the components possess equivalence at a mathematical level. It is this generality that leads people to suggest that even simple model studies may sometimes be of relevance to seemingly much more complicated situations. In the future we are bound to see the statistical mechanics approach to emergent phenomena being applied to a much broader range of problems than was traditionally the case. We see attempts in fields like biology and economics, but also in linguistics, to develop pertinent statistical mechanics models. Examples include phenomena ranging from gene regulation to the social behavior of insect colonies and from stock market fluctuations to management of logistics. So far statistical mechanics has mainly been developed under the influence of physics and methods like the Renormalization Group arose. It will be very interesting to follow how statistical mechanics broadens its arsenal of tools as emergent phenomena in other fields are approached from the view point of statistical mechanics.

Acknowledgments

I am deeply grateful to many people with whom I have collaborated and learnt from during many years concerning matters relevant to the topic of the present article. In particular it is a pleasure to mention Hans Fogedby, Peter Minnhagen, and Hans Weber.

Bibliography

1. Anderson PW (1984) Basiv notions of condensed matter physics. Benjamin/Cummings, Menlo Park
2. Anderson PW (1972) More is different. Science 177:393–396
3. Wilson K (1982) The renormalization group and critical phenomena. Available via DIALOG. http://nobelprize.org/nobel_prizes/physics/laureats/1982/wilson-lecture.html. Accessed 10 Jul 2008
4. Reif F (1965) Fundamentals of statistical and thermal physics. McGraw-Hill, New York
5. Shang-Keng M (1985) Statistical mechanics. World Scientific, Singapore
6. Kadanoff LP (2000) Statistical Physics. Statics, Dynamics and Renormalization. World Scientific, Singapore
7. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) The theory of critical pnenomena. An introduction to the renormaliztion group. Oxford University Press, Oxford
8. Nelson DR (2002) Defects and geometry in condensed matter physics. Cambridge University Press, Cambridge
9. Chaikin PM, Lubensky TC (1995) Principles of condensed matter physics. Cambridge University Press, Cambridge
10. Tinkham M (2004) Introduction to superconductity, 2nd edn. Dover Publications, New York
11. Sornette D (2004) Critical phenomena in natural sciences. Chaos, fractals, selforganization and disorder: concepts and tools. Springer, Berlin
12. Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) Self-organization in biological systems. Princeton University Press, Princeton
13. Solé R, Goodwin B (2000) Signs of life, how complexity pervades biology. Basic Books, New York
14. Kosterlitz JM, Thouless DJ (1973) Ordering, metastability and phase transitions in two-dimensional systems. J Phys C Solid State Phys 6:1181–1203
15. Cataudella V, Minnhagen P (1990) Simple estimate for vortex fluctuations in connection with high-t_c superconductors. Phys C 166:442–450
16. Weber H, Jensen HJ (1991) Crossover from three- to two-dimensional behavior of the vortex energies in layered *xy*-models for high-t_c superconductors. Phys Rev B 44:454–457
17. Minnhagen P, Olsson P (1991) Monte carlo calculation of the vortex interaction for high-t_c superconductors. Phys Rev B 44:4503–4511
18. Kosterlitz JM (1974) The critical properties of the two-dimensional xy model. J Phys C Solid State Phys 7:1046–1060
19. Olsson P, Minnhagen P (1991) On the helicity modulus, the critical temperature and monte carlo simulations for the two-dimensional *xy*-model. Phys Scr 43:203–209
20. Jensen HJ, Weber H (1992) A phenomenological study of vortices in a two dimensional *xy*-model in a magnetic field. Phys Rev B 45:10468–10472

21. Jensen HJ, Fogedby HC (1985) Phonon-kink interference in the φ^4 model. Phys Scr 31:210–214
22. Minnhagen P (1987) The two-dimensional coulomb gas, vortex unbinding, and superfluid-superconducting films. Rev Mod Phys 59:1001–1066
23. Ambegaokar V, Halperin BI, Nelson DR, Siggia ED (1978) Dissipation in two-dimensional superfluids. Phys Rev Lett 40:783–786
24. Ambegaokar V, Halperin BI, Nelson DR, Siggia ED (1980) Dynamics of superfluid films. Phys Rev B 21:1806–1826
25. Kadin AM, Epstein K, Gildman AM (1983) Renormalization and the Kosterlitz–Thouless transition in a two-dimensional superconductor. Phys Rev B 27:6691–6702
26. Huberman BA, Myerson RJ, Doniach S (1978) Dissipation near the critical point of a two-dimensional superfluid. Phys Rev Lett 40:780–782
27. Doniach S, Huberman BA (1979) Topological excitations in two-dimensional superconductors. Phys Rev Lett 42:1169–1172
28. Beasley MR, Mooij JE, Orlandon TP (1979) Possibility of vortex-antivortex pair dissociation in two-dimensional superconductors. Phys Rev Lett 42:1165–1168
29. Barabási A-L, Stanley HE (1995) Fractal concepts in surface growth. Cambridge University Press, Cambridge
30. Milotti E 1/*f* noise: a pedagogical review. Available via DIALOG. http://arxiv.org/pdf/physics/0204033. Accessed 10 Jul 2008
31. Press WH (1978) Flicker noises in astronomy and elsewhere. Comments Mod Phys C 7:103–119
32. Weissman MB (1988) 1/*f* noise and other slow, nonexponential kinetics in condensed matter. Rev Mod Phys 60:537–571
33. Grinstein G, Hwa T, Jensen HJ (1992) $1/f^{\alpha}$ noise in dissipative transport. Phys Rev A 45:R559–R562
34. Schweitzer F, Lao K, Family F (1997) Active random walkers simulate trunk trail formation by ants. Biosystems 41:153–166
35. Bonabeauc E, Theraulaz G, Deneubourg J-L, Franks NR, Rafelsberger O, Joly J-L, Blanco S (1998) A model for the emergence of pillars, walls and royal chambers in termite nests. Phil Trans Royal Soc B 353:1561–1576
36. Feltell D, Bai L, Jensen HJ (2006) An individual approach to modeling emergent structure in termite swarm systems. Int J Model, Identif Control 1:43–54
37. Benzatti D http://ai-depot.com/CollectiveIntelligence/Ant.html

Composites, Multifunctional

ANETTE M. KARLSSON[1], MOSOBALAJE O. ADEOYE[2]

[1] Dept. of Mechanical Engineering,
University of Delaware, Newark, USA

[2] Dept. of Materials Science & Engineering,
Obafemi Awolowo University, Ile-Ife, Nigeria

Article Outline

Glossary

Biocomposite Biological composite materials made through biological processes usually in living organisms. Biocomposites formed through biomineralization processes are referred to as biomineral composites

Biomimetics Also known as biomimicry, is the study of biological systems with the aim of applying the methods and processes in these systems to the design of engineering materials and systems to produce engineering devices and structures with superior or comparable functionalities.

Biomineralization Mineralization carried out through biological processes to convert organic materials to inorganic materials to form biominerals. Biomineral composites are composed of inorganic minerals formed through biomineralization processes by living organisms in organic matrices of proteins and polysaccharides.

Composite A combination of two or more monophase materials arranged into a material (or "material structure") to produce one or more particular properties that are superior to the same properties exhibited by the individual component materials.

Multifunctionality This is the ability of a material or device to perform two or more functions simultaneously or consecutively.

Self healing The ability of a structure to repair damage without external intervention. For example, a small cut on a human skin will be repaired automatically by the body. Man-made structures with such ability are under early development.

Definition of the Subject

A composite material is a combination of two or more monophase materials arranged into a single entity material (or "material structure") to produce one or more particular properties that are superior to the same properties exhibited by the individual component materials. Until recently, most composites were designed with a single function in mind, most commonly focusing on improving strength or durability of the material. Many high performance composites have been developed and are performing excellently in this respect. For example, carbon fiber reinforced epoxy is now commonly used in many engineering applications. Carbon fibers have high theoretical strength and high stiffness, but are brittle and therefore readily break if a very small flaw is initiated. By embedding the fiber in epoxy, which is relatively ductile (less flaw sensitive) compared to the carbon fiber, the composite structure combines the strength and stiffness of the carbon fibers with the ductility of the epoxy. A natural extension of this concept is to develop materials that are optimized for engineering applications in conditions that demand more than one function being performed by a single material. For example, a material acting as a structural support may as well function as a thermal sensor at the same time.

A multifunctional composite, therefore, is a composite material that is capable of performing two or more functions simultaneously or consecutively. The need for improved performance in current engineering application and in entirely new areas of application has been the major driving force behind the design and development of multifunctional composites. Multifunctionality is achieved in a composite by deliberately and purposefully engineering the microsructures of the component materials and the composite as a whole usually at the micro- and nano-range levels (that is at the length scales of μm and nm). When the reinforcement scale is in the nano-region the resulting composites are commonly referred to as *nanocomposites*, which is a new frontier in materials science and engineering. Today, much effort is being geared towards research and development of multifunctional composites at various materials research departments and institutions across the world.

Introduction

Multifunctional composites are designed, through microstructural modification, to enhance or introduce new material properties in order to improve or increase the functionalities of a structure, with respect to a broad range of properties. This can include various combinations of magnetic, electronic, electrical, optical, chemical, thermal and mechanical properties. In other words, a multifunctional composite provides two or more functionalities simultaneously or sequentially with improved performance, with less complexity, cost and weight compared to a structure where these functions are provided by individual components. In many cases, a multifunctional composite is the only means by which the combination of the desired functions can be achieved.

Requirements for high performance, durability, conservation of natural resources, low cost, and in many applications miniaturization have led to research and development of multifunctional composites with high specific properties, e. g., high strength per unit weight. These have especially been the driving forces for applications in space exploration, aerospace, information technology and energy production and transmission. Thus, materials science is a critical tool to bring together different materials as composites that can perform more than one function. Such composites are designed and engineered at various scales, ranging from the atomic level and up. For example, load-bearing composites have been developed to simultaneously act as thermal sensors by utilizing carbon nanotube reinforced polymers [109].

A very important multifunctional composite group that has been researched over the past two decades is the group of carbon nanotube composites because of the remarkable mechanical, electrical, thermal and structural properties exhibited by the fiber-like carbon nanotubes reinforcing component of the composites [35]. Carbon nanotubes are added to polymeric matrices to form carbon nanotube polymer composites with excellent mechanical and electrical properties [105]. With the nanotubes the scale of reinforcement is now in the nano region hence these composites are termed nanocomposites, and will be discussed in Sect. "Functionalized Carbon Reinforced Polymer Matrix Composites".

Living organisms are made up of multifunctional materials. An example is the human skin that functions as a container and protector for all internal organs and the human structure. It also serves as a heat sensor, touch sensor, and an outlet for sweat and oil. The human skin, made up of various cells and layers of cells consisting of blood vessels, sensory receptors, glands, and hair follicles, is therefore a natural multifunctional composite. Moreover, human skin has the amazing ability of self healing. Consequently, nature serves as an inspiration for the development of multifunctional materials. A second class of multifunctional natural composites is the biomineralized materials found in living organisms. Biomineral composites are natural materials, a group of bioceramic-biopolymer composites, produced through cell-mediated processes [40]. They are composed of inorganic nano- or micro-scale amorphous or crystalline minerals formed through biologically induced or biologically controlled mineralization processes by living organisms in organic matrices of proteins and polysaccharides [42,116]. Their functions include structural support, mechanical protection, movement, grinding, and gravity or magnetic field sensing. This class of multifunctional composites is of importance in biomimetics and is of interest to both biologists and materials scientists. They are tough materials combining high hardness with high fracture resistance. Examples are bone, dentine, enamel, shells, scales, eggshells and sponge silica skeletons. There are many other biocomposites that are not organic-inorganic such as skin, wood and leaf – these are organic and are also multifunctional.

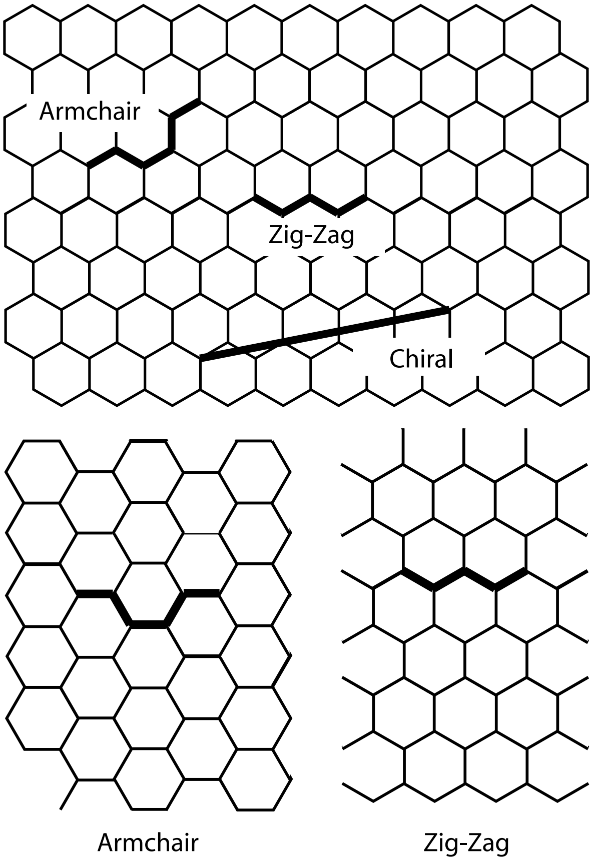
Many other types of composite materials are now being developed aiming towards multifunctional composites as demands are increasing for smaller, lightweight but smarter products. Reinforcing components could be particles (0-dimensional), fibers (1-dimensional) or plates (2-dimensional). The orientation of these reinforcing components in a matrix could be random, unidirectional or bidirectional, and could be laminar. They are in the form of distinct phases in the matrix with sizes varying from macrophases to nanophases. When the size of the reinforcement is in the nano-range, the composite is referred to as a nanocomposite. When the nanoreinforcements are composed of a functional nanophase the composite is then a multifunctional nanocomposite in which the nanophase provides an advanced functional behavior through enhanced properties such as mechanical, chemical, biological, electrical, magnetic, optical properties.

A Short Note on Carbon

Carbon is a critical, enabling material and consequently an important component of many multifunctional composites. We will, therefore, outline some of the key properties for this element.

Carbon is one on the most common elements found on earth. It appears in three major forms: diamond, graphite, and as various forms of "fullerenes" and "carbon nanotubes." The two former versions occur naturally, whereas the latter is typically artificially produced. Diamond, with a cubic crystal structure where the atoms are arranged in a tetrahedral configuration, is the hardest material found in nature. It is an electric insulator. In graphite, the atoms are arranged in hexagonal planes that are stacked parallel to each other. This layered nature of the structure makes graphite anisotropic and relatively soft, since fracture easily occurs along these layers. In this form, carbon is electrically conductive.

A third form of carbon was categorized in 1985 [54] (even though earlier work indicated their existence) and will play a key role in functionalizing composite materials. This third form is a variant of graphite, where one layer of graphite is wrapped to form a sphere or cylinder. The initial form described was a 60-carbon atom sphere, where the carbon atoms are arranged in 20 hexagons and 12 pentagons on the surface of the sphere. Soon after this discovery, variants of the 60-atom sphere were found, where, in particular, the tubular form, carbon nanotube (CNT) have provided an exciting new research field [50]. There are endless variants of how CNTs can be assembled, but the major categories can be summarized as single- and multi-wall tubes along with their major division by the atomic structure (armchair or zigzag; if arbitrary direction "chiral" see Fig. 1). CNTs are one of the most promising materials for functionalizing composites, due to their unique properties. For example, the armchair styled CNT are metallic, but the other versions are semiconductors. In particular, single CNTs are excellent conductor, and can therefore serve as nano-sized electrical wires. CNTs also have excellent mechanical properties, for example as high as 1 TPa elastic modulus has been reported [48,60]. Multiwalled CNTs consist of either single walled tubes layered



Composites, Multifunctional, Figure 1
Illustration of the various atomic configurations a carbon nanotube can assume. A carbon atom is located in each of the corners of the hexagons

over each other ("Russian doll-style"), or as a continuous, rolled sheet forming a wrapped layered sheet (parchment-style). There are several interesting review papers discussing CNT tubes, for example by Thostenson et al. [106] focusing on mechanical properties, by Gooding [43] focusing on electrochemistry and by Bandaru [5], discussing electrical properties of CNTs. Two recent review papers on theoretical aspects of the thermomechanical properties of carbon nanotubes can be found in [20,21].

Functionalized Carbon Reinforced Polymer Matrix Composites

Carbon reinforced, polymer matrix composite are now commonly used in engineering applications, due to their relatively high strength and low weight. The carbon fibers give the composite their strength and stiffness, for which the structures traditionally have been optimized. However, recent developments make it possible to utilize the ability of carbon fibers to be electrically conductive. This opens up a broad range of opportunities for functionalizing carbon reinforced polymer matrix composites.

Electromagnetic Shielding

Electromagnetic interference (EMI) is caused by electromagnetic radiations emitted from sources that carry electrical current. A common source for EMI is an electronic device, but natural sources such as the Northern Lights and the Sun can also cause EMI. EMI results in anything from annoying "sound sparkles" on the television or cell phone to a malfunction of a device or a system (e. g., an aircraft). In warfare, EMI can be used to disrupt communications by the hostile side. Thus, it is important to shield aircrafts and communication devices from EMI. Carbon reinforced polymer matrix composites can readily be designed to shield EMI. Since carbon is electrically conducting, the carbon fibers can be used to reflect the radiations. By adding conductive fillers into the polymer matrix, the composite material can work efficiently as an EMI shield. It is now commonly used in many consumer products to shield them from EMI, both for protecting the machine and from spreading EMI the machine is generating. For example Bagwell et al. [4] added short copper fibers to increase EMI. A review of EMI shielding can be found in reference [27] and a review of conducting polymer composites that are primarily optimized for EMI shield is found in [101].

Electrified Carbon Fiber Polymer Matrix Composites

Interestingly, research from the groups of Sierakowski and Zhupanska have shown that the material properties in a carbon fiber polymer matrix composite may change when subjected to an electromechanical field [97,98,121]. Early work [98] indicated that the strength, in particular the resistance to fracture and delamination, increases when the composite is subjected to an electric current. Several factors contributes to this, including that (i) the mechanical and electromagnetic fields are coupled when mechanical and electromagnetic loads are imposed simultaneously; (ii) the heat generated in the conducting carbon fibers are transferred to the polymer matrix; and possibly that (iii) the failure mechanisms change when the structure is subjected to an electromagnetic field. Recent work [97,121] where the impact resistance was investigated, show that the gains are short-term. The impact resistance initially can increase as much as 30%, but for a structure subjected to long-term exposure to an electromagnetic field, the gain is reduced back to the initial properties. The losses appear to be caused by the increasing temperature of the polymer due to the heating of the carbon fibers. Nevertheless, a structure can temporarily be strengthened by imposing an electromagnetic field. Moreover, the heat generated in the carbon fibers could potentially be used to activate self healing mechanisms (self healing mechanisms are discussed in Sect. "Self-healing Composites")

Functionalized Composites with Carbon Nanotubes

As discussed in the introductory section of carbon, carbon nanotubes are probably one of the single most promising materials to functionalize composites. The possibilities appear to be endless and a few limited examples will be discussed here.

Single-walled carbon nanotubes (SWNTs) have great promise for functionalizing composite materials. They are light weight and have high mechanical strength, high thermal and electric conductivity and unique optoelectronic properties. They are also light weight, with a small diameter and high aspect ratio. However, these properties may be compromised when incorporated into a polymer matrix. This is primarily caused by the SWNT not being compatible with the polymer matrixes. This results in the SWNTs tending to agglomerate into clusters. When the SWNT are not bonded properly to the polymer matrix and/or appear in clusters, their unique properties may not be transferred to the composite materials. Therefore, significant efforts are being aimed towards improving the dispersion and bonding.

Chen et al. [19] suggest that molecular engineering is a viable approach to achieve good mechanical strength and retain the electric conductivity of CNT. They pointed out that the problem with dispersion and bonding is due to the smooth surface of the SWNT. In a mix of SWNT and a polymer matrix, the mechanical load can be transferred though mechanical interactions between the SWNT and the polymer matrix, via van der Waals' interactions or covalent bonds, or via special non-covalent bonds, such as hydrogen bonding. The covalent bonding is in general the strongest type of bonding. However, when this is implemented, the electrical and thermal properties are often seriously challenged since these bonds tend to interfere with the SWNT structure [19].

CNTs can be used for energy absorption, which have been shown by Chen and co-workers [22,45,86]. In their work, they developed a solid-liquid composite which combines a non-wetting liquid with a hydrophobic nanoporous solid. The basic premise is that a liquid is absorbed into nanopores (such as the inside of a CNT) when a pressure is applied on the system. They showed that this infiltration absorbs and converts mechanical work into solid-liquid interface energy, with high energy absorption (10–100 J/g). Due to the ultra-high specific surface area of the nanopores, this is several orders of magnitude higher than conventional energy absorption materials. Moreover, by varying the interface energy, the energy absorption performance may be adjusted in a wide range, suitable for damping protections, vibration proof, or blast resistance. The interfacial energy can be changed by using chemical admixtures, or using viscous liquid. If the load rate can be controlled, this can also change the interfacial energy. In addition, by using functional liquids (such as electrolytes), the ion density at the nanopore-liquid interface may be perturbed by external mechanical or thermal fields. Thus, the multifunctional solid-liquid nanocomposite may harvest thermal and mechanical energies into electricity [46,87].

Polyaniline (PANi) is formed by polymerizing aniline (phenylamine, aminobenzene) which is an aromatic amine with the formula $C_6H_5NH_2$. PANi is a conductive polymer, and consequently has great potential to be a useful material component in multifunctional composites. When aniline is polymerized with the presence of multi-walled carbon nanotubes (MWNT), to form a polyaniline-carbon nanotube composite, the MWNT are coated with PANi and form a three-dimensional network within a matrix of PANi. This results in a composite with excellent electro-optical properties [90]. Preliminary work has shown that PANi can also work as a biological sensor [61]. Here the SWNT was wrapped with a single-stranded DNA and mixed with a self-doped polyaniline, poly(anilineboronic acid). The composite is able to detect nanomolar concentration of dopamine (a naturally occurring hormone). The sensitivity for detecting dopamine was increased with a factor of four by adding the SWNT compared to the self-doped polyaniline [61].

Polypyrrole (PPy) is formed from synthesized (connected) pyrrole, where pyrrole is a heterocyclic aromatic organic compound, C_4H_4NH. PPy have been used for corrosion protection of metals, discussed in Sect. "Multifunctional Coatings". In a similar manner as PANi, PPy is conducting. An interesting application is to use PPy with MWNT, where supercapacitive properties have been measured [49]. To achieve this, the MWNTs must be aligned and then coated by an appropriate layer of PPy. Alignment of the MWNTs can be obtained by growing them on a quartz glass under appropriate conditions, described for example by Hughes et al. [49]. Measurement of the charge storage capacity of the aligned-MWNT-PPy composite film show several times charge storage than either PPy or MWNT alone (e. g., 2.55 F/cm^2 for the composite film compared to 0.62 F/cm^2 for pure PPy film) [49]. Thus, aligned MWNTs coated with PPy have potential applications for supercapacitors and batteries, as well as sensors.

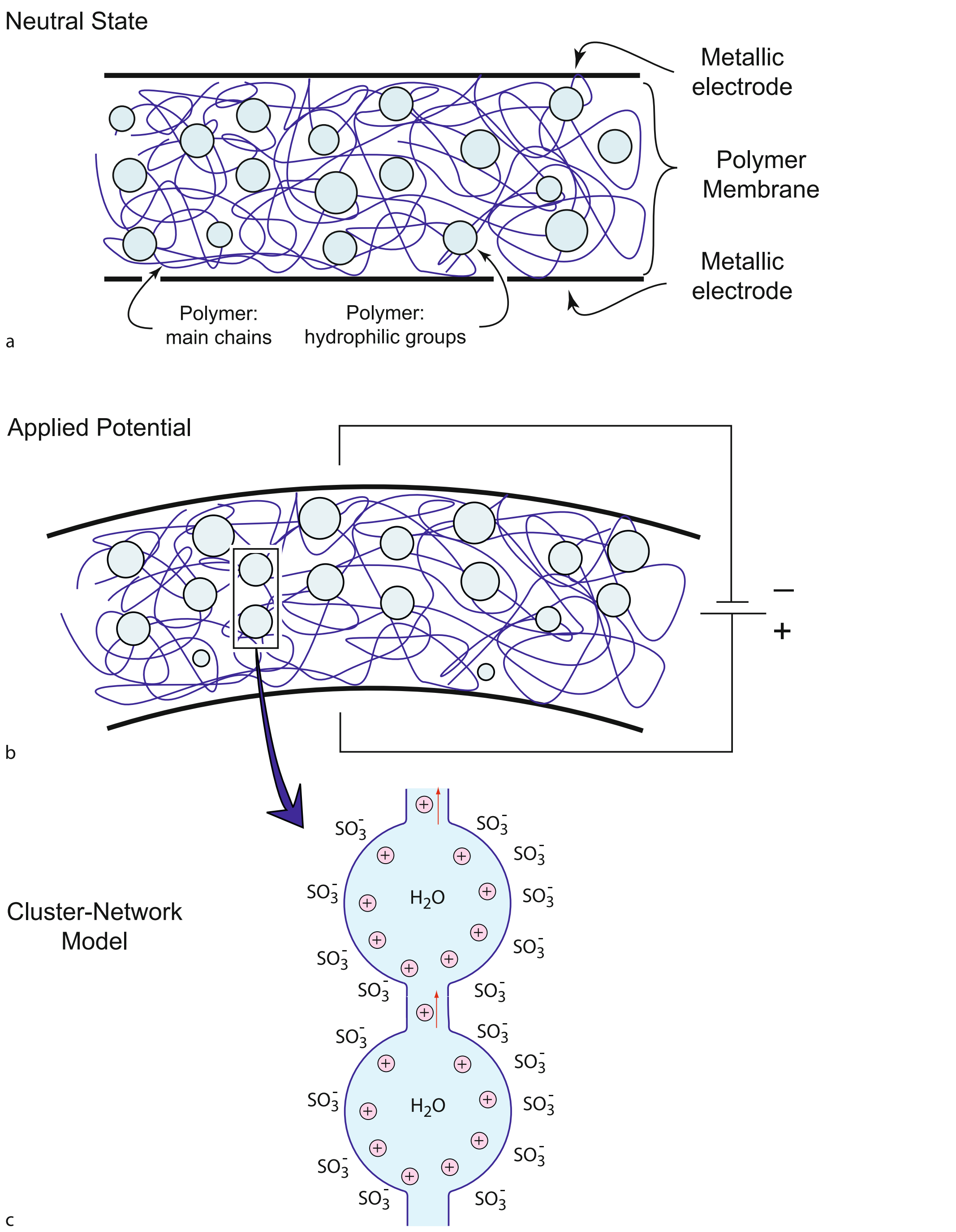
Composites as Actuators and Sensors

Ionic polymers have received significant attention during the last decade, based on their ability to work as electromechanical transducer and potentials to work as sensors and/or actuators, and also as electrolytes in polymer fuel cells. Today, the preferred polymer materials are perfluorosulfonic acid (PFSA) polymers, a class of fluoropolymers consisting of a hydrophobic polytetrafluoroethylene (PTFE) backbone attached to hydrophilic sulfonic acid groups (SO_3^-) or carboxylate groups via the fluorocarbon polymer side chains. Upon water uptake, the hydrophobic groups ionize and attach to the hydronium ions (H_3O^+) to provide a conductive path for proton transport, while the polymer network maintains the overall structure of the membrane. Commercial material products include Nafion®[1] membranes and Flemion®[2] membranes. These membranes are also used as the proton exchange membranes (PEM) for fuel cell applications. A PEM functions as a "filter" (an electrolyte), letting protons through the polymer membrane, whereas the electrons are forced to take a path outside the PEM. From the path of the electrons, electric energy can be harvested.

Actuators and sensors can be made by an ionic polymer metal composite (IPMC) [3,8,9,56,57,74,75,76,89],

[1] Nafion® is a registered trademark of E.I. DuPont De Nemours & Co.

[2] Flemion® is a registered trademark of Asahi Glass Group in Japan.



Composites, Multifunctional, Figure 2

A schematic of the principle of an IPMC actuator. a Geometry before a voltage is applied highlighting the morphology of the polymer chains (hydrophobic) and the hydrophilic groups forming a cluster. b The structure deforms when an external voltage is applied. c The cluster-network model for Nafion® membranes, illustrating the cation migration. Adapted from [47,67,76,95]

with a potential application as artificial muscles [3,8,9,10, 56,57,67,76,94,95,112,114]. In this case, the ionic polymer is sandwiched between two metallic electrodes. The ionic polymer is hydrated, typically with water or ionic liquids such as salts [3], to achieve its functionality. The metal electrodes are flexible layers, resulting in a soft and flexible actuator which can perform large dynamic deformation if placed in an alternating electric field, see Fig. 2.

Currently, platinum (Pt) and gold (Au) are the preferred electrode materials, which defuses into the ionic polymer membrane, resulting in a material gradient over the thickness of the IPMC. The actuation is governed by the mobile cations (positively charged) moving towards the fixed cathode (negatively charged), resulting in a biased morphology and consequently a bending of the membrane as indicated in Fig. 2. When the current is switched, the location of the cathode is reversed and the cations will consequently move towards the other side, causing the actuator to move in the reverse. Alternatively, the IPMC can work as a sensor, where it generates a voltage if it is suddenly bent.

Nafion® polymer based IPMCs relaxes (reduces its deflection), whereas Flemion® polymer based IPMCs slowly increases its deflection under constant voltage. This is attributed to the mobile cations initially repelling the sulfonic acid groups in Nafion® polymer (which gives a fast actuation), but when the polarization is held constant, the cations relocates slowly, relaxing the IPMC. The carboxylate in Flemion® polymers are weaker in polarizing the structure, and therefore the relaxation is not observed [75,76].

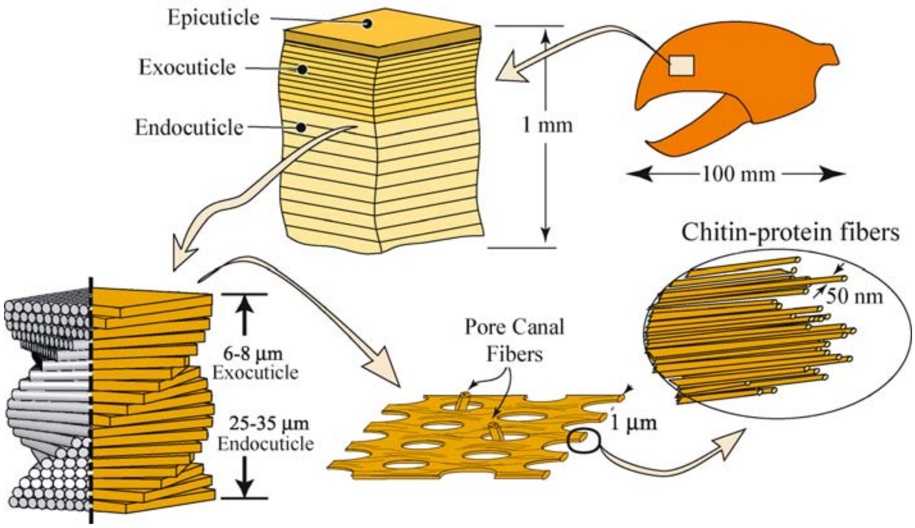
These actuators can strain up to 3% for voltages less than 7 V [67]. This induces a significant bending, where stresses up to 30 MPa are reported [67]. Future research efforts are focused on reducing the negative effect of dehydration (an ionizing liquid is needed for the function), as well as addressing the reduced efficiency over time. Current applications range from fins to robotic fish to artificial eyes [67].

Biomineral Composites

Nature has presented us with a variety of biological composites (biocomposites) such as skin, bone, shell, leaf and wood, which are all multifunctional in living organisms. Some are organic in nature, e. g. leaf and wood, whereas some are organic-inorganic, e. g. bone. Attempting to mimic nature, the design of these biocomposites are now intensely studied, including the structure and microstructure and the physical, chemical, electrical, magnetic and even optical properties. For example, using a biomimetics process M. C. Chang and co-workers have been able to develop a synthetic bone [17,18].

Biocomposites formed through a biomineralization process are referred to as biomineral composites. Examples of biomineral composites include bone, dentine, enamel, mollusk shells, crustacean exoskeletons, eggshells, sponge silica skeletons, and a variety of transition metal minerals produced by different bacteria (see references in references [42] and [33]). The functions of the biomineral composites include structural support, mechanical protection and movement, anchoring (to another body or to ocean floor) grinding, filtering, gravity or magnetic field sensing, optical and piezoelectric [33]. Biomineral composites are composed of an organic matrix of proteins, lipids and polysaccharides. The structure consist of a nano- or micro-scale amorphous or crystalline minerals formed by a biologically induced or controlled mineralization processes, through complex chemical interactions between organic and inorganic matrices [2,42,91,116]. The structure is usually complex with the organic and the mineral components tightly interwoven at the nanoscale level, highly ordered and hierarchical to give high strength, rigidity along with mechanical and chemical stability, that are superior to synthetic materials made from the same materials. Biomineral composites are normally designed to function under a narrow range of environmental conditions, such as narrow temperature regimes and restricted mechanical loads. However, the compositions of biocomposites can be altered over time when a gradual change in environmental conditions occurs, to achieve necessary properties for the survival of the organisms [63]. Mollusks and sponges are known to make use of sophisticated biomineralization mechanisms to obtain structures that exhibit attractive combinations of strength, stiffness, resilience, and energy absorbing capabilities [63]. Even though the mechanisms of biomineralization are yet to be fully understood [69], biomineral composites are of much interest to warrant in-depth studies by biologists and chemists as well as material scientists. The structures of a few biomineral composites that are multifunctional are briefly described below.

The crustacean exoskeleton is a layered structure made up of the epicuticle layer, which is the topmost layer, the exocuticle layer and the endocuticle layer, which is the innermost [25]. The epicuticule is a relatively thin layer of about 2–4 μm. It is waxy, acting as a diffusion barrier [77]. The exocuticle is about 150–180 μm in *Homarus americanus* (American lobster) claw with the endocuticle 3–4 times this thickness. As a comparison, for the *Callinected sapidus* (Atlantic blue crab) claw, the exocuticle is about 40–50 μm and the endocuticle 6–8 times thicker [25]. The exocuticle and the endocuticle are the major load bearing structures of the exoskeleton and are made up of multiple fibrous layers arranged parallel to the surface. The fibrous layer consists of chitin-protein (chitin is a biological polysaccharide with the generic formula $(C_8H_{13}O_5N)_n$) fibrils bonded by a matrix of minerals and other proteins. Each of these fibrous layers is rotated by a small angle relative to the next layer in parallel, building up to a band of



Composites, Multifunctional, Figure 3

The cross-section (*top*) of the exoskeleton of a Homarus americanus (American lobster) taken from the claw (*top right*). The outer layer (epicuticle) acts as a diffusion barrier while the exocuticle and endocuticle layers are load bearing structures built of mineralized fibrous chitin protein. The helicoidal nature of the arrangement of the fibrous layers of the exocuticle and the endocuticle layers is shown on the *left*. The pore canals and the pore canal fibers in a layer are shown (*bottom*). Each layer is composed of chitin protein fibers (*bottom right*). Note that the exocuticle layer is denser than the endocuticle layer

layers that is twisted by 180° to form a helicoidal architecture (Fig. 3). The exoskeleton has through-the-thickness holes, *pore canals*, through which chitin-protein macrofibrils fibers, *pore canal fibers* run perpendicularly to the layers. Even though some disagreement exists in the literature, the pore canal fibers appear to run from the bottom of the endocuticle to the top of the exocuticle [25,28]. The pore canals and the pore canal fibers fill important functions in building the exoskeleton after molting. Moreover, Cheng et al. [25] showed that the pore canal fibers are important for strength of the exoskeleton. In all, the multiscaled structure of the exoskeleton, a biomineral composite, provide the crustacean with a strong structural support, an impervious defense covering for the body of the crustacean, and also serve as a carrying, holding and tearing tool in case of attack or feeding.

Nacre is another natural biomineral nanocomposite; it is also known as mother-of-pearl. It is the iridescent lining on the inside of the shells of many sea-going bivalves and gastropods such as oysters, mussels and abalones. Like many other biomineral composites, nacre has a hierarchical structure. It is composed of about 95% inorganic hexagonal platelets of aragonite (a crystallographic form of $CaCO_3$) 5 to 8 μm wide and 0.2 to 0.5 μm thick [6], arranged in a continuous parallel lamina in 5% organic matrix composed of elastic biopolymers (such as chitin, lustrin and silk-like proteins and polysaccharides). The organic biopolymer is typically 5 to 20 nm thick. Nacre has received significant attention in recent years due to its high ductility, enhanced toughness and fracture strength, along with its low weight, resulting in excellent specific properties. Its fracture resistance is about 1000 to 3000 times greater than that of its component aragonite crystals [30,31]. Its high toughness is as a result of the ductility of the organic matrix in connection with the repeated unfolding of the protein molecules. The nanostructure resembles a brickwork arrangement with a significant overlap of the platelets and the organic matrix serving as the mortal [15]. This architecture is a critical factor that is responsible for the high fracture strength observed in nacre [6,41]. In addition, Li et al. [59] showed that the rotation of the nano-sized grain during loading is a key contributor to the high ductility. Many studies have been carried out on nacre using various experimental and modeling techniques to study its formation, its structure and morphology, and its deformation and properties, especially the mechanical properties [6,16,52,55,62,63,65,73,85].

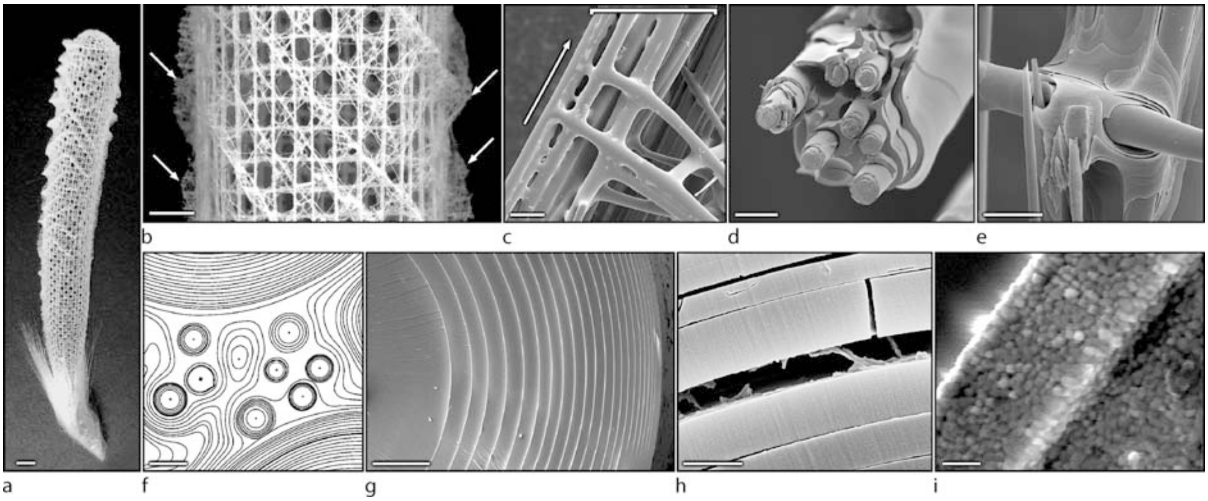
Mimicking this material is of interest in the design of high performance materials such as impact resistance armor [7,55,92]. Using layer by layer assembly technique, Podsiadlo et al. [84] have been able to prepare a nanostructured analogue of nacre from nanometer sized sheets of montmorillonite clay and a polyelectrolyte. Artificial nacre has also been synthesized [103]. Mimicking nacre in creating one nano-layer of material at a time, Nicholas Kotov and his team have evolved a process that allows the creation of materials one nano-layer at a time. They use this process to produce a new material from clay nanosheets and a water-soluble polymer that shares chemistry with white glue. The material is transparent, very strong, yet lighter in weight [64].

Animal skeletons are made up of bones which are hard and rigid tissues. Bone is self-healing and can continuously regenerate itself. Moreover, bone is relatively stiff and tough, and can withstand and adapt over time to local stresses. These properties make bone a reliable biological structural material. Bone is considered as a nanocomposite of minerals and proteins [18,26]: It is composed of a matrix impregnated with calcium carbonate, calcium sulphate and small amounts of sodium and magnesium. The matrix, consist of collagen fibers impregnated with crystals of hydroxyapatite, $Ca_5(PO_4)_3(OH)$, and water. Among the functions of the bone is support, locomotion, protection for soft and delicate organs (like the skull protecting the brain), manufacturing of blood cells and homeostasis. An example of bone is the femur; it has a covering of a tough, strong membrane, called periosteum which is richly supplied with blood vessels. Next to the periosteum is a layer of compact bone and bone forming cells (osteoblasts), arranged in concentric layers (lamellae) with round small interconnecting canals (the Haversian canals) that contain blood vessels, nerves and lymph vessels. Embedded in this hard bone matrix are osteocytes, which are associated with bone deposition and bone remodeling. Inside the compact bone is a thin soft membrane known as the endosteum that encloses the marrow cavity that contains soft tissues, the yellow marrow [88,108]. The growth and strengthening of bone is stimulated by mechanical stresses through strain detection. This function is carried out by specialized cells within the bone that are sensitive to and respond to strains. In the absence of mechanical stresses bone becomes weak and less developed. Thus, the more the bone performs its intended functions the stronger it becomes. Mimicking this ability will produce excellent synthetic engineering materials for structural or load-bearing applications.

In a study of the sponge *Euplectella sp.*, Aizenberg et al. [2] described the structure of its skeleton as being hierarchical in nature. It has a layered arrangement that gives it a high resistance to crack propagation. The skeleton is thus composed of a layered biocomposite material. The structure is described in Fig. 4, after Aizenberg et al. [2]. The microstructure is made up of consolidated hydrated silica nanoparticles forming sets of concentric rings glued together by an organic matrix to form spicules. This layer approach provides toughness and resistance to crack propagation. These spicules are then assembled in parallel into bundles within a silica matrix to form struts. The struts are arranged to make the cylindrical cage with the ability to resist tensile and shearing stresses and a significant capacity for recovery after deformation of the skeleton.

Aizenberg et al. [1] showed that the tunic spicules of the ascidian *P. pachydermatina* are a biocomposite material with well-defined domains of both amorphous and crystalline calcium carbonates, separated by an insoluble organic layer. The crystalline calcium carbonate is polycrystalline calcite, and it forms around the amorphous calcium carbonate which serves as the core. The calcium carbonate layers contain magnesium and proteins with a higher content in the amorphous than in the crystalline. The amino acid compositions of macromolecules associated with the two mineral phases are also quite different.

Wood is a natural biocomposite material with a multifunctional capability. It differs from the above mentioned bicomposites in that it is *not* a biomineral composite. Wood is a naturally hard and tough biocomposite material that forms the trunk or stem of trees. The material consists essentially of elongated hollow cells that carry nutrients from the roots to the leaves. The cells make up the cellulose fibers that are arranged generally in the grain direction, parallel to the surface of the trunk. Cellulose, with a generic chemical formular $(C_6H_{10}O_5)_n$, is one of the major chemical constituents of wood, others are hemicellulose and lignin. It is a linear polymer with thousands of mers in a single molecule and it constitutes about 40 to 50% of wood. The cross section of wood is composed of several layers: the outer bark, the inner bark, the cambium, the sapwood, the heartwood, and the pitch. The thickness of each layer depends on the age of the tree, the species to which it belongs, and on the particular tree. The cambium layer is microscopically thin and it grows by cell division to increase the diameter of the trunk. The tree trunk increases in diameter by addition of new peripheral growth layers that constitute the growth rings. The sapwood layer conducts moisture, minerals, oxygen, and nitrogen. As the stem (or trunk) grows in diameter, the sapwood progressively forms the heartwood. The heartwood is the thickest of the layers and it is the one that provides the structural strength. It is usually darker in color because of the mineral deposits, gums and resins that are present in it. Cutting across these layers horizontally are tissues called wood rays radiating out from the center outward. They help in storing and transferring nutrients. Botanically, woods are classified as softwoods and hardwoods depending on their basic cellular structure and on how moisture moves within



Composites, Multifunctional, Figure 4

Structural analysis of the mineralized skeletal system of Euplectella sp. a Photograph of the entire skeleton, showing cylindrical glass cage. Scale bar: 1 cm. b Fragment of the cage structure showing the square-grid lattice of vertical and horizontal struts with diagonal elements arranged in a chessboard manner. Orthogonal ridges on the cylinder surface are indicated by *arrows*. Scale bar: 5 mm. c Scanning electron micrograph (SEM) showing that each strut (enclosed by a *bracket*) is composed of bundled multiple spicules (the *arrow* indicates the long axis of the skeletal lattice). Scale bar: 100 µm. d SEM of a fractured and partially HF-etched single beam revealing its ceramic fiber-composite structure. Scale bar: 20 µm. e SEM of the HF-etched junction area showing that the lattice is cemented with laminated silica layers. Scale bar: 25 µm. f Contrast-enhanced SEM image of a cross section through one of the spicular struts, revealing that they are composed of a wide range of different-sized spicules surrounded by a laminated silica matrix. Scale bar: 10 µm. g SEM of a cross section through a typical spicule in a strut, showing its characteristic laminated architecture. Scale bar: 5 µm. h SEM of a fractured spicule, revealing an organic interlayer. Scale bar: 1 µm. i Bleaching of biosilica surface revealing its consolidated nanoparticulate nature. Scale bar: 500 nm. (Taken from reference [2] with permission from AAAS)

the living tree. Softwoods are mainly made up of long cells of between 3 and 5 mm called tracheids. Hardwoods, on the other hand, are mainly made up of two kinds of cells, wood fibers (0.7 to 3 mm long) and vessel elements (with wide ranging lengths).

The important physical properties of wood are moisture content, permeability, shrinkage, density. These give it the multifunctional capability such as serving as a super-structure, acting as a nutrient storage and transport medium, the ability to withstand harsh weather, and self-healing. The properties, however, vary greatly across species and also depend on factors such as the age of the tree, stem form, type of soil and climate. Wood is anisotropic with the mechanical properties varying across the growth rings and along the height up the tree. The mechanical, electrical and thermal characteristics of wood make it a popular excellent engineering material over ages.

Leaf is an organic biocomposite that is flat, broad and thin. It is a plant organ in which photosynthesis is carried out. The upper surface of the leaf is waxy for the purpose of water-proofing. It performs functions such as converting sunlight to chemical energy in the mesophyll, transporting glucose, water and minerals through out the plant by the vascular bundle; it is water-proof and provides shade for the tree. The cross-section is made up of different layers in this order from the top: upper cuticle, upper epidermis, palisade mesophyll, spongy mesophyll, lower epidermis and lower cuticle. Embedded in the mesophyll layers is the vascular bundle (phloem and xylem) and air spaces for the supply of air (carbon dioxide) and moisture that comes in through the stomata that dotted the lower epidermis through the lower cuticle. The broadness of leaf allows it to gather as much sunlight as possible as a supply of the energy needed for photosynthesis. Leaf provides a system that could be mimicked in designing materials for energy conversion and at the same time distributes the product.

Biocomposites materials produced by nature have properties that could be beneficial when reproduced in synthetic materials. The design, manufacture from simple raw materials, economical use of raw materials and energy, multifunctionality and degradability of biocomposites are inspirational to biomimetics or biomimicry in the design and manufacture of synthetic engineering materials. Another inspiration from nature is the building from bottom up, from atomic or molecular level to the macro structural level. This provides for efficient use of raw materials. A thorough study of nature's biocomposite materials could, therefore, yield viable procedures and techniques for the design and manufacture of synthetic engineering materials with excellent combination of properties that will provide for multifunctionality in them.

Self-healing Composites

Biological systems have an outstanding ability in self-healing; that is, automatically detecting and repairing damaged tissue. The repair is made by a material similar to the original (causing a scar tissue), or identical tissue (leaving the damage area undetectable after repair). For humans, the latter can for example be observed in bone, whereas the former on skin. Moreover, biological systems can adapt to new conditions. Humans build more muscles and bones if we increase our daily exercise regime and a tree grows branches to find the most sunlight.

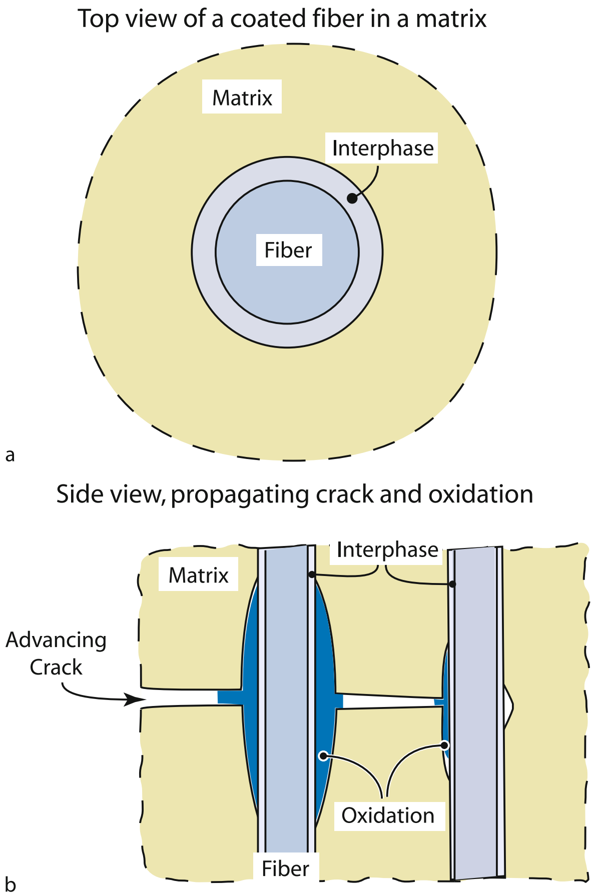
Man-made structures traditionally do not have these features. If damage occurs, damage will continue to accumulate until it is detected via human intervention or until the structures fails. Today, there are many sophisticated means of detecting failures. Even though tremendous advances have been made towards detecting and repairing damage in structures, *self-healing materials* would significantly improve the reliability of structures. During the last decade, significant advances have been made towards developing self-healing composites. The current approaches results in a "scar tissue" in the sense that the original material is not exactly reproduced, but that the structure will function satisfactory. Two materials systems will be discussed here; (i) ceramic matrix composites and (ii) polymer matrix composites.

Ceramic materials typically have high strength and stiffness, and retain their strength even at high temperatures. Since they are brittle and therefore are considered unreliable (tend to break without warning signs that metals exhibit, such as plastic deformations, especially under tensile stresses) their uses have been limited. To negotiate the brittle response, ceramics can be reinforced. Most commonly, ceramics are reinforced with a second ceramic, forming a ceramic matrix composite (CMC). The bonding between the reinforcement and the matrix is a key parameter that governs the toughness of the CMC. By controlling the interface material, referred to as "interphase", the interfacial bonding can be optimized, allowing for the load transfer between the matrix and the fibers, deflects matrix cracking and serve as a barrier towards diffusion. When considering the strength of the interface material, it must be optimized to be "just right," not too strong and not too weak.

Self-healing of CMCs can be achived through high temperature oxidation in silicon carbide based CMCs. Silicon carbide (SiC) reinforced with SiC fiber (SiC/SiC composites) or carbon fibers (SiC/C composites) is mechanically durable up to 1500°C and is therefore considered a promising composite for high temperature applications such as engines and gas turbines [70]. Even though the fibers are brittle in a similar manner as the matrix, the fibers and matrix work synergistically to create a ductile material. Micro-cracks develop during loading, but these micro-cracks are stopped by the microstructural features, simulating a non-linear response paralleling yielding in metals. When silicon is subjected to elevated temperatures, it quickly forms a protective coating of silica (SiO_2). Unfortunately, the micro-cracks that are associated with the CMC become pathways for oxygen to penetrate into the structure, causing internal oxidation. This can weaken the structure. By coating the fibers with carefully selected materials, the oxidation can be controlled and result in self-healing of the micro-cracks. Graphitic carbons ("pyrocarbons," "PyC") and boron nitride (BN) have emerged as the most prominent interphase [39,44,70,71,72]. When oxygen diffuses through the micro-cracks, a fluid oxide is formed due to the oxidation, filling the cracks, Fig. 5. These glassy oxides that form can be optimized through the interphase. For example, when a borosilica glass is used as an additional coating on the fibers, no loss in composite strength was observed after 200 h at elevated temperatures [39]. Thus, a self-healing mechanism in silicon carbide (SiC) matrix composite reinforced with SiC or carbon fibers has been observed, which is caused by oxidation at high temperatures. The oxidation occurs at temperatures above 800°C. The self-healing can continue until the reducing material has been consumed.

Polymer composites might be the most promising systems where self-healing mechanisms can be developed. There are several reasons for this. Polymer based systems are in general less expensive than ceramic based systems, and tends to be easier to work with. All self-healing ceramic systems are based on activating the healing process through subjecting the material structure to heat. Even though some self-healing polymer systems are based on heating, the temperature regime for healing polymer is significantly lower than for ceramics, thus simplifying the process. Furthermore, many self-healing approaches to polymer systems are not dependent on heating. Lastly, several different approaches for self-healing of polymer have been developed so far, thus inviting alternative approaches for self-healing.

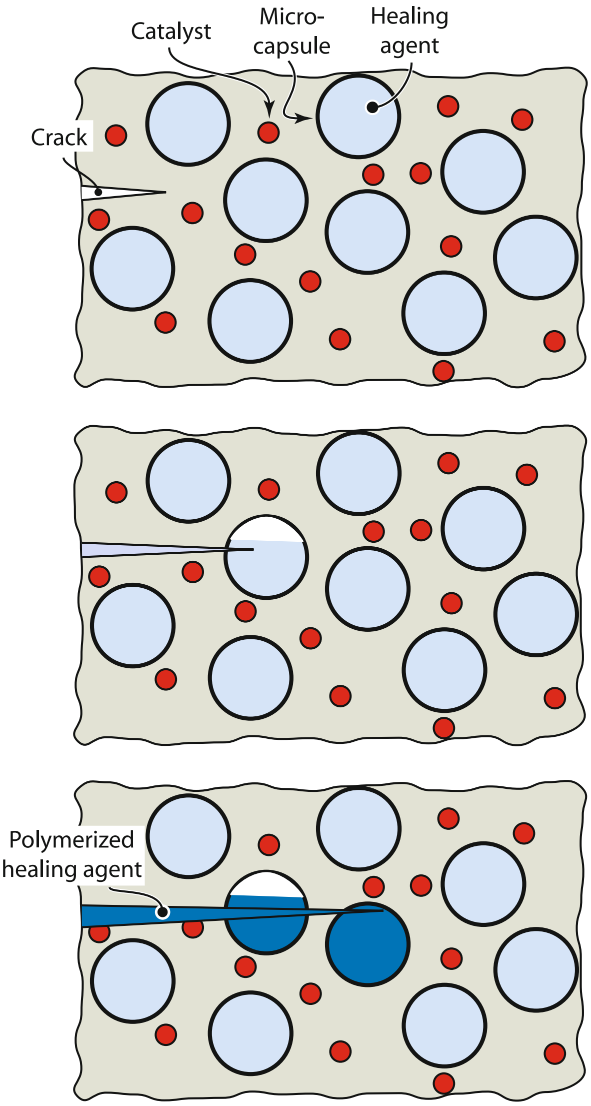
One simple concept of healing a damaged structure is to subject the material to elevated temperatures as was


Top view of a coated fiber in a matrix
Matrix
Interphase
Fiber
a
Side view, propagating crack and oxidation
Matrix
Interphase
Advancing Crack
Oxidation
Fiber
b

Composites, Multifunctional, Figure 5
Examples of interphases used in SiC matrix composites a schematic top view of a fiber with a coating (interphase) in a matrix b schematic side view of an advancing crack exposing the interphase that oxidizes, resulting in self-healing

seen for ceramics. This idea is based on that the elevated temperature activates a chemical process that results in healing. For example, by utilizing a particular thermally reversible reaction (e. g., a selected Diels–Alder reaction) for cross linking linear polymer chains, a fractured material can be healed, as demonstrated by Chen and coworkers [23,24]. For one material system investigated, after fracture and subsequent heat treatment, the original strength was regained. In this case, the polymer (a furan-based monomer synthesized with 1.8-bis(maleimido)-3.6-dioxaoctane) was subjected to mechanical testing leading to fracture followed by heat treatment at 130°C for 30 min. In this case, there was no loss in the original strength when the structure was re-tested [24]. This approach to self-healing has the clear advantage that no additional material is needed to be added: the material is intrinsically self-healing. However, only a very limited set of materials that can self-heal by cross-linking the polymer chains upon reheating. An alternative approach for healing via heating utilizes an additional material phase incorporated into the original material. When subjected to sufficient heating, this additional material is activated, and can mend the damage structure. For example, a composite made of glass fiber reinforced epoxy retained its stiffness after the heated repair [120]. By adding a heat-activated material to the composite structure, a design engineer would not be strictly limited to a narrow set of materials. However, a major disadvantage with both of these heat activated healing methods is that the healing is not automatic; rather, the structure needs to be treated in a separate process. (This may be differentiated from the ceramic self-healing that was discussed above. The ceramic is operating at the temperature at which oxidation occurs, but this polymer operates at temperature lower than where self-healing appears.) Nevertheless, depending on the application, this approach can be quite useful.

A more convenient approach to self-healing of a structure compared to the heat activated systems described above is a system that heals itself without active interference. Most approaches aiming to achieve this are based on introducing one or more phases into the composite material. These additional material phases are automatically activated when damage occurs. A successful approach have been developed by White, Sottos and co-workers, where spherical microcapsules containing a "healing agent" and a second phase containing a catalyst are embedded in a polymer matrix composite [12,13,14,53,115]. When a crack propagates, the micro capsules in the crack path burst and release their healing agent into the crack, Fig. 6. As the healing agent fills the crack, it will eventually contact the catalysts. When this occurs, the healing agent will polymerize, filling the crack and effectively healing the crack. In the work by White, Sottos and co-workers, the agent was a dicyclopentadiene (DCPD) monomer and the catalyst a bis(tricyclohexylphoshine)benzylidine rethenum (IV) dichloride (a Grubbs' catalyst) [12,13, 14,53,115]. This results in a ring-opening metathesis polymerization (ROMP) of the DCPD, resulting in a highly cross-linked polymer. When stabilized with 100–200 ppm p-tert-butylcatechol, the DCPD has a long life and healing can be achieved even for aging structures. The micro capsules containing the DCPD ranged from 40–240 μm and were made with poly-ureaformaldehyde. In early work, curing for 48 h was required to retain 45% of the initial strength (if the curing occurred at 80°C, up to 80% of initial strength was achieved) [12,115]. Subsequent studies showed that 10 h were sufficient to achieve full polymer-



Composites, Multifunctional, Figure 6
Schematic of a propagating crack in a polymer with micro-capsules filled with a healing agent

ization (full healing or full strength), and that the fatigue life can be increased with over 200% if the structure is allowed to rest sufficiently for the healing agent to polymerize [13,14]. Most engineering structures are allowed to "rest" between operations. For example, cars are normally used for commuting to work and get sufficient time to "rest," both day and night, whereas airplanes are scheduled for regular maintenance that keeps them grounded for many hours that may be sufficient for the polymerization to take place. Thus, this system is a promising approach towards extending the lifetime of polymer matrix composites.

The micro-sphere approach has the clear advantage of being possible to be incorporated into a range of materials, and that no particular treatment is needed to activate the healing processes. There are, however, some drawbacks. These include that the shell of the micro-capsules have to be designed so that it breaks when a crack has developed in the bulk material, and that the healing agent comes across the catalyst. Moreover, the up to 200 μm diameter spheres can possibly interfere with the reinforcement of the polymer, including introducing an unwanted waviness of the fiber reinforcement. The latter drawback can be addressed by replacing the micro-spheres with hollow micro-cylinders [11,36,68,82,83]. The current state-of-the-art for hollow micro-cylinders focuses on using commercial hollow glass fibers embedded in composite materials [11,82,83]. In a similar manner as to the case of micro-spheres, the hollow cylinders are filled with a "healing agent" that is activated once the fiber breaks. A two-phase epoxy system is used, where the epoxy resin is stored in one set of cylinders and the hardening agent is stored in a second set of cylinders. In a layered composite material, the hollow glass fibers are aligned with the reinforcement fibers, for example the fibers with the epoxy resin are aligned with the 0°-ply and the fibers with the hardening agent with the 90°-ply. When cracks develop and propagate, the glass tubes break, allowing the epoxy and the hardening-agent to fill the damaged zone. The materials are selected so that the epoxy cures at ambient conditions. The major challenge with this approach relates to the difficulty of finding suitable glass tubes. Ideally, the properties of the hollow glass tubes should match that of the original reinforcement, so they can replace or enhance the composite structure. To address this, Pang and Bond [83] purchased commercial borosilicate glass tubing and using in-house facilities drew the fibers to external diameter of 60 μm and inner diameter of approximately 42 μm. The fibers were filled with a commercial epoxy repair agent (MY750 Ciba–Geigy) and the corresponding hardening agent respectively [83]. In this case, about 90% of the strength of the original strength is retained after repair, but the strength degrades with time.

With the exception of the reversible cross linking polymers, the repair schemes discussed so far for polymer composites are all based on one-time repairs; once a micro-capsule or micro-fiber breaks the healing agent is consumed, and no further healing will occur if another crack should develop again at the same point. In contrast, self-healing in biological systems can occur multiple times for repeated injuries, assuming a reasonable frequency of injuries. In animals, this is possible by the continuous flow of an intelligent mixture of biochemicals in the vascular network, which is related to the circulatory system. Some attempts are made to mimic vascular network for self-healing [104,107], where a constant supply of healing agent could potentially be provided. There are several manufacturing issues involved here, and opens up many potential research avenues.

Multifunctional Coatings

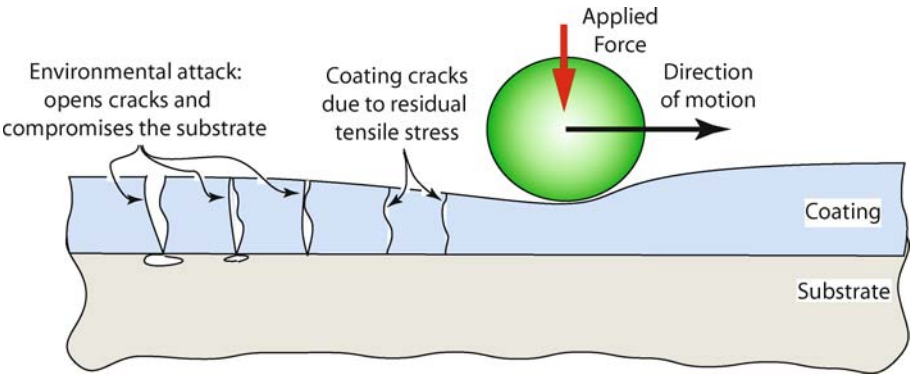
Applying coating on a structure is many time a cost effective way of obtaining a multifunctional composite material. There are many examples that illustrate this, for example environmental barrier coatings, coatings for increased wear resistance, and thermal barrier coatings. Most coatings combine several functions by having multiple layers where each layer contributes a particular function.

Coatings for corrosion resistance (sometimes referred to as environmental coatings) are probably one of the most common classes of coatings, for example covering steel with zinc to obtain a galvanic protection, (the zinc is sacrificed to protect the underlying steel). Even though functional, these cannot be categorized as multifunctional and will not be discussed here. More advanced coatings are now being developed to protect steel and other metals. Some of these coatings have multifunctional capacities. Some particular interesting coating materials are electropolymerized polymer composites including polyaniline (PANi) [51,78,93,102] and polypyrrole (PPy) [51,102]. Polyaniline (PANi) is formed by polymerizing aniline (phenylamine, aminobenzene), $C_6H_5NH_2$ and polypyrrole (PPy) is synthesized pyrrole, C_4H_4NH. The polymers are typically deposited through an electrochemical synthesis in which the thickness can be controlled. The corrosion resistance depends on the deposition parameters including applied potential and the feeding rate of the monomers. By producing a composite of PPy and PANi, the corrosion rate can be reduced with more than two order of magnitudes compared to unprotected materials [51]. To achieve this improved rate, it is crucial to ensure a proper bonding of the structure of the polymer composites deposited on the metal, which must be controlled through the processing parameters [102]. This class of coatings protects in sulfuric acid (H_2SO_4), not so well in hydrochloric acid (HCl), and not at all in a Sodium chloride (NaCl) solution [93]. Thus, care must be taken when using this type of coating for corrosion protection, but evidently, it can be quite useful for a range of applications.

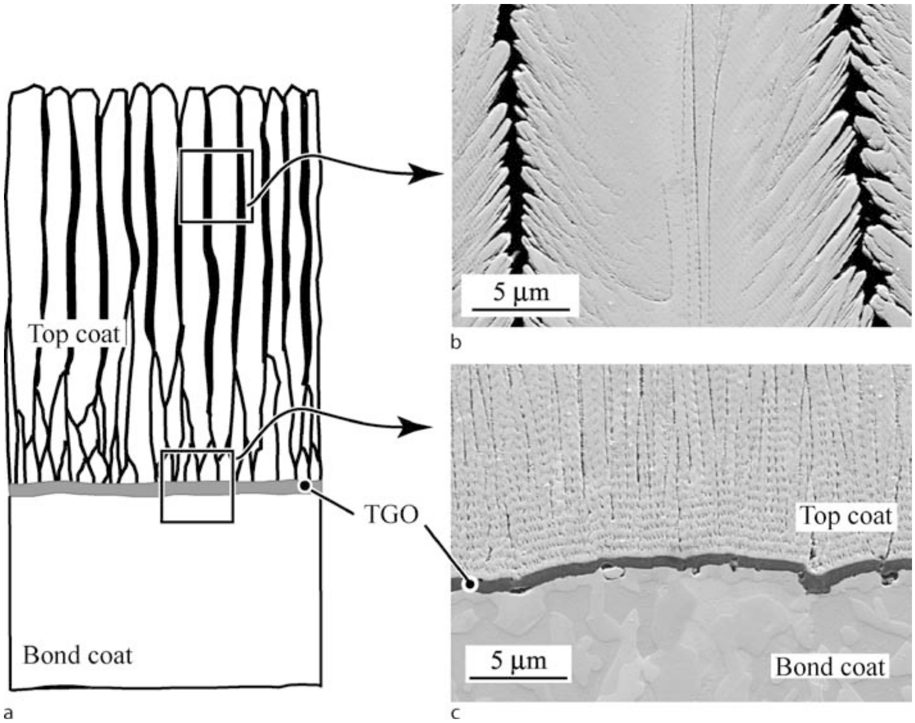
A second class of important coatings is coatings used to ensure low friction and increased wear resistance of the underlying structure, tribological coatings. These coatings are critical for a range of applications, including moving contacts (e. g., bearings), materials processing (e. g., drilling), and applications where addition of lubricants or materials debris from wearing is unacceptable (e. g., food processing, medical implants). Also, by reducing friction in moving parts in vehicles, the fuel efficiency of the vehicle can be significantly increased. When optimizing a coating for wear resistance, the goal is to reach as high hardness as possible [110]. When combining wear resistance with low friction, many other aspects much be considered. There are now several systems used as solid lubricants which allows for both low friction and wear resistance. These include diamond and diamond-like carbon, graphite, molybdenum disulfide, hexagonal boron nitride, boric acid as well as soft metals [37]. An interesting example of a low friction wear-resistant coating consists of a composite coating made from a titanium nitride matrix, TiN, with molybdenum sulphides, MoS_x, dispersed as a second phase. Up to 8% (by weight) addition of MoS_x does not effect the hardness of the coating (thus promoting wear resistance), but decreased the coefficient of friction with more than a factor of two, and consequently increasing the life up to 500 times compared to the TiN coating alone [29].

Of high interest is to combine the corrosion resistance coating with the tribological coatings. This multifunctional coating would then resist both corrosion, wear and provide a low coefficient of friction [113,117]. This would increase the lifetime of many engineering applications, and for example, increase the fuel efficiency of vehicles since it reduces energy losses due to friction. Tribological coatings under stress and at the same time in aggressive environments degrade at a significantly faster rate compared to if they were subjected to wear alone or to the aggressive environments alone, since the two conditions aid each other to aggravate the deterioration of the coatings, as illustrated in Fig. 7. In a recent review by Wood [117] it is clear that this is a research area that requires significant attention in the future.

The last class of coatings we will discuss is coatings used for high temperature protection: thermal barrier coatings (TBCs), used to protect the underlying metallic substrate. TBCs are commonly used for protecting superalloys in gas turbines (both stationary for energy production and mobile for propulsion of airplanes). These systems are a material system with multiple layers, in which each layer is optimized for a particular function. It is made up of a *bond coat* being deposited on a metallic base, after which a ceramic *top coat* is deposited, see Fig. 8. In gas turbine engines, a thermal gradient over the top coat of up to 150°C is achieved from active cooling of the superalloy and by selecting top coat materials with relatively low thermal conductivities [38,66,79,99,100,118]. The metal-


Environmental attack: opens cracks and compromises the substrate
Coating cracks due to residual tensile stress
Applied Force
Direction of motion
Coating
Substrate

Composites, Multifunctional, Figure 7
Illustration of a degradation caused by combined wear and corrosion of a coated structure. Wear (illustrated by a sphere moving on the surface with an applied normal force) causes cracks in the coating. As the cracks develop, the environment can reach the substrate and deteriorate the substrate quickly



Composites, Multifunctional, Figure 8
An example of a thermal barrier coating produced by electron-beam physical vapor deposition (EB-PVD). a Sketch of the system; b scanning electron microscopy (SEM) image of the cross-section of the top coat, close to the surface; c SEM image of the interface between the top coat, thermally grown oxide (TGO) and the bond coat. Note that b and c are of the same scale. The top coat (yttria stabilized zirconia) is intentionally made porous to allow for strain tolerance during thermal cycling

lic bond coat provides oxidation protection to the superalloy by sacrificing itself by supplying aluminum to form an alpha-alumina scale (α-Al_2O_3) between the bond coat and the top coat (Fig. 8). As the alumina scale grows, the aluminum content decreases in the bond coat, ultimately changing the bond coat properties [32,96]. Thus, it is important to control the chemical content of the materials since even small amounts of critical trace elements can enhance or reduce the interfacial fracture toughness of the structure. Even though TBCs have been used for more than a decade, improvements of these multifunctional materials are still critical. By improving their reliability and durability, gas turbine powerplants and aircraft engines can become significantly more efficient, reducing their fuel consumption and reducing the pollution associated with energy production and propulsion.

Other Multifunctional Composites

Polyaniline (PANi) as discussed earlier has also been used as nanoparticles (rather than the matrix) to achieve multifunctional composites. For example Dispenza et al. [34] used PANi particles in a hydrogel matrix obtaining a biocompatible nanocomposite with properties suitable for the development of optoelectronic devices. The composite was obtained by a multistep process, starting with water dispersion polymerization of aniline, followed by γ-irradiation. The γ-irradiation cross links the PANi to the steric stabilizers (either poly-vinyl-pyrrolidone, PVP or poly-vinyl-alcohol, PVA). Depending on the processing conditions, various properties can be obtained, but the PANi particles remain spherical [34]. The composites can undergo two optical transitions and the fluorescence signals can vary in wave-length as a function of pH-value.

Composites with 3-dimensional reinforcements have been developed in order to eliminate a number of shortcomings like low shear and transverse stiffness and strength exhibited by laminated composites, with application in areas such as the aerospace industry. 3-dimensional braided reinforcements give reinforcing support in all the three directions thereby suppressing delamination in the composite and giving a higher damage tolerance [80,81,119].

Another group of composites that are multifunctional are the hybrid composite materials. A hybrid composite is made by combing two or more types of fibers in a single matrix material or in two or more types of matrix materials. This gives a greater possibility of achieving multi-functions by changing the combinations of fibers and/or matrix materials [111]. However, the interactions of the constituent components of a hybrid composite and large number of design variables involved make the design of such a composite complex. Examples are carbon-aramid reinforced epoxy, glass-carbon reinforced epoxy, and carbon-Kevlar reinforced epoxy. Lee [58] classified hybrid composites as (1) interply or tow-by-tow, in which tows of the two or more constituent types of fiber are mixed in a regular or random manner; (2) sandwich hybrids, also known as core-shell, in which one material is sandwiched between two layers of another; (3) interply or laminated, where alternate layers of two (or more) materials are stacked in a regular manner; (4) intimately mixed hybrids, where the constituent fibers are made to mix as randomly as possible so that no over-concentration of any one type is present in the material; (5) other kinds, such as those reinforced with ribs, pultruded wires, thin veils of fiber or combinations of the above [58].

Future Directions

Biomimetics is seeking to mimic nature to design and produce materials comparable or better than the ones nature has produced. The goal is to be able to synthesize organs for replacement in the body. Such organs should be able to perform and grow just like the natural one being replaced. This will eliminate looking for or waiting for donors wanting to donate such a needed organ which, in some cases, the body of the patient may reject in the end.

For non-biological applications, mimicking multifunctional biocomposites should be directed at designing synthetic composite materials that can simultaneously perform more than one function. Self-healing is another aspect being targeted to be achieved in material design. The future goal is not only to achieve multifunctionality in composite materials but also such materials should be self-healing. One great lesson from nature is that nature does not waste materials in building its structures yet with incredible relevant properties. This is because nature builds from bottom up, from molecular level to macro level. Ability to control the design and structure of the material on the molecular level will allow production and fabrication of components, devices and structures with incredible properties and functionalities without excessive use of material or energy inputs. This is a great advantage in conservation of material and energy.

Bibliography

Primary Literature

1. Aizenberg J, Lambert G, Weiner S, Addadi L (2002) Factors involved in the formation of amorphous and crystalline calcium carbonate: A study of an ascidian skeleton. J Am Chem Soc 124:32–39
2. Aizenberg J, Weaver JC, Thanawala MS, Sundar VC, Morse DE, Fratzl P (2005) Skeleton of Euplectella sp.: Structural hierarchy from the nanoscale to the macroscale. Science 309:275–278
3. Akle BJ, Bennett MD, Leo DJ (2006) High-strain ionomeric-ionic liquid electroactive actuators. Sens Actuators A-Phys 126:173–181
4. Bagwell RM, McManaman JM, Wetherhold RC (2006) Short shaped copper fibers in an epoxy matrix: Their role in a multifunctional composite. Compos Sci Technol 66:522–530
5. Bandaru PR (2007) Electrical properties and applications of carbon nanotube structures. J Nanosci Nanotechnol 7:1239–1267

6. Barthelat F (2007) Biomimetics for next generation materials. Philos Trans R Soc A-Math Phys Eng Sci 365:2907–2919
7. Barthelat F, Tang H, Zavattieri PD, Li CM, Espinosa HD (2007) On the mechanics of mother-of-pearl: A key feature in the material hierarchical structure. J Mech Phys Solids 55:306–337
8. Bennett MD, Leo DJ (2004) Ionic liquids as stable solvents for ionic polymer transducers. Sens Actuators: A-Physical 115:79–90
9. Bennett MD, Leo DJ, Wilkes GL, Beyer FL, Pechar TW (2006) A model of charge transport and electromechanical transduction in ionic liquid-swollen Nafion membranes. Polymer 47:6782–6796
10. Biddiss E, Chau T (2006) Electroactive polymeric sensors in hand prostheses: Bending response of an ionic polymer metal composite. Med Eng Phys 28:568–578
11. Bleay SM, Loader CB, Hawyes VJ, Humberstone L, Curtis PT (2001) A smart repair system for polymer matrix composites. Compos Pt A-Appl Sci Manuf 32:1767–1776
12. Brown EN, White SR, Sottos NR (2004) Microcapsule induced toughening in a self-healing polymer composite. J Mater Sci 39:1703–1710
13. Brown EN, White SR, Sottos NR (2005) Retardation and repair of fatigue cracks in a microcapsule toughened epoxy composite – part 1: Manual infiltration. Compos Sci Technol 65:2466–2473
14. Brown EN, White SR, Sottos NR (2005) Retardation and repair of fatigue cracks in a microcapsule toughened epoxy composite – Part II: In situ self-healing. Compos Sci Technol 65:2474–2480
15. Bruet BJF, Qi HJ, Boyce MC, Panas R, Tai K, Frick L, Ortiz C (2005) Nanoscale morphology and indentation of individual nacre tablets from the gastropod mollusc Trochus niloticus. J Mater Res 20:2400–2419
16. Bruet BJF, Qi HJ, Boyce MC, Panas R, Tai K, Frick L, Ortiz C (2005) Nanoscale morphology and indentation of individual nacre tablets from the gastropod mollusc Trochus niloticus (vol 20, Pg 2400, 2005). J Mater Res 20:3157–3157
17. Chang MC, Ikoma T, Kikuchi M, Tanaka J (2001) Preparation of a porous hydroxyapatite/collagen nanocomposite using glutaraldehyde as a cross-linkage agent. J Mater Sci Lett 20:1199–1201
18. Chang MC, Ko CC, Douglas WH (2003) Conformational change of hydroxyapatite/gelatin nanocomposite by glutaraldehyde. Biomaterials 24:3087–3094
19. Chen J, Ramasubramaniam R, Xue C, Liu H (2006) A versatile, molecular engineering approach to simultaneously enhanced, multifunctional carbon-nanotube-polymer composites. Adv Functi Mater 16:114–119
20. Chen X, Cao G (2007) Review: Atomistic studies of mechanical properties of carbon nanotubes. J Theor Comput Nanosci 4:823–839
21. Chen X, Huang Y (2008) Nanomechanics modeling and simulation of carbon nanotubes. J Eng Mech 134:211–216
22. Chen X, Surani FB, Kong X, Punyamurtula VK, Qiao Y (2006) Energy absorption performance of a steel tube enhanced by a nanoporous material functionalized liquid. Appl Phys Lett 89:241918
23. Chen XX, Dam MA, Ono K, Mal A, Shen HB, Nutt SR, Sheran K, Wudl F (2002) A thermally re-mendable cross-linked polymeric material. Science 295:1698–1702
24. Chen XX, Wudl F, Mal AK, Shen HB, Nutt SR (2003) New thermally remendable highly cross-linked polymeric materials. Macromolecules 36:1802–1807
25. Cheng L, Wang L, Karlsson AM (2008) Image analyses of two crustacean exoskeletons and implications of the exoskeletal microstructure on the mechanial behavior. J Mater Res (in press)
26. Choi H, Sofranko AC, Dionysiou DD (2006) Nanocrystalline TiO2 photocatalytic membranes with a hierarchical mesoporous multilayer structure: Synthesis, characterization, and multifunction. Adv Funct Mater 16:1067–1074
27. Chung DDL (2001) Electromagnetic interference shielding effectiveness of carbon materials. Carbon 39:279–285
28. Compere P, Goffinet G (1987) Ultrastructural shape and 3-dimensional organization of the intracuticular canal systems in the mineralized cuticle of the green crab Carcinus–Maenas. Tissue Cell 19:839–857
29. Cosemans P, Zhu X, Celis JP, Van Stappen M (2003) Development of low friction wear-resistant coatings. Surf Coat Technol 174:416–420
30. Currey JD (1977) Mechanical-properties of mother of pearl in tension. Proc Royal Soc London Ser B-Biol Sci 196:443–463
31. Currey JD (2005) Hierarchies in Biomineral Structures. Science 309:253–254
32. Darzens S, Karlsson AM (2004) On the microstructural development in platinum-modified nickel-aluminide bond coats. Surf Coat Technol 177:108–112
33. De Stasio G, Schmitt MA, Gellman SH (2005) Spectromicroscopy at the organic-inorganic interface in biominerals. Am J Sci 305:673–686
34. Dispenza C, Leone M, Lo Presti C, Librizzi F, Spadaro G, Vetri V (2006) Optical properties of biocompatible polyaniline nanocomposites. J Non-Cryst Solids 352:3835–3840
35. Dresselhaus MS, Dresselhaus G, Eklund PC (1996) Science of fullerenes and carbon nanotubes. Academic Press, New York
36. Dry C (1996) Procedures developed for self-repair of polymer matrix composite materials. Compos Struct 35:263–269
37. Erdemir A (2005) Review of engineered tribological interfaces for improved boundary lubrication. Tribol Int 38:249–256
38. Evans AG, Mumm DR, Hutchinson JW, Meier GH, Petit FS (2001) Mechanisms controlling the durability of thermal barrier coatings. Prog Mater Sci 46:505–553
39. Ferraris M, Montorsi M, Salvo M (2000) Class coating for Sicf-Sic composites for high-temperature application. Acta Mater 48:4721–4724
40. Fincham AG (2007) Structural Biology of the Enamel Proteins. http://www.usc.edu/hsc/dental/Info/Research/9.html. Accessed 19 Sep 2007
41. Ghosh P, Katti D, Katti K, Mohanty B, Verma D (2005) Mechanical properties of biological nanocomposite nacre: Multiscale modeling and experiments on nacre from Red Abalone. Mater Res Soc Symp Proc 898:7–12
42. Gilbert PUPA, Abrecht M, Frazer BH (2005) The organic-mineral interface in biominerals. Rev Miner Geochem 59:157–185
43. Gooding JJ (2005) Nanostructuring electrodes with carbon nanotubes: A review on electrochemistry and applications for sensing. Electrochimica Acta 50:3049–3060
44. Guo QG, Song JR, Liu L, Zhang BJ (1999) Relationship between oxidation resistance and structure of B4c-Sic/C composites with self-healing properties. Carbon 37:33–40
45. Han A, Chen X, Surani FB, Qiao Y (2008) Rate-dependent in-

filtration of a viscous liquid in nanopores. Appl Phys Lett in press

46. Han A, Qiao YT (2007) Hermal effects on infiltration of a solubility- sensitive volume memory liquid. Phil Mag Lett 87:25–31
47. Heitner-Wirguin C (1996) Recent advances in perfluorinated ionomer membranes: Structure, properties and applications. J Membr Sci 120:1–33
48. Hernandez E, Goze C, Bernier P, Rubio A (1998) Elastic properties of C and Bxcynz composite nanotubes. Phys Rev Lett 80:4502–4505
49. Hughes M, Shaffer MSP, Renouf AC, Singh C, Chen GZ, Fray J, Windle AH (2002) Electrochemical capacitance of nanocomposite films formed by coating aligned arrays of carbon nanotubes with polypyrrole. Adv Mater 14:382–385
50. Iijima S (1991) Helical microtubules of graphitic carbon. Nature 354:56–58
51. Iroh JO, Zhu YR, Shah K, Levine K, Rajagopalan RR, Uyar T, Donley M, Mantz R, Johnson J, Voevodin NN, Balbyshev VN, Khramovb AN (2003) Electrochemical synthesis: A novel technique for processing multi-functional coatings. Prog Org Coat 47:365–375
52. Katti K, Katti DR, Tang J, Pradhan S, Sarikaya M (2005) Modeling mechanical responses in a laminated biocomposite – Part II: Nonlinear responses and nuances of nanostructure. J Mater Sci 40:1749–1755
53. Kessler MR, Sottos NR, White SR (2003) Self-healing structural composite materials. Compos Pt A-Appl Sci Manuf 34:743–753
54. Kroto HW, Heath JR, Obrien SC, Curl RF, Smalley RE (1985) C-60 – Buckminsterfullerene. Nature 318:162–163
55. Kumar P, Nukala VV, Šimunović S (2005) Statistical physics models for nacre fracture simulation. Phys Rev E 72:041919
56. Lee MJ, Jung SH, Kim GS, Moon I, Lee S, Mun MS (2007) Actuation of the artificial muscle based on ionic polymer metal composite by electromyography (Emg) signals. J Intell Mater Syst Struct 18:165–170
57. Lee S, Kim KJ, Park IS (2007) Modeling and experiment of a muscle-like linear actuator using an ionic polymer metal composite and its actuation characteristics. Smart Mater Struct 16:583–588
58. Lee SM (1989) Dictionary of composite materials technology. Technomic Publishing Company Inc, Lancaster
59. Li XD, Xu ZH, Wang RZ (2006) In situ observation of nanograin rotation and deformation in nacre. Nano Lett 6:2301–2304
60. Lu JP (1997) Elastic properties of carbon nanotubes and nanoropes. Phys Rev Lett 79:1297–1300
61. Ma YF, Ali SR, Dodoo AS, He HX (2006) Enhanced sensitivity for biosensors: Multiple functions of DNA-wrapped single-walled carbon nanotubes in self-doped polyaniline nanocomposites. J Phys Chem B 110:16359–16365
62. Ma ZJ, Huang J, Sun J, Wang GN, Li CZ, Xie LP, Zhang RQ (2007) A novel extrapallial fluid protein controls the morphology of nacre lamellae in the pearl oyster, Pinctada Fucata. J Biol Chem 282:23253–23263
63. Mayer G (2005) Rigid biological systems as models for synthetic composites. Science 310:1144–1147
64. McGee T (2007) Biomimicry: Nacre inspires transparent strong as steel plastic. Science & Technology (biopolymer) http://www.treehugger.com/files/2007/10/plastic_steel.php. Accessed 1 Nov 2007
65. Metzler RA, Abrecht M, Olabisi RM, Ariosa D, Johnson CJ, Frazer BH, Coppersmith SN, Gilbert PUPA (2007) Architecture of columnar nacre, and implications for its formation mechanism. Phys Rev Lett 98:268102
66. Miller RA (1984) Oxidation-based model for thermal barrier coating life. J Am Ceram Soc 67:517–521
67. Mirfakhrai T, Madden JDW, Baughman RH (2007) Polymer artificial muscles. Mater Today 10:30–38
68. Motuku M, Vaidya UK, Janowski GM (1999) Parametric studies on self-repairing approaches for resin infused composites subjected to low velocity impact. Smart Mater Struct 8:623–638
69. Mount AS, Wheeler AP, Paradkar RP, Snider D (2004) Hemocyte-mediated shell mineralization in the eastern oyster. Science 304:297–300
70. Naslain R (2004) Design, preparation and properties of non-oxide CMCS for application in engines and nuclear reactors: An overview. Compos Sci Technol 64:155–170
71. Naslain R, Guette A, Rebillat F, Pailler R, Langlais F, Bourrat X (2004) Boron-bearing species in ceramic matrix composites for long-term aerospace applications. J Solid State Chem 177:449–456
72. Naslain RR, Pailler R, Bourrat X, Bertrand S, Heurtevent F, Dupel P, Lamouroux F (2001) Synthesis of highly tailored ceramic matrix composites by pressure-pulsed CVI. Solid State Ion 141:541–548
73. Nassif N, Pinna N, Gehrke N, Antonietti M, Jager C, Colfen H (2005) Amorphous layer around aragonite platelets in nacre. Proc Natl Acad Sci USA 102:12653–12655
74. Nemat-Nasser S, Li JY (2000) Electromechanical response of ionic polymer-metal composites. J Appl Phys 87:3321–3331
75. Nemat-Nasser S, Wu YX (2003) Comparative experimental study of ionic polymer-metal composites with different backbone ionomers and in various cation forms. J Appl Phys 93:5255–5267
76. Nemat-Nasser S, Wu YX (2006) Tailoring the actuation of ionic polymer-metal composites. Smart Mater Struct 15:909–923
77. Neville AC (1975) Biology of the arthropod cuticle. Springer, Berlin
78. Ozyilmaz AT, Erbil A, Yazici B (2004) Investigation of corrosion behaviour of stainless steel coated with polyaniline via electrochemical impedance spectroscopy. Prog Org Coat 51:47–54
79. Padture NP, Gell M, Jordan EH (2002) Materials science – Thermal barrier coatings for gas-turbine engine applications. Science 296:280–284
80. Pandey R, Hahn HT (1996) Designing with 4-step braided fabric composites. Compos Sci Technol 56:623–634
81. Pandey R, Hahn HT (1996) Visualization of 4-step braided fabric composites. Compos Sci Technol 56:161–170
82. Pang JWC, Bond IP (2005) 'Bleeding Composites' – Damage detection and self-repair using a biomimetic approach. Compos Pt A-Appl Sci Manuf 36:183–188
83. Pang JWC, Bond IP (2005) A hollow fibre reinforced polymer composite encompassing self-healing and enhanced damage visibility. Compos Sci Technol 65:1791–1799
84. Podsiadlo P, Liu ZQ, Paterson D, Messersmith PB, Kotov NA (2007) Fusion of seashell nacre and marine bioadhesive analogs: High-strength nanocompoisite by layer-by-layer assembly of clay and L-3,4-dihydroxyphenylalanine polymer. Adv Mater 19:949–955

85. Qi HJ, Bruet BJF, Palmer JS, Ortiz C, Boyce MC (2006) Micromechanics and macromechanics of the tensile deformation of nacre. In: Holzapfel GA, Ogden RW (eds) Mechanics of biological tissue. Springer, Berlin, pp 189–203
86. Qiao Y, Cao G, Chen X (2007) Effect of gas molecules on nanofluidic behaviors. J Am Chem Soc 129:2355–2359
87. Qiao Y, Punyamurtula VK, Han A (2007) Mechanoelectricity of a nanoporous monel – Electrolyte solution system. J Power Sources 164:931
88. Royal Society of Chemistry (RSC) (2006) Glass bones, education in chemistry. http://www.rsc.org/Education/EiC/issues/2006Nov/GlassBones.asp. Accessed 25 Sep 2007
89. Sadeghipour K, Salomon R, Neogi S (1992) Development of a novel electrochimically active membrane and "smart" material based vibration sensor/damper. Smart Mater Struct 1:172–179
90. Sainz R, Benito AM, Martinez MT, Galindo JF, Sotres J, Baro AM, Corraze B, Chauvet O, Dalton AB, Baughman RH, Maser WK (2005) A soluble and highly functional polyaniline-carbon nanotube composite. Nanotechnology 16:S150–S154
91. Sarikaya M, Fong H, Sunderland N, Flinn BD, Mayer G, Mescher A, Gaino E (2001) Biomimetic model of a sponge-spicular optical fiber – Mechanical properties and structure. J Mater Res 6:1420–1428
92. Sarikaya M, Liu J, Aksay IA (1995) Nacre: Properties, crystallography, morphology and formation. In: Sarikaya M, Aksay IA (eds) Biomimetics design and processing of materials. American Institute of Physics, New York pp 34–90
93. Sathiyanarayanan S, Devi S, Venkatachari G (2006) Corrosion protection of stainless steel by electropolymerised pani coating. Prog Org Coat 56:114–119
94. Shahinpoor M, Kim KJ (2000) The effect of surface-electrode resistance on the performance of ionic polymer-metal composite (IPMIC) artificial muscles. Smart Mater Struct 9:543–551
95. Shahinpoor M, Kim KJ (2001) Ionic polymer-metal composites: I. fundamentals. Smart Mater Struct 10:819–833
96. Shi J, Darzens S, Karlsson AM (2005) Aspects of the morphological evolution in thermal barrier coatings and the intrinsic thermal mismatch therein. Mater Sci Eng A-Struct Mater Prop Microstruct Processing 392:301–312
97. Sierakowski RL, Telitchev IY, Zhupanska OI (2008) On the impact response of electrified carbon fiber polymer matrix composites: Effects of electric current intensity and duration. Compos Sci Technol 68:639–649
98. Snyder DR, Sierakowski RL, Chenette ER, Aus JW (2001) Preliminary assessment of electro-termo-magnetically loaded composite panel impact resistance/crack propagation with high speed digital laser photography. 24th international congress on high-speed photography and photonics. Proc SPIE 4183:488–513
99. Stiger M, Yanar N, Topping M, Pettit F, Meier G (1999) Thermal barrier coatings for the 21st century. Z Metallkunde 90:1069–1078
100. Strangman TE (1985) Thermal barrier coatings for turbine airfoils. Thin Solid Films 127:93–105
101. Strumpler R, Glatz-Reichenbach J (1999) Conducting polymer composites. J Electroceram 3:329–346
102. Tan CK, Blackwood DJ (2003) Corrosion protection by multilayered conducting polymer coatings. Corros Sci 45:545–557
103. Tang ZY, Kotov NA, Magonov S, Ozturk B (2003) Nanostructured artificial nacre. Nat Mater 2:413–418
104. Therriault D, White SR, Lewis JA (2003) Chaotic mixing in three-dimensional microvascular networks fabricated by direct-write assembly. Nat Mater 2:265–271
105. Thostenson ET, Chou TW (2006) Processing-structure-multifunctional property relationship in carbon nanotube/epoxy composites. Carbon 44:3022–3029
106. Thostenson ET, Ren ZF, Chou TW (2001) Advances in the science and technology of carbon nanotubes and their composites: A review. Compos Sci Technol 61:1899–1912
107. Toohey KS, Sottos NR, Lewis JA, Moore JS, White SR (2007) Self-healing materials with microvascular networks. Nat Mater 6:581–585
108. University of the Western Cape UWC (2007) Bone, internet BioEd project. http://www.botany.uwc.ac.za/sci_ed/grade10/mammal/bone.htm. Accessed 26 Sep 2007
109. Veedu VP, Cao AY, Li XS, Ma KG, Soldano C, Kar S, Ajayan PM, Ghasemi-Nejhad MN (2006) Multifunctional composites using reinforced laminae with carbon-nanotube forests. Nat Mater 5:457–462
110. Veprek S, Veprek-Heijman MGJ, Karvankova P, Prochazka J (2005) Different approaches to superhard coatings and nanocomposites. Thin Solid Films 476:1–29
111. Wan YZ, Lian JJ, Huang Y, He F, Wang YL, Jiang HJ, Xin JY (2007) Preparation and characterization of three-dimensional braided carbon/kevlar/epoxy hybrid composites. J Mater Sci 42:1343–1350
112. Wang J, Xu CY, Taya M, Kuga Y (2006) Mechanical stability optimization of flemion-based composite artificial muscles by use of proper solvent. J Mater Res 21:2018–2022
113. Wang LP, Zhang JY, Zeng ZX, Lin YM, Hu LT, Xue QJ (2006) Fabrication of a nanocrystalline Ni-Co/Coo functionally graded layer with excellent electrochemical corrosion and tribological performance. Nanotechnology 17:4614–4623
114. Weiland LM, Leo DJ (2005) Computational analysis of ionic polymer cluster energetics. J Appl Phys 97:10
115. White SR, Sottos NR, Geubelle PH, Moore JS, Kessler MR, Sriram SR, Brown EN, Viswanathan S (2001) Autonomic healing of polymer composites. Nature 409:794–797
116. Wilt FH (2005) Developmental biology meets materials science: Morphogenesis of biomineralized structures. Dev Biol 280:15–25
117. Wood RJK (2007) Tribo-corrosion of coatings: A review. J Phys D-Appl Phys 40:5502–5521
118. Wright P (1998) Influence of cyclic strain on life of a PVD TBC. Mater Sci Eng A 245:191–200
119. Wu XQ, Li JL, Shenoi RA (2006) Measurement of braided preform permeability. Compos Sci Technol 66:3064–3069
120. Zako M, Takano N (1999) Intelligent material systems using epoxy particles to repair microcracks and delamination damage in GFRP. J Intell Mater Syst Struct 10:836–841
121. Zhupanska OI, Sierakowski RL (2007) Effects of an electromagnetic field on the mechanical response of composites. J Composite Mater 41:633–652

Books and Reviews

Bar-Cohen Y (2006) Biomimetics: Biologically inspired technologies. CRC Press Taylor and Francis Group LLC, New York

Currey J (1984) The mechanical adaptations of bones. Princeton University Press, Princeton

Forest Products Laboratory (1999) Wood handbook – Wood as an engineering material. Gen Tech Rep FPL–GTR–113. US Department of Agriculture, Forest Service, Forest Products Laboratory, Madison
Lowenstam HA, Weiner S (1989) On biomineralization. Oxford University Press, New York
Mann S (2001) Biomineralization. Oxford University Press, New York
Wainwright SA, Biggs WD, Currey JD, Gosline JM (1976) Mechanical design in organisms. Wiley, New York

Computational Chemistry, Introduction to Complexity in

DANAIL BONCHEV
Department of Mathematics and Applied Mathematics
and Center for the Study of Biological Complexity,
Virginia Commonwealth University, Richmond, USA

Defining complexity one remembers the words of St. Augustine: "I know very well what time is, until I have to explain it to somebody else." Intuitively so clear, complexity is so difficult to define rigorously. Complexity is sometimes mistaken with size, large systems are presumably complex. Are they always? Is a big piece of rock more complex than a tiny living cell? What makes a system complex is not the number of its components, but the number and kind of their interactions. As a result of these interactions the system formed has properties and behavior different from those of its components. The dynamics of a complex system's interactions is described by more sophisticated mathematical models (nonlinear differential equations). But how much complexity is needed to observe such emergent properties and dynamics? Given the number of components how many interactions would produce a new behavior?

We may not be able to specify an ideal definition of complexity or a detailed reply to these questions. However, what we can and must do is to quantify complexity, to measure it, as we do with electric charge without knowing what exactly it is. In the realm of computational chemistry, this has been done during the last 30 years defining a variety of measures of *molecular complexity* – compositional, structural, and dynamic. While molecular compositional complexity is straightforward for calculation from the number of atoms of different elements, the structural complexity of molecules can be quantitatively assessed by a variety of descriptors (termed also topological indices), based on graph theory and information theory. Presenting molecular structure as a graph gives rise to a multitude of quantitative descriptors. However, only few of them satisfy the set of criteria formulated in this field. A major requirement for a complexity descriptor is to match in a series of molecules certain patterns of increasing complexity, based on the number of complexifying elements – branches, cycles, central organization, and others (see ▶ Topological Complexity of Molecules). New methods consider as effective complexity measures the overall subgraph count, the sum of vertex degrees of all subgraphs ("overall connectivity"), and the count of all random walks. Shannon's information theory has also been broadly used in quantifying complexity of chemical systems. Shannon's function was found to be very useful in characterizing different distributions of atoms in molecules. It reflects the increase in complexity with the increase in the diversity of atoms and bonds in molecules, the diversity being caused not only by the elemental composition of atoms, but also by the vertex degrees and vertex total distances distributions in molecular graphs. Moreover, Shannon's information indices calculated from electron distributions in atoms and nucleon distributions in atomic nuclei were shown to be an important tool for predicting the properties of superheavy chemical elements, and their isotopes (see ▶ Information Theoretic Complexity Measures). The universality of the graph theoretical and information theoretical methods developed in the field of computational chemistry makes them also of interest in network complexity analysis in biology, social sciences, computer sciences, and other branches of science and technology.

A higher-level theoretical framework of molecular similarity and complexity is the quantum mechanical, density-functional formalism (DFT). This quantitative approach enables comparing and ordering of molecular structures, which finds application in a quantum version of the linear and nonlinear QSPR (see ▶ Quantum Similarity and Quantum Quantitative Structure-Properties Relationships (QQSPR)). Tuning the relative reaction rates of the different steps of a complex chemical reaction would enable the optimal design of many industrially important reactions. The DFT technique was recently used to address such a fine tuning procedure (see ▶ Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis). Such an approach promises to provide in the near future tools for effective screening of potential new materials with tailored catalytic properties.

Drug discovery is a major field of interest for applied computational chemistry. The lengthy process of design of a new drug starts with a search for the best computational model that relates molecular structure, composition, and properties to certain biological activity that has the desired therapeutic effect. This vast area of research, which spans over the last 50 years and is known as QSAR (Quantitative Structure-Activity Relationships) modeling, reduces strongly the potential drug candidates selected for synthesis and testing. While composition and properties are experimentally determined quantities, the translation of molecular structure into numbers is not unique, and includes a multitude of structural parameters termed molecular descriptors (see ▶ Drug Design, Molecular Descriptors in). Different classes of descriptors capture different aspects of molecular structure. The manner in which atoms are connected to each other is characterized by graph theory-based structural invariants called topological or two-dimensional (2D) indices. Three-dimensional (3D) descriptors are used to account for the metric and geometry of molecules, while conformational (or 4D) parameters complete the picture by accounting for the presence of a variety of 3D structures in molecules having free rotating atomic groups. Distinct classes include quantum chemical descriptors, such as atomic charges, bond orders, superdelocalizabilities and dipole moments, as well as physicochemical properties like solubility and binding constants. No theory exists to predict which of these descriptor classes would work the best for a given class of biologically active compounds, thus making all of them useful tools in drug design.

The search for new promising drug candidates experienced a quantum leap, when the intelligent guesses based on analogies with naturally occurring compounds were replaced by automated searches in huge databases of chemical compounds. The use of such combinatorial libraries has led to an explosive increase in the number of identified therapeutic targets, and enabled producing tailor-made drug molecules. All this increased considerably the complexity of the *in silico* methods used in drug discovery. The QSAR modeling reached a stage of maturity, with rigorously defined criteria for selection of molecular descriptors, as well as selection of methods for model validation and determining the applicability domains of the models derived (see ▶ QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern). It was realized that no particular QSAR modeling method has an advantage over the others, as well as that each specific technique and each specific class of descriptors have their unique contribution to the relationship with the examined biological activity. This resulted in the creation of combinatorial QSAR methodology, which produces a multitude of models, satisfying a set of acceptability criteria.

In parallel with the optimization of traditional QSAR methods and their adaptation to the complex datasets available, the field of drug design benefited tremendously by using the methods of artificial intelligence. Most of these methods have been inspired by the manner in which the living things function and evolve (see ▶ Drug Design with Artificial Intelligence Methods). Genetic and evolutionary algorithms are stochastic methods that solve optimization problems by evolving solutions based on concepts of DNA genetics and Darwinian evolution. The ant colony optimization (ACO) algorithms mimic the behavior of some ant species when searching for the shortest path to a food source. The particle swarm optimization (PSO) algorithm mirrors the social behavior of large groups of individuals, such as bird flocking, fish schooling, and animal herding. Swarm intelligence algorithms are used in drug design for diverse application, including selection of structural descriptors for QSAR models, enzyme-inhibitor docking, and gene expression. Another class of algorithms, termed Artificial Immune Systems (AIS), makes use of learning and memory principles used in the human immune system. AIS models found application in diverse biological and medical problems, such as prediction of protein structure, cancer diagnosis, classification of gene expression data, and recognition of ECG arrhythmia. Machine learning (decision trees, support vector machines, k-nearest neighbors, etc.) is an important field of artificial intelligence methods. It extracts information from experimental data by computational and probabilistic methods, and by using a set of rules it predicts the properties of objects not included in the learning set (see ▶ Drug Design with Machine Learning). Machine-learning techniques, such as artificial neural networks (ANNs), have been widely applied to QSAR data (see ▶ Drug Design with Artificial Neural Networks). Its success story began with the creation of algorithms for the training of multilayer feed-forward (MLF) artificial neural networks. MLF was widely used to model biological and toxicological properties of chemical compounds. Other networks used in drug design are self-organizing maps, counter-propagation networks, and probabilistic neural networks.

Biochemistry became for biology what physics has been for chemistry – the fundament to build on and explain. The essence of life, and even its origin, is nowadays in a process of redefining, proceeding from the self-organization of biomolecules (DNA, RNA, proteins, metabolites) in biochemical networks. All aspects of biological development, such as cell differentiation, tissue multilayering, segmentation and left-right asymmetry, can be related to the physicochemical processes of self-organizing. The complex dynamics of self-organizing is characterized by oscillations, pattern formation, and the emergence of multiple steady states. Origin of life is the domain where chemical self-organization and biological evolution meet (see ▶ Biological Development and Evolution, Complexity and Self-organization in). Networks of different kinds and sizes appear at different stages of molecular evolution. (see ▶ Molecular Evolution, Networks in). The evolution of networks is not yet studied in detail, yet it is already established that gene mutations cause addition and deletion of nodes in protein–protein interaction networks, transcriptional regulatory networks, and metabolic networks, whereas rewiring of nodes is an evolutionary mechanism observed only in the first two networks.

Complexity of biological systems is multifaceted. Traditionally, the genome sequence complexity is viewed as a universal method for comparing species complexity, despite the controversy with some of the complexity measures examined. The Kolmogorov's information, Shannon's information, and Jensen–Shannon divergence are among the most frequently used measures (see ▶ Biological Complexity and Biochemical Information). Other types of biocomplexity, based on the structure and functionality of biological systems, have also been analyzed. Recently, the advance in the study of biological networks has offered new options for the assessment of biological complexity, including modeling the evolution of complexity in artificial life systems. Essential for each specific quantitative measure of network complexity is its divergence from the random network having the same size and the same average connectivity. Modular and motif (subgraph) information content of biological networks were shown to reflect well the complex network organization. Complexity of the dynamics of biochemical reactions is closely related to the cell fractal structure, and has important consequences for the dynamic behavior of cells. The processes of organizing the intracellular macromolecular systems obey chaotic dynamics. Non-linear interactions in and between spatial and temporal domains and over wide ranges of scales underlie the emergent properties of complex biological systems (see ▶ Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in). With their nonlinearity, fractal dimensions, and attractors-controlled dynamics, the processes of the self-organizing chaos are of major importance for complexity theory. The high complexity of biological systems is a challenge for the traditional modeling of dynamics based on continuous models and ordinary differential equations (ODE). The cellular automata method (see ▶ Cellular Automata Modeling of Complex Biochemical Systems) appears as a viable alternative for modeling of processes within biological systems, including their metabolic, protein, and gene regulatory networks, as well as for identifying dynamic patterns, and devising of strategies for pathway control.

Chemistry and chemical engineering have entered into the 21st century with a broad spectrum of new classes of sophisticated and even "smart" materials of high complexity. Nanotechnology offers structures of unusual properties based on atomic clusters, nanotubes, and DNA. Composite materials and special polymers extend essentially the areas of application of classical materials. Many of the new materials are created with the technology of molecular self-assembly, adopted from the living nature. The high complexity of nanoscale materials (see ▶ Nanoscale Atomic Clusters, Complexity of) stems from the wealth of possible structures, and the very strong size-dependence of their properties. Clusters have great potential as a means of optimization of existing technologies, developing principally new technological processes, materials, and devices.

DNA provides basic building blocks for constructing nanostructures with specific functional features like molecular recognition, self-assembly, and predictable structure (see ▶ DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires). Periodic protein arrays can be constructed by templated self-assembly onto DNA nanogrids. This enables producing target materials with predictable 3D structure like highly conductive nanowires. The principles of biomolecular self-assembly can be extended to the modern materials synthesis, leading to a broad range of new materials and processes with significant technological impact (see ▶ Self--assembled Materials). The complex nature of self-assembling processes makes their modeling a nontrivial task. The methods used cover from atomic scale (molecular dynamics and Monte Carlo simulations) to mesoscale (kinetic Monte Carlo and phase field modeling) to lattice methods, to Random Sequential Adsorption (RSA) and cellular automata techniques. The development of complex, multi-scale modeling approaches is under way and shows great promise.

The mass-production of industrial polymers was dominated for a long time by polymers with linear macromolecular structure like polyethylene and polypropylene,. The appearance of nonlinear polymers with a variety of branched-tree topologies offered new classes of materials with properties unseen in linear polymers (see ▶ Polymers, Non-linearity in). The entanglement of branches in the star- and comb-shaped, and hyper-branched (known also as dendrimers) polymers, results in complicated conformational statistics and dynamics. The latter is modeled surprisingly well using graph theory, providing reliable prediction of the materials' properties in melt and solid states. Even more unusual properties are offered by composites. Composite materials combine the best features of different materials and are ideally suited to achieve multifunctionality and to form materials that have a broad spectrum of desired properties (see ▶ Composites, Multifunctional). Carbon fiber and carbon nanotube-based composites can be formed with materials as diverse as polymers, cement, and ceramics. Perhaps the most unusual property demonstrated by some complex composites is the self-healing. Inspired from the technological secrets of the living things, such dream materials not requiring continuous support will become one of the most exciting novelties of 21st century.

Computational and Theoretical Nanoscience, Introduction to

YONG S. JOE
Center for Computational Nanoscience,
Department of Physics and Astronomy,
Ball State University, Muncie, USA

Nanoscience and nanotechnology change the nature of almost every human-made object in this century. Advances in the field of nanoscience empower us with new tools for proving electronic devices with ever-decreasing scale. Many people have projected that nanometer-scale devices will continue this trend, bringing control of matter to unprecedented scales. This includes scale reduction not only in microelectronics, but also in fields such as quantum-switch-based computing in the shorter term. These advances have the potential to change the way we engineer our environment, construct and control systems, and interact in society.

Computational science, which has emerged as a third way of doing research, one that complements theory and experiment, plays a key role in developing our understanding of materials at the nanometer scale and in the development "by-design" of new nanoscale materials and devices. Hence, modeling and simulation are now integral components of scientific research.

It is essential to have a detailed understanding of quantum effects in electronic transport to design devices effectively at the nanoscale, sustain the miniaturization trends of integrated circuits, and create new engineered nanostructures [see ▶ Quantum Phenomena in Semiconductor Nanostructures]. The resonance phenomena has a special attention in the electronic transport of non-interacting electrons through the infinite rectilinear quantum wires with impurities and one-dimensional rings with impurities connected to current leads [see ▶ Resonances in Electronic Transport Through Quantum Wires and Rings]. The Fano resonance, which is a manifestation of the interference between a localized state and the continuum states, is investigated in an Aharonov–Bohm ring and in an open three-terminal interferometer with a quantum dot [see ▶ Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring]. It is also important to understand quantum transport and Kondo physics in coupled quantum dots with Anderson impurity [see ▶ Quantum Impurity Physics in Coupled Quantum Dots]. The complex quantum dots, treated as artificial molecules, are an excellent model system for studying dynamical symmetries and Kondo effect. The simplest of such complex objects is the double quantum dot which consists of two islands with confined electrons [see ▶ Tunneling Through Quantum Dots with Discrete Symmetries]. Since the quantum interference of the system depends not only on the dynamical phase but also on the geometric phase, the investigation of the geometric phase and its effect on nanostructures in a dynamical process of the environment is of particular interest [see ▶ Geometric Phase and Related Phenomena in Quantum Nanosystems]. In studying complex systems we have to deal with coupled phenomena and processes at a multitude of different spatial and temporal scales. Understanding interactions in the low dimensional semiconductor nanostructures and its response at multiple scales is a fundamental quest of modern science. [See ▶ Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology].

The spin phenomena in mesoscopic transport have moved only recently into the focus of attention, as one branch of the field of spintronics. The interplay between quantum coherence with confinement- or interaction-effects gives rise to a variety of unexpected spin phenomena in mesoscopic conductors and allows moreover to control and engineer the spin of the charge carriers. The spin interference is often the basis for spin-valves, spin-filters and detectors, and spin-switches. Their underlying mechanisms may gain relevance on the way to future semiconductor-based spin electronics. A quantitative and theoretical understanding of spin-dependent mesoscopic transport calls for developing efficient and flexible numerical algorithms.

The electronic and transport properties of quantum dot spin transistors are studied with emphasis on single-electron tunneling and shell structure using comprehensive modeling approach. For example, self-consistent calculation of the Poisson and Schrödinger equations within the spin-density-functional theory and the exact diagonalization of the many-body Schrödinger equation can be employed to describe spin transport properties of quantum dots [see ▶ Quantum Dot Spin Transistors, Self–consistent Simulation of]. With rising interest in spin-dependent transport, the interplay of the electron spin and charge degree of freedom has been exploited in a variety of spin interference devices. Charge and spin transport through phase-coherent conductors of mesoscopic scales carry imprints of wave interference as predominant and characteristic features [see ► Spin-Polarized Quantum Transport in Mesoscopic Conductors: Computational Concepts and Physical Phenomena]. Novel physical phenomena that may lead to improved memory devices and advances in quantum information processing are closely related to spin-orbit interactions. Hence, it is important to investigate a semiclassical wave-packet description of spin transport in the presence of electric fields by explaining how the microscopic motion of carriers gives rise to a spin current [see ► Semiclassical Spin Transport in Spin-Orbit Coupled Systems].

In the meantime, the theory of exchange-correlation energy is of great interest in modern density-functional many-body approaches. Hence, the spin dependent exchange-correlation energy in an interacting system such as the two-dimensional electron layers is of critical importance in the proper design of modern nano-structure devices and quantum-well lasers from a technological point of view [see ► Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers]. In addition, the time revolution of correlation functions in disordered spin systems is also worth to investigate [see ► Spin Dynamics in Disordered Solids].

Carbon nanotubes are referred to as the fabric of nanotechnology, and will play a central role in the future development of this technology. Understanding the properties of nanotubes, via computational simulation studies, has been one of the most intensive areas of research in physical sciences. Carbon nanotubes form the fourth allotrope of crystalline carbon after graphite, diamond, and a variety of caged-like fullerene molecules. Their mechanical properties make them stronger than steel, and their thermal conductivity is faster than copper. They have very exotic electronic-conduction properties by changing their geometry or by introducing topological defects into their structure and therefore, their electronic conductance can change from metals to semi-conductors. In general, carbon nanotubes show very different mechanical, thermal, electronic and optical properties.

The investigation into nanotube properties has prompted an intensive experimental and theoretical/computational research. One of the most active areas has involved the use of predictive computational modeling of their mechanical, thermal and mass transport properties [see ► Carbon Nanotubes, Thermo-mechanical and Transport Properties of]. One of the computational methods includes atomistic simulations using tight-binding molecular dynamics to study the nucleation and formation of carbon fullerenes and single-walled carbon nanotubes [see ► Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation]. Furthermore, the non-equilibrium Green's Function approach can be employed to study nanowire field effect transistors (FETs) such as a silicon nanowire FETs and carbon nanotube FETs [see ► Quantum Simulations of Ballistic Nanowire Field Effect Transistors]. Therefore, carbon nanotubes are viewed as highly relevant nanostructures with extensive potential applications.

The physical implementation of the quantum bit (or qubit) has been, and still is, the starting point of any proposal for quantum information processing device. The research of quantum computing focuses on qubits and quantum entanglement based on both a large spectrum of systems and solid-state devices. Hence, it is interesting to consider the practical realizability of quantum computers based on solid-state flying qubits and a simple quantum-gate network [see ► Charge-Based Solid-State Flying Qubits].

For device sizes in the range of tens of nanometers, the atomistic granularity of constituent materials cannot be neglected. Effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance. In an atomistic simulation, one takes into account both the atomistic/granular and quantum properties of the underlying nanostructure. The variety of geometries, materials, and doping configurations in semiconductor devices at the nanoscale suggests that a general Nanoelectronic Modeling tool (NEMO 3-D) is needed [see ► Multimillion Atom Simulations with Nemo3D].

There is a persistent need of miniaturization of machines and energy conversion devices for various engineering applications. In the field of robotics, there is a need to develop nano-scale actuators, motors and other machine components [see ► Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines]. These nano-scale robots, machines, and sensors can be used to deliver drugs to specific locations in the body, or detect individual cancer cells, or be used as molecular filters to separate minute particles from the environment. On the other hand, the field computation, which is a model of computation that information is represented primarily in fields, is to provide a mathematical language for describing information processing in the brain and in other natural and artificial systems [see ► Field Computation in Natural and Artificial Intelligence]. This field computation can be expected to provide an increasingly important analytical and intuitive framework for understanding massively parallel analog computation in natural and artificial intelligence.

In conclusion, there is no doubt that computational and theoretical nanoscience research such as quantum electronic and spintronic nanodevices will yield new applications in the near future for sensing, information processing, and quantum computation.

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis

Anton Kokalj
Department of Physical and Organic Chemistry,
Jožef Stefan Institute, Ljubljana, Slovenia

Article Outline

Glossary

Surface site is a position on the surface. High-symmetry positions are usually named as top, bridge, and hollow sites and designate positions over a surface atom, over two neighboring surface atoms, and over the void in between neighboring surface atoms, respectively.

Surface reaction center is a group of atoms on the surface where an elementary reaction occurs. The term reaction site will be used occasionally.

Adsorption is a process where molecules from the gas-phase or from liquid-solution bind to a solid or liquid surface. In this article, adsorption from the gas-phase to a solid surface is considered exclusively.

Chemisorption is a type of adsorption where a molecule binds to the surface through a direct chemical bond, whereas in physisorption the binding is due to van der Waals interaction.

Catalyst's activity is related to the rate of catalytic reaction, and can be expressed as amount of product made per unit time per active site (i.e. turnover frequency).

Catalyst's selectivity is a ratio between the amount of desired product obtained per amount of consumed reactants.

DFT Density Functional Theory

GGA Generalized Gradient Approximation

PES Potential Energy Surface

MEP Minimum Energy Path

NEB Nudged Elastic Band

fcc Face-Centered Cubic

IS Initial State

TS Transition State

FS Final State

BEP Brønsted–Evans–Polanyi

MD Molecular Dynamics

TM Transition Metal; term "early TM" designates transition metals with less than half-filled *d*-states, whereas "late TM" are those with more than half filled *d*-states.

DOS Density of States

STM Scanning Tunneling Microscopy

OMC Oxametallacycle

EO Ethylene epoxide

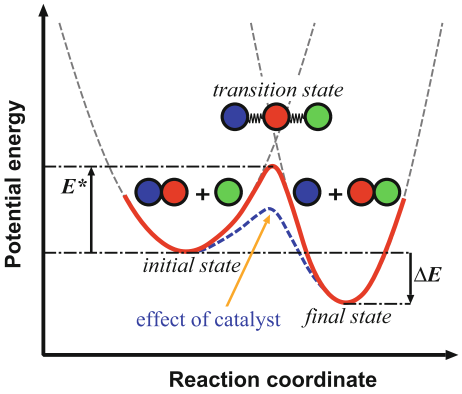
Ac Acetaldehyde

Definition of the Subject

First-principles (in Latin *ab initio*) quantum-mechanical-based computer simulations can deliver an atomic-level insight into elementary constituents of phenomena such as chemical reactions. The knowledge and understanding thus obtained may be used to propose new reaction-centers with on-demand tailored catalytic properties. The purpose of this article is to demonstrate with several examples how to gain such an insight and on this basis how to help design new *heterogeneous* reaction centers. In particular, surfaces of transition metals are considered, which are traditionally known as good heterogeneous catalysts. However, *ab initio* calculations of the properties of individual reaction centers present only the first part of the endeavor, because dynamical properties (e. g. overall reaction rates) are, in general, determined by the interplay between many elementary atomic-scale processes. The recipe for calculating the properties of such a dynamical system may therefore be accomplished by a hierarchy approach, where different computational methods are utilized for different time and lengths scales. For example, the information provided by the atomic-scale quantum-mechanical calculations can be utilized by statistical mechanics methods like kinetic Monte Carlo to follow the time evolution of the system. Such (statistical) methods are beyond the subject of this article and are just shortly outlined.

Introduction

A chemical reaction is a process of conversion of substances into other substances. Consider for example a simple chemical reaction, where molecules A and B react so as to form molecules C and D. The molecules A and B that enter into reaction are called reactants, and the molecules C and D are the reaction products. To understand the mechanism of reaction we need to envisage the atomic structure of the involved molecules and the trajectory during their inter-conversion. In general, during chemical reaction some bonds are broken and new bonds are formed. Perhaps the simplest picture of a chemical bond—a bond linking two atoms together—is that of a mechanical spring governed by Hook's law. This means that atoms "feel" a harmonic potential around equilibrium positions. Imagine for example that during chemical reaction an *a*–*b* bond breaks and *b*–*c* bond forms. This means that the *a*–*b* bond will be stretched first (the potential energy will increase) and eventually broken (point of maximum potential energy), and then a new *b*–*c* bond will start to form (the potential energy will decrease). This process is schematically shown in Fig. 1. Already this simple picture shows that chemical reaction is an activated process, that is, the system has to cross an energy barrier when passing from reactants to products. In reality the majority of chemical reactions are activated. Moreover, in many cases, the energy barrier of the reaction is so large that the reaction does not take place: more precisely the rate of reaction is too slow to be observed. Namely, the rate of chemical reaction depends exponentially on the energy barrier: the larger the barrier, the smaller is the rate. To accelerate a chemical reaction, a "third" substance, called a catalyst, is used. A catalyst is a substance that accelerates the rate of a chemical reaction, but is itself not consumed by the reaction. In terms of the above discussion of activated process, this means that the catalyst reduces the height of the activation barrier (see Fig. 1), for example, by means of providing a new reaction mechanism.



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 1
Reaction energy profile of a simple reaction (*red curve*): *ab* + *c* → *a* + *bc*. Around equilibrium positions (local minima) atoms feel harmonic potential (indicated by *two dashed gray parabolas*). As a consequence chemical reaction (bond-breaking and bond-making) will involve overcoming an energy barrier, *E. The effect of a catalyst on the chemical reaction is to reduce the *E** (*blue curve*) and correspondingly to enhance the rate of reaction**

A catalyst can be either homogeneous or heterogeneous. A homogeneous catalyst is in the same phase (solid, liquid, and gas), whereas heterogeneous catalyst is in a different phase with respect to reactants. A typical example of heterogeneous catalysis is where a solid catalyst is used to accelerate the gas-phase reaction. At present, about 90% of all chemicals are produced as heterogeneously catalyzed processes [62], where catalysts are solids, and provide a range of products such as fuels, fertilizers and plastics. Moreover heterogeneous catalysis is used to clean poisonous emissions from power plants, cars, and industrial production. Because of the technological and economical importance of catalysis, it makes it worthwhile to improve the catalysts even if the gain is as minor as only a few percentage points or less in performance. The efficiency of the production process strongly depends on the specific properties of the catalyst. A deeper knowledge of the interactions between the reactants and the catalyst and the understanding of the reaction mechanism would allow for the optimization of the production of many chemicals.

To understand the reaction mechanism of a given chemical reaction in detail, it is clear that the above-mentioned spring-picture of the chemical bonds holding the atoms within the molecule together will not suffice. Moreover, when a catalyst is used to accelerate the rate of reaction, the complexity of the system rises, because the number of degrees of freedom increases substantially. The progress attained in the last few decades in the field of computer simulations of matter at the atomic scale have opened new perspectives toward an atomic-scale understanding of the catalysts. In particular, electronic structure calculations, which for the description of solids are mainly based on Density Functional Theory (DFT), can provide information that are otherwise hard to obtain from experiment, such as the details of the reaction mechanisms and identification of the nature of the transition states. Therefore they help elucidate the factors that determine the activity and selectivity of catalysts, and consequently provide a way for their improvement.

Ab Initio Electronic Structure Calculations

The dynamics of a chemical reaction could, in principle, be treated by solving the time-dependent Schrödinger equation. In practice, this is not feasible (except maybe for a few trivial cases), and therefore several approximations are made. For example, one could proceed with a time-independent Schrödinger equation that is simplified with Born–Oppenheimer approximation. The time-independence stems from the fact, that, for example, the Hamiltonian (the energy operator) of an isolated molecule or crystal (in the absence of any external time-varying "*influence*") does not depend on time, whereas the Born–Oppenheimer approximation allows one to decouple the electronic motion from the motion of nuclei. This is made possible, because the mass of the nucleus is about three orders of magnitude heavier than the mass of the electrons and therefore the time scale of ionic dynamics is much larger than the electronic one. This means that one picks a given fixed configuration of nuclei (kinetic energy of nuclei is thus neglected) and then solves the electronic problem (hence the name *electronic structure theory*). Formally, this can be written as:

$$\widehat{H}_{\{\mathbf{R}\}}\Psi_{\{\mathbf{R}\}} = E(\{\mathbf{R}\})\Psi_{\{\mathbf{R}\}}\,, \tag{1}$$

where $\widehat{H}_{\{\mathbf{R}\}}$ and $\Psi_{\{\mathbf{R}\}}$ are the Hamiltonian and wave-function of n interacting electrons moving in the field of N fixed nuclei with coordinates $\{\mathbf{R}\}$. $E(\{\mathbf{R}\})$ is the energy of the Hamiltonian $\widehat{H}_{\{\mathbf{R}\}}$, and can be seen as the potential energy that the nuclei "feel" at configuration $\{\mathbf{R}\}$. As for the analysis of chemical reactions, it is often only the ground-state $E(\{\mathbf{R}\})$ that is of interest. The $E(\{\mathbf{R}\})$ is therefore a function of the coordinates of nuclei. Solving Eq. (1) for many different configurations of the nuclei, would allow for the construction of the potential-energy-surface (PES), that can be used to analyze the motion of nuclei. Once the PES is constructed the forces acting on atoms can be calculated as $\mathbf{F}_I = -\frac{\partial E(\{\mathbf{R}\})}{\partial \mathbf{R}_I}$ (I designates a given atom), and the dynamics (e. g. motion of atoms) can be studied, for example, by integrating the classical equation of motions—a method known as *ab initio* molecular dynamics (MD). However, the construction of PES is only feasible for very simple systems, for others *ab initio* MD consists of solving Eq. (1) at each snapshot during the *discretized* time evolution (t, $t+\delta t$, $t+2\delta t, \ldots$) of the system. At each snapshot the forces are calculated by Hellman–Feynman theorem as $\mathbf{F}_I = -\frac{\partial E(\{\mathbf{R}\})}{\partial \mathbf{R}_I} = -\langle\Psi_{\{\mathbf{R}\}}|\frac{\partial H_{\{\mathbf{R}\}}}{\partial \mathbf{R}_I}|\Psi_{\{\mathbf{R}\}}\rangle$, and the system is propagated by δt accordingly. Note that in such MD treatment the quantum effects of nuclei (e. g. tunneling) are neglected.

For practical purposes, even solving Eq. (1) is too complicated (analytical solution does not exist, except for a few trivial cases), and the n-electron equation is further simplified into a set of one-electron equations with the following form:

$$\left[-\tfrac{1}{2}\nabla^2 + \upsilon_{\text{eff}}(\mathbf{r})\right]\phi_i = \varepsilon_i\phi_i\,, \quad i = 1, n\,, \tag{2}$$

where $\upsilon_{\text{eff}}(\mathbf{r})$ is an effective one-electron potential (for notational simplicity, nonmagnetic insulator is considered). One possible scheme is the Hartree–Fock (HF) method, where the many-electron wave-function Ψ is approximated by an anti-symmetrized product (*Slater determinant*) of one-electron orbitals, ϕ_i, thus satisfying the Pauli exclusion principle according to which the wave-function has to be anti-symmetric with respect to exchange of two electrons. In HF method the effective potential $\upsilon_{\text{eff}}(\mathbf{r})$ is given by:

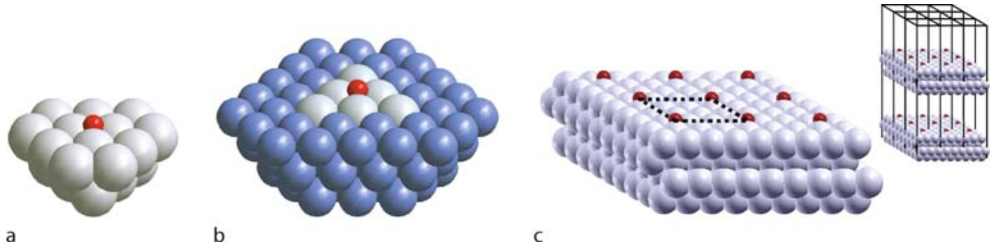
$$\widehat{\upsilon}^{\text{HF}}_{\text{eff}}(\mathbf{r}) = \upsilon(\mathbf{r}) + \int_{\mathbf{r}'} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' - \widehat{K}(\mathbf{r})\,. \tag{3}$$

The $\upsilon(\mathbf{r})$ is the *external* potential due to nuclei charges, $\rho(\mathbf{r}')$ is the electron density (the integral accounts for the electron-electron Coulomb interaction), and $\widehat{K}(\mathbf{r})$ is the exchange operator that arises due to "anti-symmetry" of the wave-function. The HF method, although computationally appealing, gives too poor results due to neglect of the correlation effects. The procedure for treating correlations is well established, but the corresponding wave-function-based methods are computationally very demanding [104].

Another computationally appealing method is provided by Density Functional Theory, where the electronic structure problem is simplified by realizing that the ground-state energy is a functional of the electron ground-state density [35]. The practical DFT approach also consists of solving a set of one-electron equations (*Kohn–Sham* equations [46]), similar to those of HF, but the DFT formalism—although formally exact—allows one to treat the correlation effects, in practice, in an approximate way. In DFT the effective potential $\upsilon_{\text{eff}}(\mathbf{r})$ is given by:

$$\upsilon^{\text{KS}}_{\text{eff}}(\mathbf{r}) = \upsilon(\mathbf{r}) + \int_{\mathbf{r}'} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + \upsilon_{\text{xc}}(\mathbf{r})\,, \tag{4}$$

where the $\upsilon_{\text{xc}}(\mathbf{r})$ is the exchange-correlation potential, which is given in terms of exchange-correlation (XC) energy functional $E_{\text{xc}}[\rho]$ as $\upsilon_{\text{xc}}(\mathbf{r}) = \delta E_{\text{xc}}[\rho]/\delta\rho(\mathbf{r})$. There are two types of approximations in DFT, depending on how



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 2
Three types of surface models: a cluster, b embedded cluster, and c slab model (surface unit cell is indicated by *dashed line*). The multi-slab flavor of the slab model is shown in the *inset* of c

the $E_{xc}[\rho]$ is approximated [79]. In the local-density-approximation (LDA) the XC functional depends solely on electron density (i. e. the space is partitioned into infinitesimally small cubes, which are treated as uniform electron gas), whereas the generalized-gradient-approximation (GGA) incorporates also the density gradient terms into the XC functional. The LDA usually gives poor estimates of activation energies (poorly describes the bond-breaking and bond-making), because it favors homogeneous systems. Because a more bonded system is more homogeneous, the LDA overestimates the binding energies, and correspondingly severely underestimates the activation energy (E^*) of bond-making processes (and vice versa for bond-breaking reactions). On the other hand, the GGA gives more accurate binding and activation energies. There are many flavors of GGA functionals, for the description of solids the most popular are those of Perdew [80,81].

Modeling the Surface

In order to describe a surface (or a catalyst) with a computer, one needs to construct an appropriate model. Such models consists of a limited number of atoms, because it is not possible to treat a macroscopic number of (inequivalent) atoms. Three types of surface models are used (see Fig. 2):

Cluster model: surface is represented by a cluster of atoms that is obtained by a "finite cut" from an ideal lattice. This model builds on the assumption that adsorption is a phenomenon of strong local character. The cluster models of metallic surfaces are prone to cluster-size effects [34,86,109], in particular, the adsorption energy is known to undergo large variations as a function of the cluster size and shape [34,86].

Embedding scheme: embedding schemes divide the adsorbate/substrate system into at least two zones: (i) the cluster region and (ii) an outer region. The essence of embedding is to correctly connect the cluster to the outer part of the model so as to diminish the size effects of the cluster model. The most sophisticated embedding schemes are based on Green function formalism [6,24].

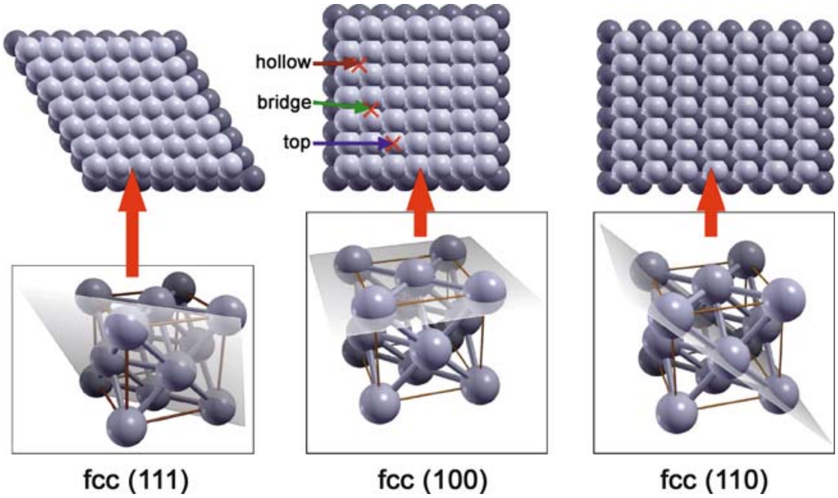
Periodic slab model: describes the surface by an infinite periodic slab in the *xy*-direction with finite thickness in the *z*-direction. The structure (and description) of such a slab is entirely given by a surface unit cell. There are two flavors: a single slab and a multi-slab model. The latter is actually an infinite 3D model, where individual slabs are separated by vacuum layers that are thick enough to prevent adjacent slab–slab interaction.

A given surface can be cut from the bulk crystal with many possible orientations. Miller indices are used to designate the surface plane by three integers as (*hkl*). These indices denote a plane that intercepts the three lattice vectors at $1/h$, $1/k$, and $1/l$, respectively. Figure 3 shows the three low-Miller index surfaces of an fcc (face-centered-cubic) crystal, which are (100), (110), and (111).

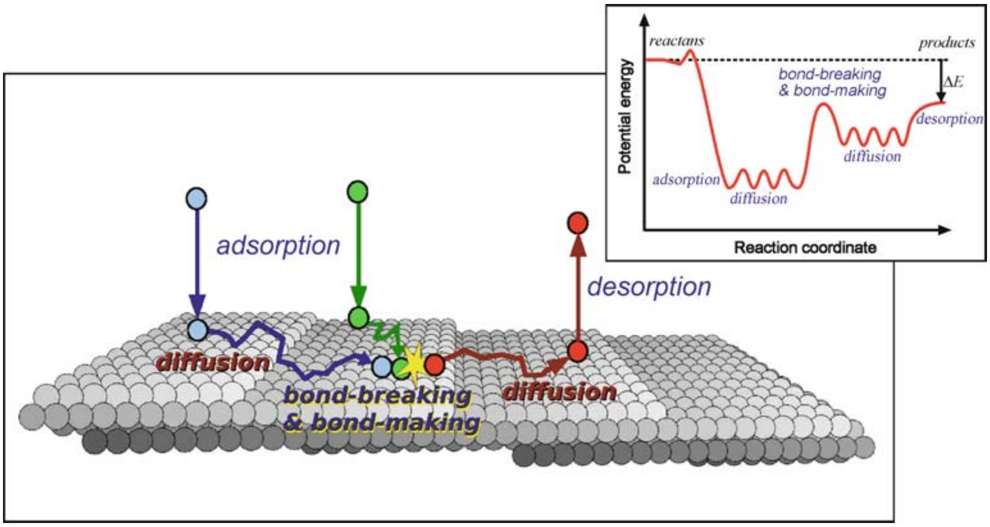
Description of Activated Processes

As mentioned in the Introduction, a chemical reaction is an activated process, and reactants have to overcome an energy barrier (E^*) during their conversion to products. The transition over the energy barrier occurs on average once per $t_{\text{slow}} \approx t_{\text{fast}} e^{E^*/kT}$, where t_{fast} is the time-scale characteristic to microscopic dynamics, and kT is the thermal energy. Usually $E^* \gg kT$ and the two time-scales differ by many orders of magnitude (the transition itself is very fast but it occurs rarely). Hence activated processes cannot be studied directly by molecular-dynamics (MD) simulations, as this would require an unfeasible number of MD integration steps before a single transition would occur (the time step used in MD must be smaller than the t_{fast}).

In general, there are several energy barriers involved in a chemical reaction. In this case the description of the overall chemical reaction is decomposed into several elementary steps, where each elementary step is associated



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 3
Three low Miller index surfaces of an fcc (face-centered-cubic) crystal: (111), (100), and (110). In the *bottom row* of plots the direction of the plane used to cut the corresponding surfaces is indicated. The positions of high-symmetry surface sites (*hollow*, *bridge*, and *top*) are also shown on the (100) surface



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 4
Chemical reactions consist of several elementary steps. The Figure shows a typical mechanism of surface catalyzed reaction (i. e. Langmuir–Hinshelwood mechanism), which consists of the following reaction steps: adsorption of reactants, their diffusion on the surface, one or more elementary steps where some bonds are broken and new ones are formed so as to form the product molecules, which eventually desorb from the surface. The *inset* shows the associated reaction energy profile

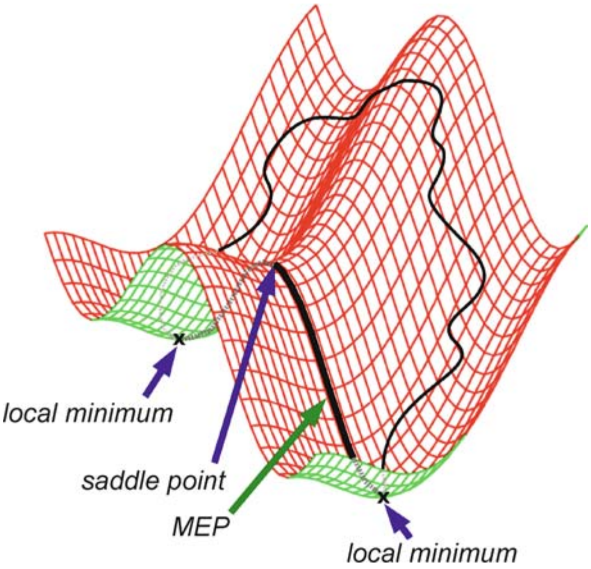
with (at most) one energy barrier. Figure 4 shows elementary steps involved in a typical surface-catalyzed reaction. Therefore, one needs to consider each elementary step in order to understand the overall chemical reaction. The hierarchical approach to the problem would consist of first analyzing each elementary step individually by electronic structure calculations, and then to assemble them into an overall picture by other methods, such as, for example, kinetic Monte Carlo (Sect. "Putting It All Together").

Let us consider an activated elementary step, such as depicted in Fig. 1. A stable molecular configuration on the left side of the barrier is called initial-state (IS) and that on the right side the final-state (FS) of an elementary step, whereas the point of the highest energy along the path from IS to FS is called a "transition state" (TS), and its energy difference with respect to IS, is the activation energy, E^*:

$$E^* = E_{\mathrm{TS}} - E_{\mathrm{IS}} \,. \tag{5}$$

But for an isolated diatomic molecule, the potential-energy-surface is multidimensional, hence there is an infinite number of trajectories going from IS to FS, and among there is also the minimum-energy-path (MEP), which is characterized by vanishing force components orthogonal to the path. The MEP's tangent is therefore parallel to the force vector and the configuration of the highest energy along the MEP is the TS—a first-order stationary point (see Fig. 5).

The calculation of PES is, due to its multidimensional nature, computationally too demanding to be sampled accurately but for a few simple cases. However, in many instances it turns out that the knowledge of the stationary points on the PES and its shape around them is sufficient to calculate reaction rates. Accurate estimates of transition rates, can be obtained, for example, by using a statistical approach known as *transition-state theory* (TST) [16,18]. According to TST the (forward) rate constant depends exponentially on the activation energy, E^*, and is given by a van't Hoff–Arrhenius type expression:



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 5

Two-dimensional potential-energy-surface (PES). Two trajectories are shown linking the two local minima: one of them is the minimum-energy-path (MEP) that is characterized by the vanishing force components orthogonal to the path and passes through the saddle point. In an elementary chemical reaction, reactant (initial-state, IS) and product (final-state, FS) are associated with local minima on the PES, and the reactant is transformed into the product by proceeding along a given trajectory on the PES that links the two minima. Statistically, the most probable trajectory displays the smallest barrier, hence it passes through the saddle point (first-order stationary point). The configuration at the saddle point is called the transition state (TS)

$$k_{\mathrm{f}} = \nu \mathrm{e}^{-E^*/kT} \,, \tag{6}$$

where k is a Boltzmann constant (please note that k and k_{f} label two distinct quantities), T is temperature, and ν is a prefactor, which depends on the shape of the PES around the stationary points. The letter "f" in the k_{f} indicates that this is the rate-constant for the forward direction of the elementary step, i. e. from IS to FS. There will also be backward jumps from FS to IS, and the net rate, r, of the elementary step will be given by the difference between the forward and backward rates, $r = r_{\mathrm{f}} - r_{\mathrm{b}}$. The TST also provides an insight into the physical origin of prefactor ν. In the harmonic TST approximation, where the PES around the IS and TS is expanded in normal modes, the prefactor ν can be described in terms of the vibration frequencies ν_i of IS and TS, and Eq. (6) can be written as:

$$k_{\mathrm{f}} = \frac{\prod_i^{3N} \nu_i^{\mathrm{IS}}}{\prod_i^{3N-1} \nu_i^{\mathrm{TS}}} \mathrm{e}^{-E^*/kT} \,. \tag{7}$$

Note that the product in the denominator is only over $3N - 1$ frequencies (N being the number of atoms), because the TS's normal mode tangent to reaction coordinate cancels out in the derivation of TST. On the basis of (harmonic) TST it is therefore clear why the knowledge of the stationary points on the PES and its shape around them suffices (in many cases) to calculate reaction rate. In the one-dimensional case, Eq. (7) would reduce to $k_{\mathrm{f}} = \nu^{\mathrm{IS}} \mathrm{e}^{-E^*/kT}$, where ν^{IS} is the oscillator frequency around the IS. Bond-stretching frequencies are on the order of 10^{12} to $10^{13}\mathrm{s}^{-1}$, and this gives an idea about the typical value of the prefactor in Eq. (6).

Identification of Transition-States

The dependence of the rate constant on the E^* is exponential and only linear on the prefactor ν, Eq. (6). It is precisely for this reason why the identification of the transition state is so crucial: knowing its structure and understanding its chemistry would allow for the design of a reaction site that would lower the energy of TS and consequently increase the rate of the elementary step. The premise behind this is that by "chemically" acting on the reaction-center, the prefactor would not change substantially, and the exponential dependence on E^* would prevail in determining the rate. Moreover, it is much more intuitive to act on the energetic stability of the TS than on the "shape" of the PES around the IS and TS. As for the overall reaction rate, one would consequently need to lower the energy barrier for all elementary steps to increase the overall rate, but it turns out that in catalytic reactions, some barriers are anti-correlated, that is, lowering a barrier for one elementary step increases the barrier for some other elementary step.

At this point one could ask: why bother with computer atomistic simulations to identify transition states involved in given chemical reactions? The reason is that transition states are very hard to identify experimentally, because of their too short lifetimes. Moreover, the identification of transition states is quite demanding even by computer simulations: usually at least one order of magnitude heavier than the identification of equilibrium configurations (local minima on PES). The reason is that the methods based solely on the first derivatives of PES—the forces —do not converge to saddle points. Among the most popular methods for the identification of saddle points is the nudged-elastic-band method (NEB) [32,33], where instead of searching solely for TS, one considers a path connecting the IS and FS. This path is then discretized into several structures called images. The images are connected by "springs" to prevent them from sliding down the PES into the nearest local minimum during the optimization procedure (the spring forces are allowed to act parallel to the path and the true forces orthogonal to it). The NEB method therefore converges the path to the MEP. The highest point along a stable MEP—which by definition is a saddle point (a first-order stationary point)—is then found by allowing one of the images to "climb up" until the forces vanish. Actually the NEB will converge to MEP's whenever these are stable (i. e. when the curvature of the PES on hyperplanes orthogonal to the MEP itself is everywhere positive). When the MEP's are not stable different techniques based on the use of collective variables should be preferred [110].

Brønsted–Evans–Polanyi Relationship

Because the PES is such an illusive object, chemists have long attempted to develop some concepts that would surpass the explicit knowledge of the PES and even of the transition states. Although these are mainly empirical, they are useful and can be used as guides or in explanatory purposes when searching for improved reaction sites. Various relationships have been proposed, and they are derived from some assumptions of the shape of the potential-energy-surface. Perhaps the most widely used, at least in the field of heterogeneous catalysis, is the *Brønsted–Evans–Polanyi* (BEP) relationship [7,17,62], according to which the activation energy for a set of related elementary reactions correlates linearly with the reaction energy, ΔE (ΔE is defined graphically in Figs. 1 and 7):

$$E^* = a + b\Delta E \, . \tag{8}$$

This relation tells us that the smaller ΔE is, the smaller will be the activation energy. Small here means negative, that is, exothermic or thermodynamically favored. In Eq. (8) there are two parameters, a and b, which give the slope and the intercept of the BEP line. These two parameters were given various names, e. g. the slope is called the reaction coefficient, whereas the intercept is the intrinsic barrier of reaction [62]. It has been shown that the BEP relation holds well for dissociation reactions at surfaces [3,71], because the transition state is *late*, product-like, and variations in the stability of dissociated products are likely to reflect the stability of the transition state. Analogously, for an *early*, reactant-like, transition state the variation in the stability of the reactant will correspondingly affect the stability of the transition state, but the barrier will be little affected. This is schematically shown in Fig. 6. For this reason the BEP slope is large for reactions with *late* transition state, and small for *early* transition state.

Description of the Overall Reaction

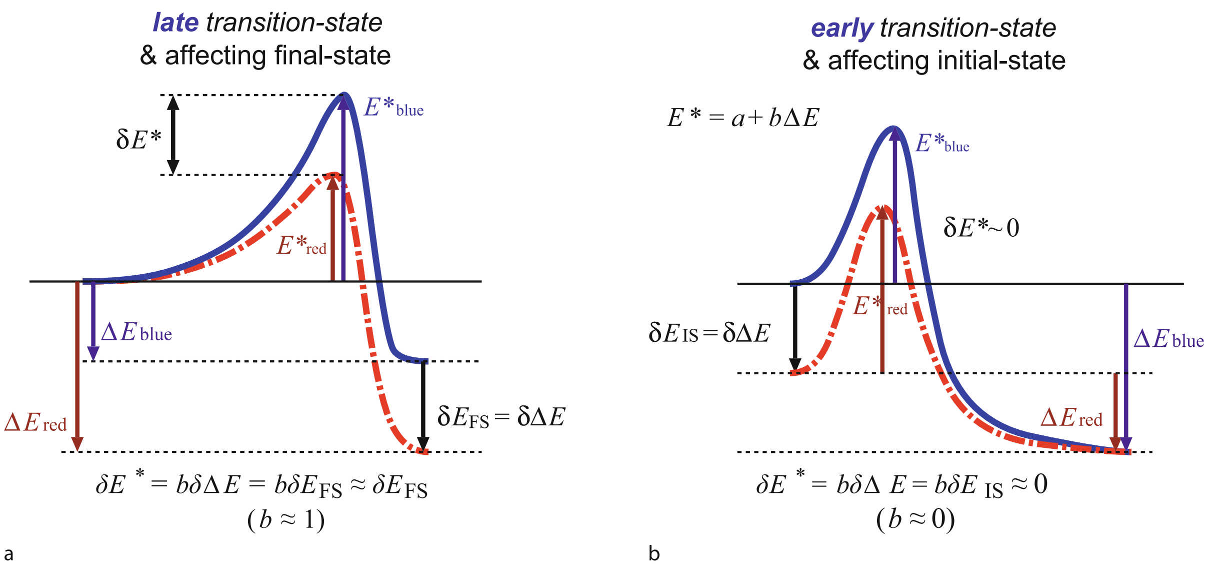
Decomposition into Elementary Steps

A heterogeneous catalytic reaction involves several elementary steps, such as the adsorption of the gas-phase reactants on the solid surface, their diffusion on the surface, one or more elementary steps involving breaking some of the reactant chemical bonds and making new ones so as to form the product molecules, which eventually desorb from the surface (see Fig. 4). Consider, for example, ammonia (NH_3) synthesis, which is among the largest industrial catalytic processes (Haber–Bosch process) and one of the most studied [94]. The overall reaction is:

$$N_2 + 3H_2 \rightarrow 2NH_3 \, , \tag{9}$$

and it consists of the following elementary steps [94] (omitting diffusion of adsorbed species):

$$\begin{array}{lr} 1. & N_{2(g)} \rightarrow N_{2(ad)} \\ 2. & N_{2(ad)} \rightarrow 2N_{(ad)} \\ 3. & H_{2(g)} \rightarrow 2H_{(ad)} \\ 4. & N_{(ad)} + H_{(ad)} \rightarrow NH_{(ad)} \\ 5. & NH_{(ad)} + H_{(ad)} \rightarrow NH_{2(ad)} \\ 6. & NH_{2(ad)} + H_{(ad)} \rightarrow NH_{3(ad)} \\ 7. & NH_{3(ad)} \rightarrow NH_{3(g)} \, , \end{array} \tag{10}$$



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 6
Effect on the variation of relative stability of final (FS) and initial state (IS) on the activation energy for *late*- and *early*-type transition state (TS), respectively, according to the BEP relation. a For *late* TS the stabilization of FS stabilizes the TS leading to a reduction of the activation energy, *E. b For *early* TS the stabilization of IS stabilizes the TS, but this affects the *E** very little**

where labels (g) and (ad) designate the gas-phase and adsorbed species, respectively. This set of elementary steps can be classified into three groups:

(i) dissociative adsorption of nitrogen (N_2) and hydrogen (H_2), Eqs. (10.1)–(10.3);
(ii) reactions of the dissociated N and H species to form the product NH_3 species: in this step, the adsorbed N is successively hydrogenated, Eqs. (10.4)–(10.6);
(iii) desorption of the product ammonia molecules, Eq. (10.7).

Sabatier Principle

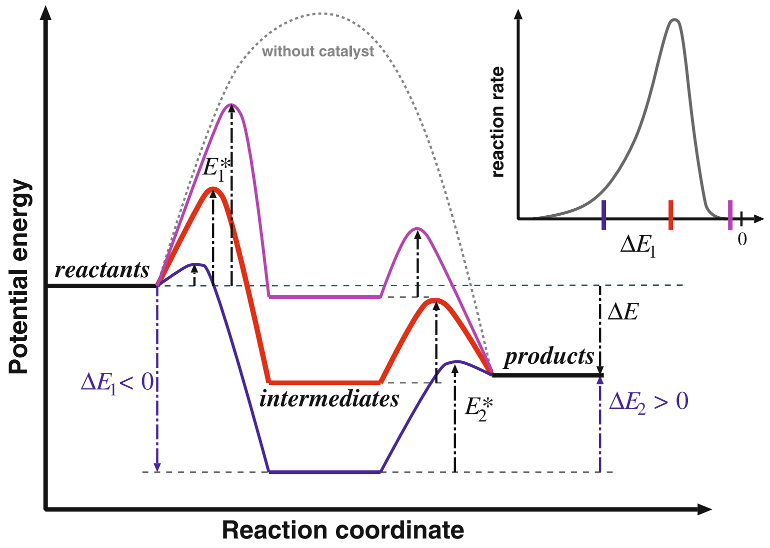
As for the above NH_3 synthesis, a good catalyst must facilitate both the dissociation of reactant N_2 and H_2 molecules, as well as the formation of the product molecules. Because the reactants are the molecules in the gas phase, the reaction energy of dissociation steps is given by the dissociative chemisorption energies. According to the BEP relation, Eq. (8), the more negative (exothermic) is the chemisorption energy the smaller will be the activation energy, and the easier (faster) the formation of intermediates. This indicates that a very reactive surface should be used that forms strong bonds with the dissociated species. However, there is an opposite trend for the formation of product molecules, see Fig. 7 (it is worth mentioning that the BEP slope for association reactions tends to be much smaller than for dissociation reactions [71]). In addition, if the surface is too reactive intermediates will fully cover it thus inhibiting the reaction. Therefore, weakly chemisorbed intermediates would be required to form the product molecules. The best catalyst will be therefore a compromise with an intermediate chemisorption energy. This is the *principle of Sabatier*, which results in the so-called volcano activity plots (see inset of Fig. 7).

Considerations presented above demonstrate why the surface acts as a catalyst: the main effect is to stabilize the intermediates, which reduces the activation energies accordingly. But the stabilization should not be overdone, because then the reaction rate would decrease again.

Putting It All Together

From the preceding section it is rather obvious that the dynamics of the overall reaction may develop pattern(s), that may not be guesses merely by considering elementary steps individually. Namely, macroscopic dynamical properties are, in general, determined by the interplay between many elementary atomic-scale processes. For this reason, after the individual elementary steps have been quantified, they need to be put into a broader context that will link them together so as to construct a coherent picture of the overall reaction.

Mean-Field Description: Rate Equations One could setup rate equations for all elementary steps. The procedure for doing that is well established [15,61,102]. Consider for example, an elementary step where adsorbed



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 7
Surface acts as the catalyst, because it stabilizes the intermediates, and therefore decreases the activation energy. The more they are stabilized (the more negative ΔE_1) the smaller is the activation energy for their formation (*left barrier*, E_1^*), but the barrier for their conversion into products will display the opposite trend (*right barrier*, E_2^*). The optimum is therefore a compromise between the two anti-correlated effects, hence the catalyst should bind intermediates moderately (*Sabatier principle*). *Inset*: the rate of reaction displays a *volcano* dependence on the intermediate chemisorption energy (ΔE_1)

species A and B react so as to form adsorbed species AB:

$$A_{(ad)} + B_{(ad)} \rightleftharpoons AB_{(ad)} \, . \tag{11}$$

Naively speaking, the forward rate, r_f, of this step will be given by the "*probability of* A *and* B *to meet at the same site*" times the "*probability to overcome the barrier*" per unit time. If we define the coverage of species x as $\theta_x = N_x/N$, where N_x is the number of adsorbed x molecules and N is the number of all surface sites, then the first probability is given by the product of the two coverages, $\theta_A\theta_B$, whereas the second probability is given by the forward rate constant, k_f. In this case the rate is normalized to the number of "jumps" per site per unit time. The reverse rate, r_b, will be analogously given by $\theta_{AB} \times k_b$, where k_b is the rate constant for the backward direction (activation energy in the backward direction is $E^*_{bck} = E^* - \Delta E$, e. g. see Fig. 1). The net rate is therefore given by:

$$\begin{aligned} r = r_f - r_b &= \theta_A\theta_B k_f - \theta_{AB} k_b \\ &= \theta_A\theta_B\left[\nu_f e^{-E^*/kT}\right] - \theta_{AB}\left[\nu_b e^{-(E^*-\Delta E)/kT}\right] , \end{aligned} \tag{12}$$

where ν_f and ν_b designate prefactors for the forward and backward direction, respectively. Note, however, that if the *probability* is expressed as $\theta_A\theta_B$ then A and B are treated as independent. Therefore, in this scheme the lateral interactions between the adsorbed species can be treated only by mean-field description through dependence of the activation energies and prefactors on the coverage. After rate equations are constructed for all elementary steps the so-obtained set of equations needs to be solved. The resulting equations may be highly nonlinear and very interesting patterns such as kinetic oscillations and chaos may develop [116]. In order to simplify the treatment several approximations may be applied, such as steady-state (SSA) and quasi-equilibrium approximation (QEA) [61].

Kinetic Monte Carlo A more detailed description of the overall reaction could be obtained with a *kinetic Monte Carlo* (kMC) simulation [19,115], but it requires larger computational effort. In kMC the surface is usually mapped onto a two-dimensional lattice and the adsorbate positions are associated with lattice sites (adsorption for example is simulated by random "arrivals" of molecules to the lattice sites). At each simulation step the rates of all considered elementary processes are evaluated. A given process is then executed by a random selection with appropriately weighted probability (i. e. the rate of jumps is proportional to the associated barriers). In kMC the time evolution is coarse-grained to time scale appropriate for the rare activated events. At each step a detailed atomic configuration is known, and the activation energies can be calculated taking into account proper local environment.

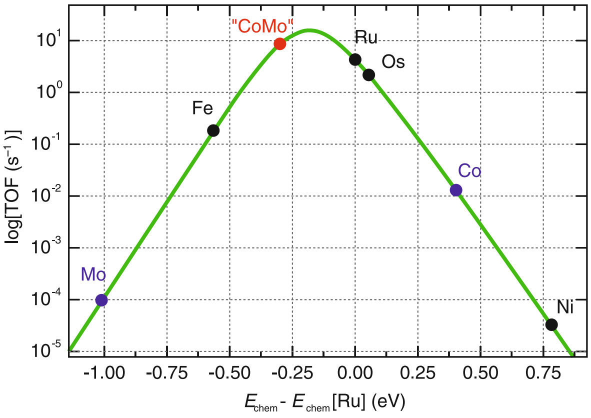
In kMC the number of particles as well as the number of steps is very large, and to calculate all the interactions and activation energies on the fly with ab-initio methods is far from feasible. Therefore, a model Hamiltonian (e.g. *lattice gas* Hamiltonian) is constructed that allows for fast on the fly calculation of required energies. A model Hamiltonian may be built by expanding it into various interaction terms that are determined from a number of *ab initio* pre-calculated structures [85].

Rate-Limiting Step An analysis of the rates of the involved elementary steps reveals that not all the elementary steps are equally significant in determining the rate of the overall chemical reaction. The step that most significantly determines the rate of the overall chemical reaction is called the rate-determining step. This may imply that the rate-determining step is the "slowest" among all elementary steps, however under steady state conditions, by definition, the rates of all the steps will be equal. For example, in the ammonia synthesis mentioned above, the dissociation of N_2 is rate-determining. There may be more than one kinetically relevant steps in a given chemical reaction. From Eq. (12), we see that the rate depends exponentially on activation energy and temperature, whereas the dependence on the other parameters is weaker. Therefore, if a given elementary step has much larger E^* than the others, it is likely to be rate-determining. Note also that under different experimental conditions (T, P) different steps may be rate-determining. For example, under large enough T the E^*/kT ratio will become small enough, and other factors, such as, the availability of sites and partial pressures of reactants may prevail.

Optimum Chemisorption Energy For a class of heterogeneous reactions, where dissociative adsorption is the rate-determining step, like the ammonia synthesis considered above, an analysis of the reaction rates gives a universal result that the optimum chemisorption energy should be, broadly speaking, in the range from -1.0 to -2.0 eV per molecule [3,76]. This universality is a consequence of the BEP dependence of the activation energy on the chemisorption energy. This implies that the stronger the bond in the diatomic molecule, the more reactive will be the "optimum" catalyst. For example, the bond of N_2 is stronger than that of oxygen, O_2, therefore a good catalyst for reactions involving dissociation of N_2 is more reactive than the catalyst for reactions involving the O_2 dissociation. The catalysts for ammonia synthesis are metals such as ruthenium (Ru) and iron (Fe), whereas oxidation reactions typically involve less reactive silver (Ag), palladium (Pd) and platinum (Pt) catalysts.

The general result presented in the paragraph above calls for an in-depth understanding of the factors that govern the dissociative chemisorption energies and these factors will be underlined in Sect. "Theory and Trends of Chemisorption and Reactivity on Transition Metals". Theory and computer simulations have played an important role here. They helped to provide an understanding of chemisorption phenomenon and moreover computer simulations based on DFT have provided systematic databases of dissociative chemisorption energies.

Although the range of optimum chemisorption energies suggested above is relatively broad, it turns out that for a given diatomic molecule this range typically involves only a few elemental metals, and none of them may even lie close to the volcano curve maximum [76]. Hence a fine-tuning of the chemisorption energy would be required so as to move toward the volcano maximum. A simple rational approach would be to form a bimetallic alloy by simple interpolation [39], i. e., by mixing a given metal that lies on the left side of the volcano peak (too strong chemisorption energy, E_{chem}) with the one on the right (too weak E_{chem}). As for the ammonia synthesis discussed above, mixing Mo—which binds N too strongly—with Co —which binds N too weakly—results in a catalyst that is closer to the volcano maximum and therefore more active than the best elemental metals Fe, Ru, and Os [39] (see Fig. 8). This is precisely what was found experimentally [38,47]. Note that both Mo and Co display lower activity than Fe, Ru, and



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 8

Calculated volcano plot [39]: catalyst's activity (i. e. turnover frequency, TOF) for ammonia synthesis as a function of chemisorption energy of nitrogen with respect to Ru, $E_{\text{chem}} - E_{\text{chem}}$[Ru] (eV per N_2 molecule). Curve is calculated by microkinetic model assuming linear BEP relation between nitrogen chemisorption energy and the activation energy for N_2 dissociation. On the basis of results from [39]

Os, and without the knowledge of the principal factor that governs the activity, this would only be guessed by error-and-trial.

For the mixed catalyst to have the interpolated properties, several requirements have to be fulfilled. In particular the constituent metals should not segregate. More issues about alloys will be discussed in Sect. "Theory and Trends of Chemisorption and Reactivity on Transition Metals".

Selectivity Issues

Tuning solely the reactivity of a catalyst—for example according to the above-mentioned Sabatier volcano plot principle—in order to optimize its activity is in many cases not sufficient. Chemical reactions are accompanied with the formation of (unwanted) byproducts. In some cases the selectivity issues may be crucial (selectivity can be defined as a ratio between the amount of desired product obtained per amount of consumed reactants), because some of the reaction byproducts may even poison the catalyst, such as, for example, coke formation during the syngas production (e. g. see Sect. "Improvement of Steam-Reforming Catalyst"). Therefore, it is clear that improving the catalyst's selectivity is of great importance.

DFT simulations can help improve the catalyst's selectivity by disentangling the mechanisms that govern the reactivity of catalysts toward competitive reaction pathways. Consider the formation of two competitive species from a given adsorbed intermediate:

$$\begin{aligned} &A_{(\mathrm{ad})} \to \mathrm{B} \\ &A_{(\mathrm{ad})} \to \mathrm{C}\,. \end{aligned} \tag{13}$$

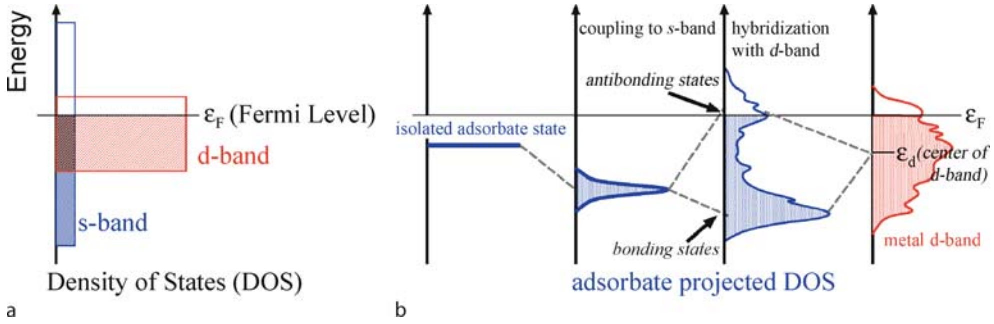
If B and C are adsorbed species, then the BEP relation suggests that in order to favor the formation of, for example, B species the reaction center should be designed such as to relatively enhance the magnitude of adsorption energy of B with respect to C. However, if the species B and C are gas-phase molecules, or if the reactions do not obey the BEP relation, the design of the proper reaction center that will favor the formation of wanted species will be more difficult. A possible strategy is to act on stabilities of involved transition states by designing such a reaction center that will stabilize the TS of the desired reaction channel with respect to those of undesired pathways. To do so the transition states have to be identified: knowing their structure and understanding their chemistry would allow one to act on their stability. In this respect, the understanding of the factors that determine the adsorbate–substrate bonding should be most helpful. For this reason a simple chemisorption model and its applicability under different circumstances will be presented below. Then it will be demonstrated with a few examples how quantum mechanical computer simulations can be used to help in designing novel reaction-centers with tailored properties.

Theory and Trends of Chemisorption and Reactivity on Transition Metals

Ab initio electronic structure methods, such as DFT, allow one to calculate the properties needed for the quantitative description of chemical reactions, such as equilibrium and transition state structures, their energies, and vibrational properties. On the other hand, one needs a simple model, which would describe in terms of a few parameters the trends of these properties, and therefore provide understanding of the most important factors that govern them.

Below a short description of the Hammer–Nørskov chemisorption model [28,30] is given, which is based on the Newns–Anderson chemisorption theory [75]. This model provides a simple description of the chemisorption trends on transition metal (TM) surfaces. Consider, for example, that a single one-electron state of an adsorbed atom interacts with a TM surface. This state will interact with all the valence electron states of the surface metal atoms, which can be described with *s*- and *d*-band (i. e. valence electrons of a given (isolated) TM atom are described by discrete *s*- and *d*-states, but in a solid these states form *s*- and *d*-band). The former is very broad and approximately half filled for all TM (containing one electron/TM-atom). The *d*-band is instead much narrower, as shown in Fig. 9a. The delocalized *s*-states interact weakly with the adsorbate state, resulting in its broadening, whereas the localized *d*-states (which can be approximately considered as molecular state) interact strongly forming the adsorbate–surface bonding- and antibonding-states. This is schematically shown in Fig. 9b. Because the *s*-band and its interaction with the adsorbate is similar for all TM, the variation in the adsorbate–surface bonding is described merely by considering the *d*-band in the Hammer–Nørskov model. The variation of the adsorbate–surface bonding is determined mainly by two effects: (i) the amount of empty anti-bonding adsorbate–surface states; note that anti-bonding interaction is repulsive, so the more these states are empty the stronger is the adsorbate–surface interaction, and (ii) Pauli type repulsion arising from the orthogonalization of the overlapping adsorbate and metal states. This term is proportional to the square of coupling matrix elements between adsorbate state and metal *d*-states.

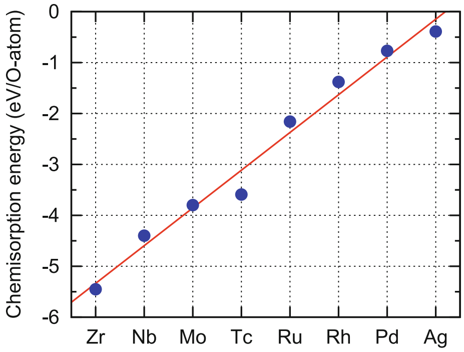
As for the effect (i), the amount of the empty antibonding states is related to the position of the *d*-band with respect to the Fermi level, ε_F (Fermi level is the energy of the highest occupied state in the metal). In the Hammer–



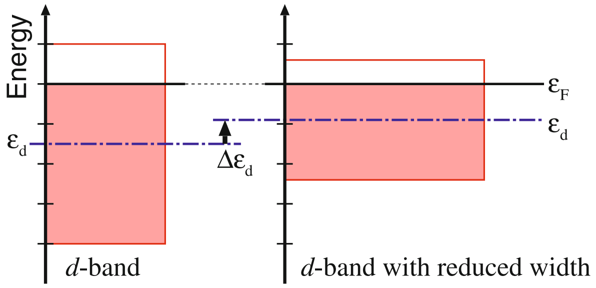
Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 9
a Electronic structure of transition metal (TM); b schematic illustration of adsorbate interaction with TM surface

Nørskov model this position is described by the center of the *d*-band, ε_d. The higher is the ε_d with respect to ε_F the more empty are the antibonding states and the stronger is the adsorbate–surface bond. This explains why the reactivity of TM decreases from left to right in the Periodic Table: the more the *d*-band is occupied the more below the Fermi level it is (lower ε_d). Figure 10 shows the trend of the DFT calculated chemisorption energies of oxygen on close packed surfaces of 4*d* transition metals. It is due to this trend and the Sabatier principle that mainly *late* transition metals (i. e. those with more than a half filled *d*-band) are catalytically interesting, because *early* transition metals (i. e. those with less than a half filled *d*-band) bind adsorbates too strongly. The reactivity also reduces when going down from the 3*d* to 5*d* row in the Periodic Table, because the latter states are more extended and therefore the effect (ii) is more repulsive.

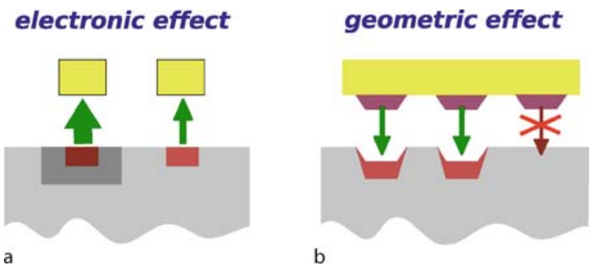
For a given adsorbate, its interaction to the surface can be affected not only by changing the metal (as shown in Fig. 10), but also by other means that are more subtle, and can be used to fine-tune the adsorbate–surface interaction. Consider what happens to the reactivity of a given *late* transition metal if the width of the *d*-band is reduced by some means. DFT calculations reveal that in such cases no charging or decharging of *d*-states occurs [30], and in order to keep the number of *d*-electrons fixed the *d*-band has to move up causing an up-shift of its center thus increasing the adsorbate–surface interaction. This is schematically presented in Fig. 11. In general, the less *effectively* a given TM atom is bonded the smaller is the width of its local *d*-band (the density-of-states projected to *d*-states of the considered metal atom), resulting in increased reactivity for late TM. The variation of the width of the *d*-band can be achieved, for example, by varying the coordination of surface atoms or by strain. Consider for example (111) and (100) low Miller index surfaces of an fcc metal (these two surfaces are shown in Fig. 3): the former has nine-fold coordinated surface atoms, whereas the latter has eight-



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 10
DFT calculated dissociative chemisorption energies of oxygen with respect to gas-phase O_2 molecule (in eV per O-atom) on closed-packed surfaces of 4*d* transition metals. On the basis of results from Refs. [13,54]. The *line* is drawn to guide the eye



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 11
Effect of the reduced band width on the center of the *d*-band for *late* (more than a half filled *d*-band) transition metal: in order to keep the number of electrons fixed (number of electrons is given by the colored area below ε_F), the *d*-band has to shift up, resulting in the $\Delta\varepsilon_d$ shift of the center of the *d*-band, ε_d. The effect would be the opposite for *early* (less than a half filled *d*-band) transition metal: the ε_d would be down-shifted



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 12
Schematic representation of electronic and geometric effects. a The adsorbate–surface interaction can be tuned by changing the electronic properties of the binding site by acting on its environment (electronic effect). b The adsorbate–surface interaction is hindered if the ensemble of the active binding site on the surface does not match the "geometric" requirements of the adsorbate (geometric effect)

fold coordinated surface atoms. Surface atoms of the (111) facet are therefore effectively more bonded than those of the (100) facet, and therefore we may anticipate that the former binds simple adsorbates less than the latter.

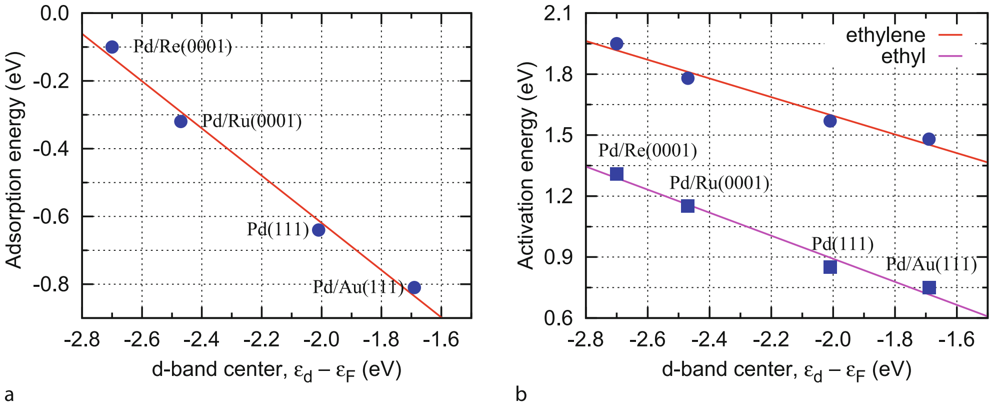
Considerations presented above were made for an atom (with a single state) interacting with the TM surface, but the arguments can be generalized to molecules by considering several adsorbate states interacting with the surface. The above *electronic* effects are still valid (e. g. see Fig. 13 for the correlation between the *d*-band center and molecular adsorption and activation energies), but in addition to that there is another effect, which is related to the geometry of the molecule and surface. Indeed, the two types of effects are termed *electronic* (or ligand) and *geometric* (or ensemble) effects, and provide a basic framework for tuning the reactivity of surface reaction centers. The way the two effects work is presented schematically in Fig. 12, and they may be described as:

- Electronic (ligand) effect is due to modifications of catalytic properties due to the interaction of a surface atom that the molecule binds to with its neighbors (e. g. shift of the d-band center).
- Geometric (ensemble) effect is related to the number of active surface atoms required for a given molecule to bind.

Above we implicitly assume that the larger the adsorbate–surface bond strength the larger the reactivity of the surface (i. e. the two terms were used casually; here surface reactivity toward dissociation reactions is meant). This can by justified by the BEP relation and also by the fact that the interaction of TS with the surface can be treated like a molecule–surface interaction, hence the stronger is this interaction the lower is the energy barrier.

Special Active Sites at Surfaces

What we have not mentioned explicitly so far is that not all sites on any real surface are alike. For example, lattice imperfections are always present on surfaces. This is particularly the case for industrial catalysts which consist of small dispersed particles on a ceramic support and display a large variety and concentration of defects [99]. It may be anticipated on the basis of the *d*-band model that such structural variations should influence the reactivity. Indeed, nowadays a large amount of experimental evidence indicates that the binding of adsorbates at surfaces is strengthened by the presence of low-coordinated

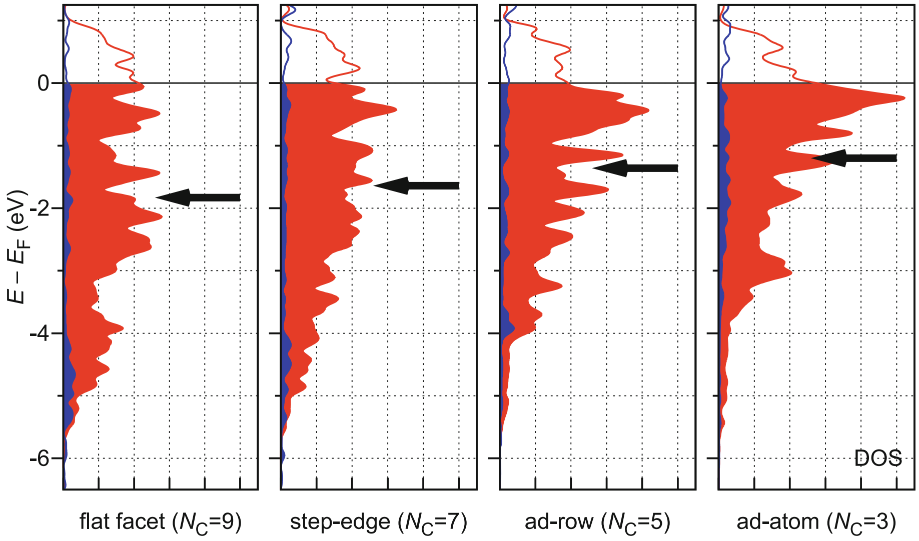


Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 13
a DFT calculated ethylene adsorption energy as a function of the *d*-band center (ε_d) of the Pd(111) surface and commensurate Pd overlayer over close packed Re(0001), Ru(0001), and Au(111) surfaces. b DFT calculated C–H bond activation energies for ethylene (C_2H_4, ●) and ethyl (C_2H_5, ■) dehydrogenation versus Pd *d*-band center for the same systems as in a. On the basis of results from [77]

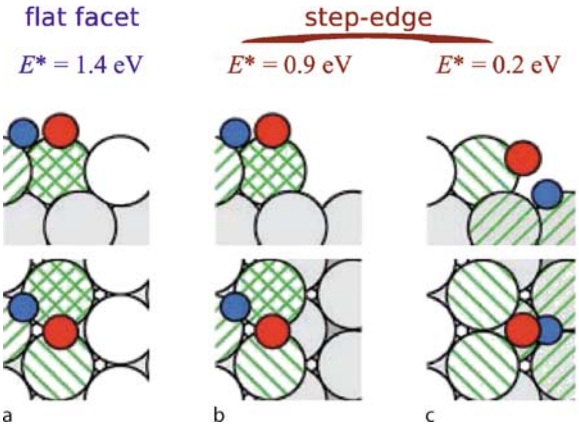
defects [83,93,98,112]. Below we shortly survey how the reactivity can be tuned by strain, defects and alloying.

Strain With computer simulations the effect of strain can be very easily modeled by merely increasing/decreasing the lattice spacings. Although such calculations are fictitious, they are useful to discern the effects of strain on the reactivity, and show that surface reactivity increases with tensile strain and vice-versa [65]. As anticipated on the basis of the *d*-band model, this is due to the fact that a reduced overlap between the *d*-states at neighboring atoms reduces the width of the *d*-band. Experimentally the strained sites can be made, for example, by growing atomically thin commensurate overlayers on support with different lattice spacings [9,87,95]. This way both compressive and tensile strained overlayers can be achieved (depending on the relative lattice spacings of support with respect to metal constituting the overlayers). An example is shown in Fig. 13, which shows the adsorption energy of ethylene (C_2H_4) and activation energies for dehydrogenation of ethylene and ethyl (C_2H_5) on palladium overlayers over rhenium (Re), ruthenium (Ru), and gold (Au) [77]. Correlation between these energies and the *d*-band center is rather good. The Pd overlayer on Au is more reactive than pure Pd, because lattice spacings of Au are larger than that of Pd and vice-versa for Ru. So Pd overlayer experiences tensile and compressive strain on Au and Ru, respectively. Rhenium is an exception, because of very similar lattice spacings to that of Pd, yet the reactivity trend is captured by the *d*-band center, which is apparently affected by the electronic interaction between Pd and Re.

Surface Defects The importance of surface defects in determining the reactivity of heterogeneous catalysts can hardly be overestimated [30]. Not only do defects usually bind adsorbates more strongly than perfect facets do, but in certain cases the reactivity of the step defects is so much larger than that of the terraces as to dominate the dissociation reaction rate at concentrations as low as 1% of step atoms [14]. This can be attributed both to electronic and geometric effects. As for the former, the width of the *d*-band of atoms at defects is reduced due to their decreased coordination (up-shift of ε_d), as shown in Fig. 14. The reactivity is then further increased by geometric effect, because of the availability of new sites. The extent of the two effects is apparent from Fig. 15, which displays the TS for NO dissociation under three different circumstances [26]: (a) shows the TS structure over the perfect facet of Ru(0001), and in (b) the same TS structure (no geometric effect) is shown near the step-edge defect. The reduction of E^* by 0.5 eV as going from (a) to (b) is therefore due to increased reactivity of step atoms (*electronic effect*). The barrier then further decreases by 0.7 eV, by taking advantage of the geometry of the step (*geometric effect*), because the TS structure shown in (c) benefits from being highly coordinated. Note that the TS's for the dis-



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 14
Density of states (DOS) projected to progressively less coordinated metal atom at the Rh surface (N_C ≡ coordination number). From left to right: perfect (111) facet and step-edge, ad-row, and ad-atom defects thereon. The *d*- and *s*-band are colored *red* and *blue*, respectively. *Arrows* indicate the position of the *d*-band centers (ε_d). On the basis of results from [48]

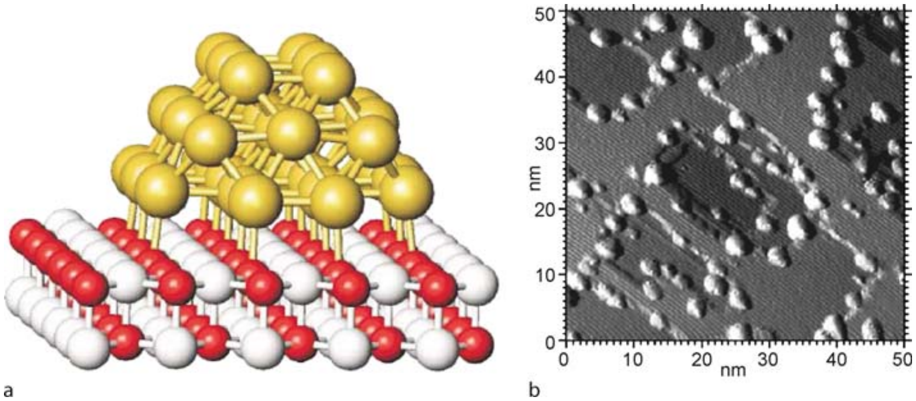


Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 15
Activation energies and transition state structures for NO dissociation over flat Ru(0001) facet a, and step-edge defect thereon b,c [26,30]. In b the TS structure is analogous to that shown in a, whereas in c the advantage of the step-edge geometry is taken into account. Note that the reaction center at flat facet a involves three surface Ru atoms, where that of step-edge c involves five surface Rh atoms. From [26,27]. With kind permission from Springer Science and Business Media

sociation of diatomic molecules such as N_2, NO, O_2, and CO are late [71,76], atomic-like, and isolated atoms are "unhappy" (reactive) and they want to be as highly coordinated as possible. So on one hand, the coordination of metal atoms at the step-edge defect is reduced, but on the other hand the step-edge defects provide the surface sites with high coordination. Therefore, at step-edges the electronic and geometric effects are cooperative resulting in their large reactivity toward dissociation reactions.

Indeed, the surface defects are so much more reactive than flat terraces, that even gold—the noblest among metals [29]—under certain circumstances becomes reactive enough to display very interesting catalytic properties [37]. For example, gold is inert toward the oxidation of carbon monoxide ($CO+\frac{1}{2}O_2 \rightarrow CO_2$), because of its inability to dissociate oxygen molecules. However, small gold nanoparticles supported on oxides (see Fig. 16) have been found to be active [31,107]. This behavior may be attributed to several effects: (i) the abundance of low coordinated reactive sites, which increases with decreasing the particle size (these sites appear at the intersections of low Miller-index planes and include edge, kink and even adatom defects) [66,84]; (ii) strain induced due to oxide-support and small particle size [66], and (iii) the appearance of new special sites at the perimeter of the oxide–metal interface (e. g. due to oxide-induced charge transfer to adsorbate) [60,72,73,74].

Alloying Alloying presents a very interesting way of changing the reactivity of the surface. On one hand addition of another metal (impurity) into a given host metal changes its electronic structure, but on the other hand it opens a way to tailor a selectivity by a geometric effect. The electronic effect on the reactivity of alloys is well captured by the shift of the *d*-band center, and these shifts have been calculated by DFT and tabulated for many catalytically interesting bimetallic alloys [88], thus providing a guideline for an on-demand tuning of the electronic structure of the surface. An interesting and counterintuitive example of alloying is that in some cases the addition of an inert IB metal enhances the adsorption ener-



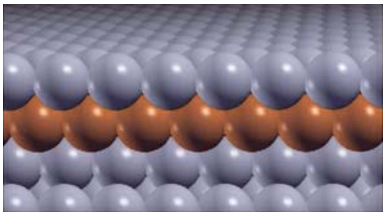
Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 16
Low coordinated surface defects are so reactive that even gold—the noblest among metals [29]—becomes catalytically active in the limit of very small nanoparticles. a A model of gold nanoparticles on a MgO(100) support. Reprinted figure with permission from [72]. b An STM (scanning-tunneling-microscope) image of gold nanoparticles on TiO_2(110). From [107], reprinted with permission from AAAS

gies on more reactive VIII group metals [28,77]. This is mainly due to strain [88], i. e., if a "small" metal atom is added into a "larger" host, it will experience a tensile strain and vice-versa (and the *d*-band center will shift accordingly).

As for the geometric effect, it has long been known from experiments that the addition of a group-IB metal (Cu, Ag, Au) to group-VIII metals (Fe, Co, Ni; Ru, Rh, Pd; Os, Ir, Pt) can lead to an increased selectivity of the catalyst [97]. In particular, it is known that such catalysts hinder hydrogenolysis reactions which involve C–C bond breaking, but do not significantly affect dehydrogenation reactions which involve C–H bond breaking [97]. For example, the hydrogenolysis reaction require sites with a larger ensemble of active atoms than dehydrogenation reactions. Already the addition of a small amount of IB metal will substantially decrease the number of sites composed of a large ensemble of active metal atoms, thus disfavoring reactions that require large ensembles of active atoms (as shown schematically by Fig. 12b) and making the dehydrogenation reactions—which require a small ensemble of active atoms—more selective [97].

Tuning the surface composition of an alloy represents an exciting way of designing new reaction centers with tailored catalytic properties. However, the actual composition cannot be varied at random. Some combinations of metals may lead to segregation, while for others the composition at the surface differs from that in the bulk (the amount of a given metal at the surface is either enriched or depleted). An extensive database of surface segregation energies have been compiled by DFT calculations [89], and the segregation tendency may be explained by surface energy and crystal structure differences between the host and impurity. Even if the two metals do not form a bulk alloy (i. e. they are immiscible), they may form an alloy on the surface. Such alloys have been shown to have interesting catalytic properties [2,22,23,44,97]. An example of a AuNi surface alloy used to catalyze steam reforming will be presented in Sect. "Improvement of Steam-Reforming Catalyst".

An exiting novel class of near surface alloys have been considered recently where the impurity metal atoms are located primarily in the subsurface layer [22,44] (see Fig. 17). The DFT calculations indicate that such alloys yield superior catalytic behavior for hydrogen related reaction, because they bind atomic hydrogen weakly, yet they are able to dissociate the H_2 molecule efficiently [22]. As for the water–gas shift reaction ($CO + H_2O \rightarrow CO_2 + H_2$), such a type of subsurface alloy was also suggested to be promising due to the ability to bind CO weakly and to dissociate water easily [44].



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 17
An idealized model of a near-surface alloy (NSA) with impurity metal in the subsurface layer. NSA's have interesting catalytic properties [2,22,23,44,97]

Examples

As described above the reactivity of special sites on the surface can be tuned in several ways, which are relatively easy to achieve with computer DFT simulations, because there we are assembling a given structure "atom by atom", and hence we have total control over it. Computer simulations allow one to consider and investigate structures under various circumstances that are otherwise unfeasible in reality with the aim to gain new understanding. One could envisage many different surface reaction centers and some of them may have very appealing catalytic properties. Below a few examples are presented with the aim to demonstrate how new reaction centers may be proposed on the basis of insight gained by careful analysis of results of quantum mechanical computer simulations.

Tuning the Relative Rates of Methane Dehydrogenation

One of the problems in the conversion of methane (CH_4)—the principal component of natural gas—to more useful liquid-phase chemicals, such as methanol, is that, although many catalysts are able to cleave the C–H bonds, they do so in a nonselective manner. The efficiency of catalysts is thus limited by the tendency of dehydrogenation to proceed until graphite is eventually formed on the surface, thus poisoning the catalyst [20]. Although a symbolic reaction equation may look as simple as $CH_4 + ? \rightarrow CH_3OH + \cdots$ it appears in reality so difficult that—despite its importance —a direct conversion of methane to methanol (or other liquid chemicals) is not yet feasible. For this reason let us ask a simpler question: is it possible to design a reactive center that will cleave the C–H bonds of methane selectively?

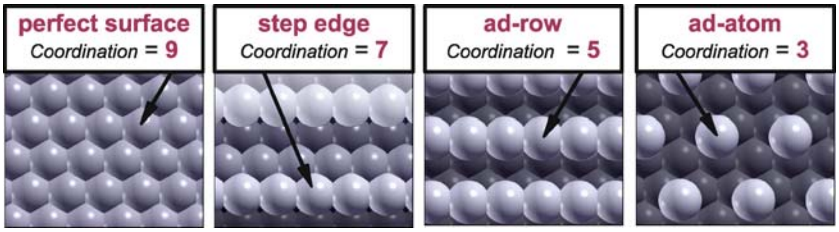
By careful analysis of the results of DFT computer simulations, it has been shown how to design such a reaction center that effectively activates the first dehydrogenation step while hindering the second one [48,49,114]:

$$\begin{aligned} &1.\quad CH_4 \rightleftharpoons CH_3 + H\,, \\ &2.\quad CH_3 \rightleftharpoons CH_2 + H\,, \end{aligned} \tag{14}$$

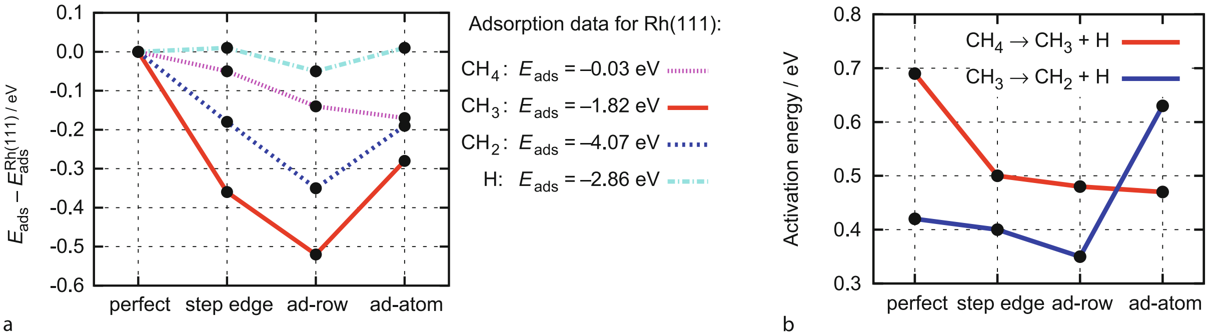
The activation energies can be used as a measure of reaction rate constants (e. g. Eq. (6), $k_f = \nu e^{-E^*/kT}$), therefore a target reaction center has to reverse the natural order of the two barriers' heights. This means that the barrier for the first step has to be lower than that for the second.

Local Geometry of the Reaction Site Kokalj et al. [48] have investigated the effect of local geometry of the reaction center on the two energy barriers. These authors considered the Rh(111) surface because Rh is known for its reactivity toward the C–H bond cleavage of CH_4 [103]. The local geometry of the reaction center was varied by "creating" progressively less coordinated defective sites on the surface, passing from perfect (111) facet with nine-fold coordinated surface atoms (coordination number, $N_C = 9$), to step-edge ($N_C = 7$), added row ($N_C = 5$), and finally to an ad-atom defect ($N_C = 3$). These surface structures are shown in Fig. 18, and the corresponding *d*-band projected density of states is shown in Fig. 14. For each one of them, the DFT calculated optimal adsorption energies of the involved CH_4, methyl (CH_3), ethylene (CH_2), and hydrogen (H) species are reported in Fig. 19a. The adsorption of CH_4, CH_3, and CH_2 is rather sensitive to the coordination of the metal atoms at the binding site, whereas the adsorption of hydrogen is not so much so.

As for CH_3 and CH_2, we may see that they are stabilized at low coordinated defects, as expected on the basis of the Hammer–Nørskov *d*-band model, except for the ad-atom defect, where the trend is reversed. Moreover, the ad-atom behaves rather unexpectedly also with respect to C–H activation energies, as shown in Fig. 19b: although the barrier for the first dehydrogenation step decreases as expected with the coordination of the reaction center, the dehydrogenation of methyl ($CH_3 \rightarrow CH_2 + H$) is hindered at an ad-atom defect, where the first dehydrogenation step is instead most favored. Therefore, the behavior of the ad-atom is so peculiar that neither can the adsorption of CH_3 and CH_2 be explained solely by the Hammer–Nørskov *d*-band model (*electronic* effect) nor can the dehydrogenation barriers be explained by the BEP relation. As will be shown below this is related to structural features of the ad-atom (*geometric* effect). We proceed by analyzing the involved structures, in particular, adsorbed CH_3, CH_2,



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 18
Different local geometries of the reaction centers on Rh(111). From *left to right*: perfect surface, step-edge, ad-row, and ad-atom. Reprinted with permission from [48]



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 19
a Adsorption energies of methane (CH_4), methyl (CH_3), methylene (CH_2) and atomic hydrogen (H) on a step-edge ($N_C = 7$), ad-row ($N_C = 5$), and ad-atom ($N_C = 3$) with respect to the Rh(111) facet ($N_C = 9$). b Activation energies for the first two steps of methane dehydrogenation as a function of local geometry of the reaction site. On the basis of results from [48]

and the corresponding TS structures for the two dehydrogenation steps.

Analysis of CH_3 Adsorption Methyl is a radical whose unsaturated C atom strongly binds to the metal surface. But in addition to this C–metal bond, methyl also displays a peculiar three-center C–H–metal bond, which is in part responsible for the peculiar behavior mentioned above. The three-center C–H–metal bond is usually referred to as *agostic bond* in organometallic chemistry [12,25,111]. The *agostic* bond results essentially from the hybridization of $1e$ C–H bonding orbitals with the d states of the metal surface [68,78]. *Agostic* bonds occur when the adsorption geometry allows for a small H–metal distance. Consider for example a methyl radical adsorbed at an fcc hollow site with H atoms pointing toward either the nearest hollow sites (no agostic bonds, Fig. 20a) or toward the nearest metal atoms (thus giving rise to agostic bonds, Fig. 20b). The difference between the two adsorption energies is rather large (≈ 0.4 eV); also note that in the latter case, the C–H distance is slightly larger than in the former. This effect is enhanced when going from the (111) facet to the step edge and ad-row: the C–H distance increases, Fig. 20, and more charge is delocalized from the C–H bond toward the Rh atom, Fig. 21. On the other hand, at an ad-atom defect the H–metal distance is quite large, indicating a smaller *agostic* interaction, and consequently explaining the reversed trend of methyl adsorption energy at the ad-atom.

Agostic bonding also stretches and correspondingly weakens the C–H bond [69], thus helping break it, and this may partially explain why the ad-atom reaction center is not so efficient for methyl dehydrogenation.

Analysis of CH_2 Adsorption To explain the reversed trend of methylene adsorption energy at the ad-atom, note that to couple the two unpaired electrons of the methylene diradical, CH_2 has to form (at least) two bonds with the substrate. The local structure of an ad-atom defect, however, is such that this requirement is not easily fulfilled. As passing from a (111) facet to a step edge and an added row, CH_2 binds to progressively less coordinated atoms, however at an ad-atom defect it bridges the ad-atom ($N_C = 3$) to a surface atom underneath ($N_C = 10$). The average coordination number of the two bridged atoms is thus $N_C = 6.5$, close to the value $N_C = 7$ of the step edge. As a result, the adsorption energy of methylene at an ad-atom and at a step edge are very similar: -4.26 and -4.25 eV, respectively (see Fig. 19a).

Structural Analysis of Transition States An analysis of a large number of identified TS structures, not only those on currently considered reaction centers, but also on Ru(0001) [11], Ni(111) [52,70,108], Pd(100) [113], and Pt(110)(1×2) [82] reveals that the TS's display some universal features [49,114], which are evident from Fig. 22, and can be described as:

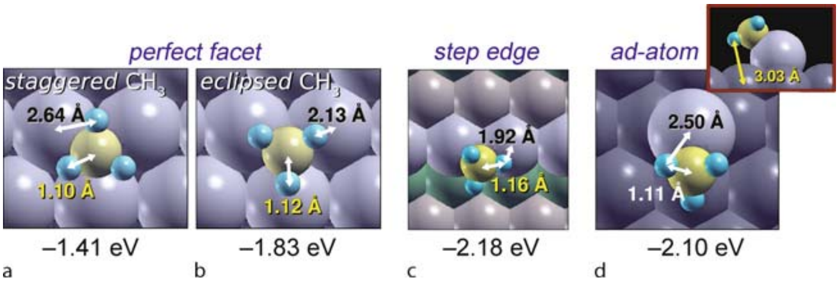
(i) For both reactions the TS is *late*, i. e., its structure is close to that of the final state: the C–H distance for the detaching H atom at the TS is in the range 1.6–1.7 Å, to be compared with an equilibrium bond length of 1.1 Å.
(ii) The TS of the first dehydrogenation step involves only one metal atom: the CH_3 fragment is located at the top site.
(iii) On the other hand, the TS for the second dehydrogenation step involves two metal atoms: the CH_2 (methylene) fragment is located at the bridge site, while the dissociating H atom is at the top site.

We further notice that although the adsorption energy of H is rather insensitive to the coordination of the metal reaction center (see Fig. 19), this is so only for the best adsorption sites corresponding to each reaction center (hollow or bridge). However, at the TS of the $CH_3 \rightarrow CH_2 + H$, the H atom is located instead close to a top site (see Fig. 22b). The hydrogen adsorption energies on top of a metal atom are also rather similar for the (111) facet, at step edge, and at ad-row (≈ -2.45 eV). However, on top of the ad-atom the adsorption is about 0.2 eV less stable, and this further contributes to the large value of the E^* at an ad-atom defect.

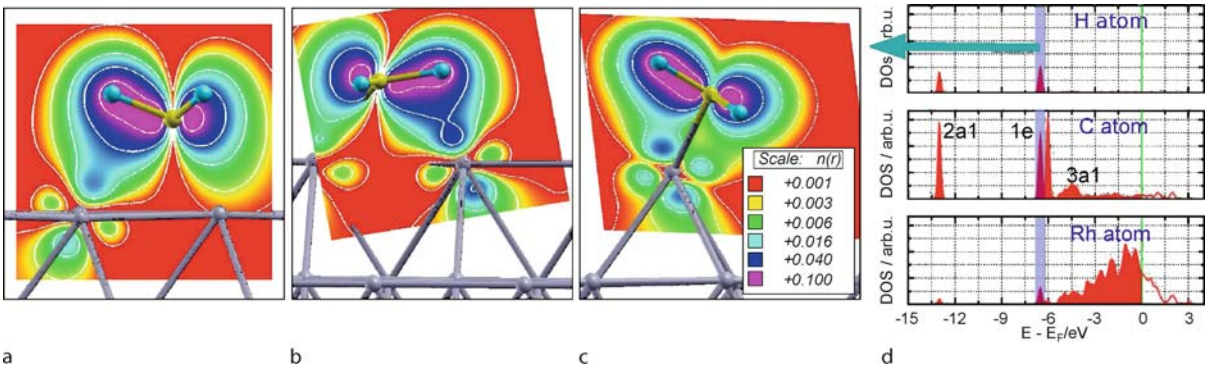
Requirements for Reaction Center The insight obtained on the basis of the analysis presented above, may be used to design new reaction centers, which would enhance the first reaction while hindering the second, by not only acting on the local geometry, but also on the local chemical composition. On the basis of the above analysis the following three requirements should be met [49]:

(i) One of the CH_2–metal bonds should be substantially weakened;
(ii) The strength of the H–metal bond at the top site should be reduced, and
(iii) *Agostic* C–H–metal bonding of CH_3 should be prevented.

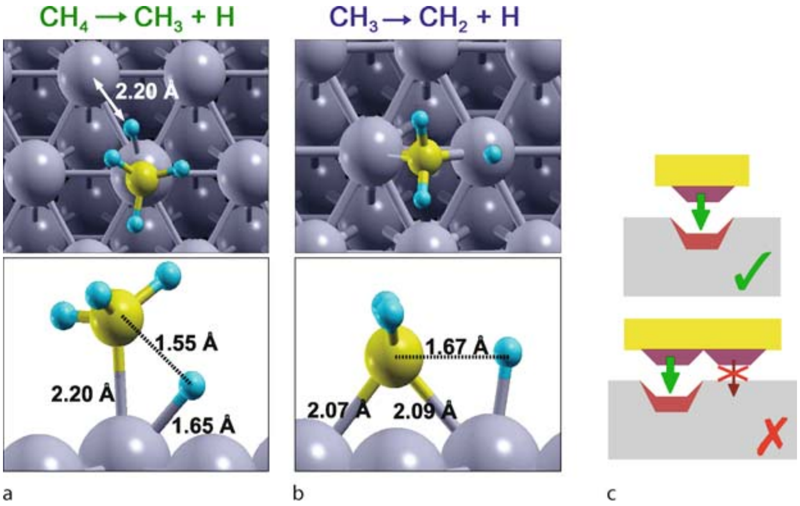
Note that *agostic* interactions are small, hence the effect of *agostic* bonding on the reaction barrier can only be small. On the other hand, the CH_2–surface and H–surface bonds are strong, and a large reduction of these bond strengths can affect the reaction barrier substantially. It turns out



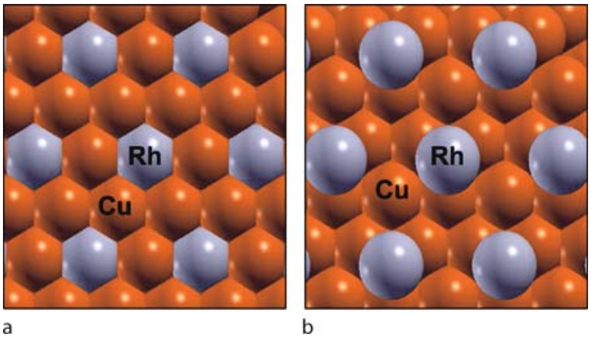
Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 20
a,b Staggered and eclipsed orientation of methyl at an fcc site of perfect Rh(111). In **a** the orientation of CH_3 is such that H atoms are staggered with the nearest metal atoms, whereas in **b** they are eclipsed thus forming agostic bonds. **c** CH_3 adsorbed at the step edge with one of the H atoms pointing toward the step metal atom thus forming the agostic bond. **d** CH_3 at the ad-atom: agostic bonding is hindered due to large H-metals distances. Adapted from [49]



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 21
a, b, c Electron charge density, integrated over the energy window marked by a *blue stripe* in **d**, illustrating the extent of three center C–H–metal *agostic* bonding of methyl adsorbed on **a** a Rh(111) facet, **b** a step edge, and **c** an ad-atom. The magnitude of ILDOS increases from *red to violet* following a *rainbow* scale. Five contours are drawn in logarithmic scale from 10^{-1} to $10^{-3} e/a_0^3$. **d** Density of states projected to H, C, and Rh atom. Adapted from [49]



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 22
a, b Main features of the transition state structures for the first two dehydrogenation reactions on a flat Rh(111) surface. **c** A reaction center composed of a single reactive site is not compatible with the structural characteristics of TS shown in **b**, and would therefore destabilize it. Adapted from [49]



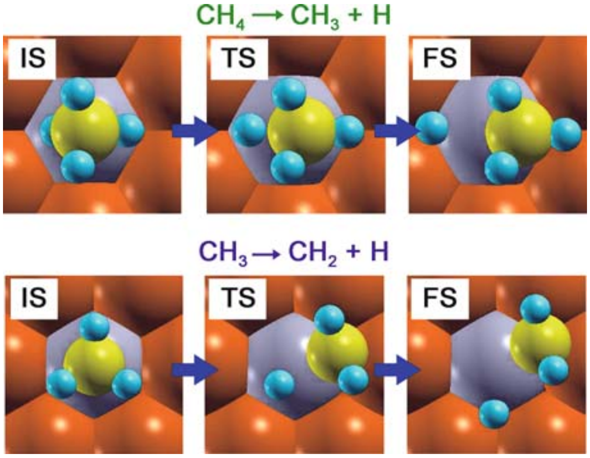
Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 23
a Rh atom substitutionally embedded into Cu(111) and b Rh ad-atom on Cu(111). Reprinted with permission from [49]

that the above requirements can be fulfilled more or less simultaneously.

Local Chemical Composition of Reaction Site The requirements just outlined, suggest that a proper reaction center should be composed of solely one reactive atom, embedded, for example, in an inert environment (shown schematically in Fig. 22c). Consider, for example, an isolated reactive Rh atom substitutionally embedded into a more inert Cu(111) surface, such as shown in Fig. 23a [49]. This can be seen as a model for a Cu-rich phase of a RhCu alloy. Bulk RhCu alloy shows two phases: a Rh-rich phase (with Cu concentration, $X_{\mathrm{Cu}} \lesssim 0.1$) and a Cu-rich phase (with $X_{\mathrm{Cu}} \gtrsim 0.8$) [63]. However, on the surface the amount of Cu is enriched, e. g., already at a small amount of Cu ($X_{\mathrm{Cu}} > 0.05$) the Cu-rich phase will form on the surface [10].

At this reaction center (Fig. 23a), the CH_2 forms one strong bond with the substitutional Rh atom, while the other bond with a Cu atom is much weaker. In addition, agostic bonding is also prevented. Therefore, it performs much better than the Rh(111): the calculated E^* for the first dehydrogenation step is 0.70 eV, whereas that for the second is as large as 0.84 eV (to be compared with 0.69 and 0.42 eV at the Rh(111), respectively) [49]. Snapshots of the two reactions are shown in Fig. 24. Given that the two reaction barriers on clean Cu(111) are predicted to be 1.7 and 1.5 eV [49], dehydrogenation would selectively occur near the Rh atom.

Tuning the chemical composition of the reaction center therefore substantially increases the barrier for the second dehydrogenation step, but hardly affects the barrier for the first step in the current case. However, by combining the structural effects with the chemical effects just outlined, it is possible to selectively increase the barrier of the second dehydrogenation step, while reducing that of the first. Consider for example a Rh ad-atom adsorbed onto a Cu(111) surface, as shown in Fig. 23b [49]. In this case, the reaction barrier for the first reaction is very small, $E^* = 0.35$ eV, while the barrier for the second reaction is quite large, $E^* = 0.89$ eV [49]. In order to illustrate the effects of the local atomic structure and local chemical composition on the two dehydrogenation steps, the reaction energy profiles at the considered reaction centers are shown in Fig. 25.

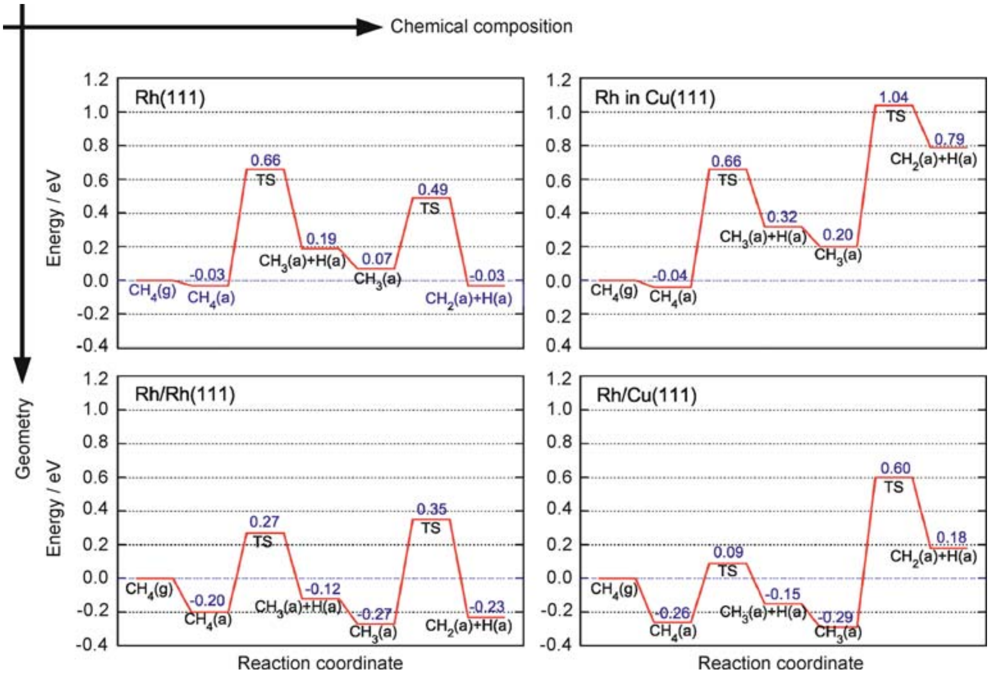


Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 24
Initial- (IS), transition- (TS), and final-state (FS) structures for the first two steps of methane dehydrogenation over the Rh atom substitutionally embedded into Cu(111). Reprinted with permission from [49]

The reaction center based on metal ad-atoms, such as just considered (Fig. 23b), is merely an academic model system, because the naked metal ad-atoms would either cluster or diffuse into the bulk. Nevertheless, the kind of arguments and analysis presented and utilized above are instrumental to the understanding of the mechanisms responsible for the reactivity of reaction centers on real catalysts and to the design and realization of new materials with tailored catalytic properties. As for the ad-atoms, the problem is of course open on how to stabilize them. For example, Zhang and Hu [114] considered isolated Pt ad-atoms on an oxide $MoO_3(010)$ surface, and calculated that the barrier for methane dehydrogenation is significantly lower than that on a Pt(111), while the further dehydrogenation of methyl is blocked.

Improvement of Steam-Reforming Catalyst

Today, the conversion of methane to liquid chemicals is utilized via an in-direct route: in a first step the methane is transformed into a syngas (a mixture of H_2 and CO). Then in the second step the liquid chemicals such as, for exam-



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 25

Reaction profiles for the two dehydrogenation reactions. *Arrows* indicate the improved parameter of the reaction center. The zero level is the energy of the gas phase methane. Difference between the labels "CH$_3$(a)+H(a)" and "CH$_3$(a)" is that for the latter the H(a) diffused to an equivalent site far from CH$_3$(a). Labels Rh in Cu(111), Rh/Rh(111), and Rh/Cu(111) stand for Rh atom substitutionally embedded into Cu(111) (Fig. 23a), Rh ad-atom on Rh(111), and Rh ad-atom on Cu(111) (Fig. 23b), respectively. Reprinted with permission from [49]

ple, methanol are synthesized from syngas. Note, however, that a direct route to methanol synthesis would surpass a very expensive syngas production.

The syngas may be produced via several processes, one of them being the steam reforming:

$$CH_4 + H_2O \rightleftharpoons CO + 3H_2 \quad (\Delta H^0_{298K} = 2.14\,\text{eV}) . \quad (15)$$

The elementary reaction steps can be classified into two groups, and can be written as (omitting adsorption, diffusion, and desorption steps):

(i) dissociation of reactants:

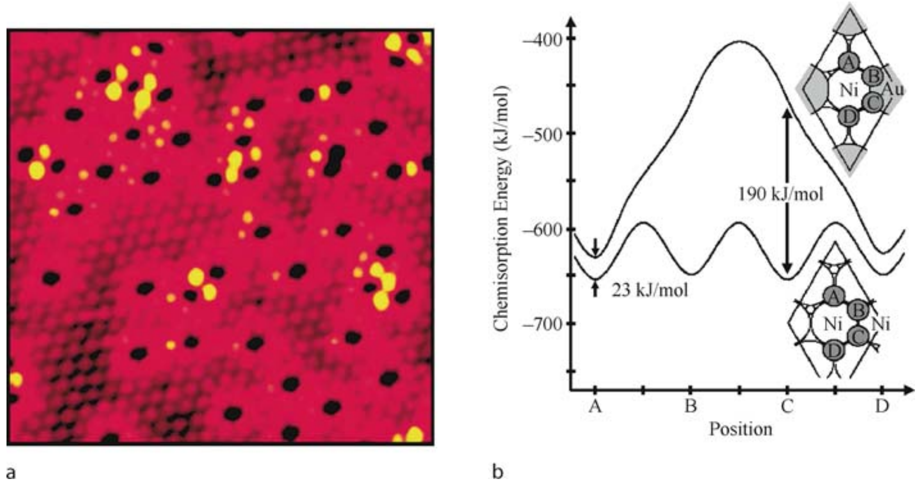
$$\begin{aligned} &1. \quad CH_4 \rightarrow CH_{3(ad)} + H_{(ad)} \\ &2. \quad CH_{3(ad)} \rightarrow CH_{2(ad)} + H_{(ad)} \\ &3. \quad CH_{2(ad)} \rightarrow CH_{(ad)} + H_{(ad)} \\ &4. \quad CH_{(ad)} \rightarrow C_{(ad)} + H_{(ad)} \\ &5. \quad H_2O \rightarrow OH_{(ad)} + H_{(ad)} \\ &6. \quad OH_{(ad)} \rightarrow O_{(ad)} + H_{(ad)} \end{aligned} \qquad (16)$$

(ii) formation of products:

$$\begin{aligned} &1. \quad H_{(ad)} + H_{(ad)} \rightarrow H_2 \\ &2. \quad O_{(ad)} + C_{(ad)} \rightarrow CO \end{aligned} \qquad (17)$$

On the basis of detailed investigation of elementary processes by means of atomistic DFT computer-simulations and surface-science experiments, Besenbacher et al. [2] designed an improved high surface area alloy catalyst for steam-reforming process. The process is obviously limited by the ability of a catalyst to activate hydrocarbon molecules, and its tendency to form an undesired graphite, $C_{(ad)} \rightarrow C_{(s)}$, which poisons the catalysts. As already stated, catalysts are able to break C–H bonds, but in a nonselective manner, which is precisely what is done in steam-reforming, where methane is stripped off all the H atoms, Eqs. (16.1)–(16.4). At this point, two issues are crucial: (i) the surface should prevent naked $C_{(ad)}$ atoms from clustering so as to avoid the graphite growth, and (ii) the $C_{(ad)}$ should not bind too strongly to the surface: recall that the less the $C_{(ad)}$ binds to the surface the easier it will react with chemisorbed $O_{(ad)}$.

A traditional catalyst in steam-reforming is based on Ni as the active element [45]. Kratzer and Besenbacher et al. [2,52] considered adding a small amount of Au into the Ni. The idea is to prevent the $C_{(ad)}$ atoms from clustering. Although the two metals do not form a bulk alloy—they are immiscible in the bulk—they do form an alloy in the



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 26

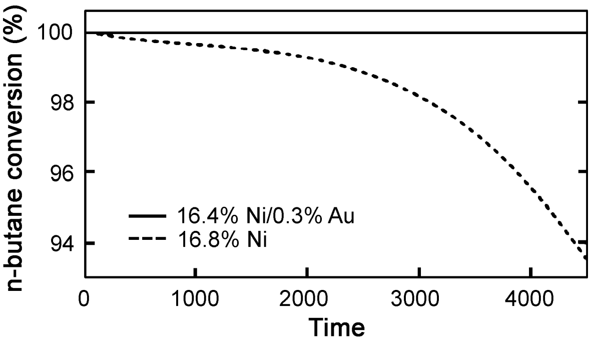
a STM image of a surface NiAu alloy with 7% of a surface Au atoms. Au atoms are black and the neighboring Ni atoms are brighter than other Ni atoms indicating the modification in electronic structure. b DFT calculated adsorption energy of a C atom as a function of its position on the surface; *bottom* curve shows the diffusion profile of the C atom on clean Ni(111) along A–B–C–D positions shown in the *bottom inset*. *Top* curve shows the same diffusion profile for a situation where one Ni atom was substituted by an Au atom as shown by *top inset*. From [2], reprinted with permission from AAAS

outermost surface layer, as shown by an STM image in Fig. 26a.

As already discussed in Sect. "Tuning the Relative Rates of Methane Dehydrogenation", the transition states for dehydrogenations are similar on various transition metals, so in the current case they will be similar to those displayed in Fig. 22 for the first two dehydrogenation steps. DFT calculations by Kratzer et al. [52] predicted that the addition of Au into Ni(111) decreases the reactivity toward the C–H bond breaking, in accord with experimental observation [36]. The activation barrier for the first dehydrogenation step increases by 0.17 eV at the reaction center involving one Au atom, and more than twice that for the center with two Au atoms. This is at variance with what was predicted on RhCu alloy [49] (Sect. "Tuning the Relative Rates of Methane Dehydrogenation"), where the barrier for CH_4 dehydrogenation was almost not affected by the presence of Cu. This can be explained by the fact that Ni and Au atoms differ substantially in size, whereas Cu and Rh do not. Addition of "larger" Au impurity onto the Ni layer therefore down-shifts the *d*-band center of Ni thus decreasing its reactivity [52]. Nevertheless, the effect of addition of Au on dehydrogenation is relatively modest compared to the effect it has on naked $C_{(ad)}$ atoms and on formation of graphite. In particular, a Au atom makes all the nearest neighboring hollow sites completely unstable for adsorption of carbon atoms (by almost 2 eV), and also destabilizes by 0.24 eV the second nearest neighbor sites, as shown in Fig. 26b. This has a tremendous consequence: a single Au atom excludes six surface Ni metal atoms from binding with $C_{(ad)}$ (corresponding to six nearest neighbor hollow sites), and additionally makes the next neighboring sites less stable and therefore somewhat more susceptible to CO formation. Therefore, even a small amount of Au prevents the $C_{(ad)}$ atoms from clustering and therefore from graphite being formed.

That actually all works as suggested by DFT calculations, the authors have performed several experiments: (i) they first synthesized a Au/Ni catalyst in a high surface area form on a $MgAl_2O_4$ support, then (ii) they performed a surface structure characterization by means of EXAFS to confirm the Ni and Au actually formed surface alloy, and finally (iii) they measured the steam-reforming activity of a catalyst. They showed that, while the Ni catalyst deactivates rapidly, the activity of the Au/Ni was almost constant under the measured time interval (see Fig. 27) [2].

It is instructive to compare the RhCu model catalyst discussed in Sect. "Tuning the Relative Rates of Methane Dehydrogenation", designed to suppress the dehydrogenation of methyl, to the current NiAu catalyst: the AuNi catalyst consists of individual inert atoms in a reactive substrate, whereas the RhCu model catalyst consists of isolated reactive atoms embedded in an inert (or less reactive) substrate. This is entirely due to the geometric (ensemble) effect: The C atom requires sites composed of a large ensemble of active surface atoms, whereas for the dehydro-



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 27
Conversion of *n*-butane on Ni (*dashed line*) and NiAu surface-alloy (*solid line*) as a function of time during steam reforming. From [2], reprinted with permission from AAAS

genation of methyl already two reactive neighboring atoms are sufficient.

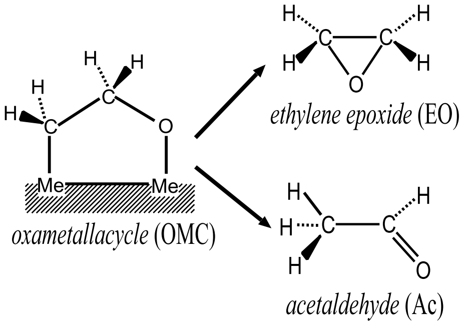
Ethylene Epoxidation

To further illustrate how quantum mechanical simulations help disentangle the mechanisms responsible for the reactivity and selectivity of a catalyst, we consider one more example, the epoxidation of ethylene, which is one of the most important selective oxidation processes based on heterogeneous metallic catalysis [1,53,92,96]. Ethylene epoxide (oxirane) is an important intermediate in the fabrication of glycols and polyols, and silver is a remarkably good catalyst for this reaction with a selectivity at high temperature and high pressure up to 80% in the presence of promoters such as Cl and Cs [53,92,96]. Although a symbolic reaction equation may look as simple as:

$$C_2H_4 + \tfrac{1}{2}O_2 \rightarrow C_2H_4O\,, \tag{18}$$

the elucidation of the reaction mechanism has been a subject of long debate. Its understanding has emerged only recently and was made possible by the synergy between the surface-science experiments and DFT computer simulations. Early studies have mainly focused on the identity of active oxygen species in the reaction and tried to resolve the longstanding question whether molecular or atomic species are the active oxidant [8,21,91]. It has been established only recently that the reaction proceeds via the surface oxametallacycle (OMC) intermediate, which then transforms either to ethylene epoxide (EO) or to acetaldehyde (Ac) [5,42,43,55,56,57,59,67,100,101], the latter reaction leading to undesired total combustion. The overall reaction can be described by the following minimal-sequence of steps:

$$\begin{array}{lr} 1. & O_{2(g)} \rightarrow O_{2(ad)} \\ 2. & O_{2(ad)} \rightarrow 2O_{(ad)} \\ 3. & C_2H_{4(g)} \rightarrow C_2H_{4(ad)} \\ 4. & C_2H_{4(ad)} + O_{(ad)} \rightarrow OMC_{(ad)} \\ 5a. & OMC_{(ad)} \rightarrow EO \\ 5b. & OMC_{(ad)} \rightarrow Ac\,. \end{array} \tag{19}$$



For a more elaborate list of elementary steps see, for example, the work of Stegelmann [101]. For ethylene epoxidation to occur, the catalyst must be reactive enough to dissociate the oxygen molecule [53], but it should be also mild enough so as to form and to stabilize an OMC intermediate, rather than to break C–H bonds, either in the ethylene or in the OMC intermediate [51]. Silver fulfills all these requirements quite well: it is among the most appropriate metals for reactions involving dissociation of oxygen (Sect. "Optimum Chemisorption Energy"), does not activate the C–H bonds [51,92,96] and also binds the OMC weakly to the surface [50,56,64] so that its transformation to EO is facile (as will be shown below). As for dissociation of oxygen, for example, palladium would be also acceptable (Sect. "Optimum Chemisorption Energy"), but at variance with Ag is able to break the C–H bonds [90] (see also Fig. 13b).

Provided that the C–H bond activation is not an issue, the main reaction byproduct involves the formation of Ac [51,56,57]. The Ac molecule is thermodynamically more stable than the EO by about 1 eV. The reason that the EO can be synthesized is because a proper catalyst makes the barrier for its formation smaller or at least comparable to the barrier for Ac formation. The epoxidation of ethylene is therefore an example of a kinetically favored reaction. Linic et al. [56,57] showed that the selectivity of a catalyst is determined by the branching reaction of OMC to form EO rather than the competitive Ac. In particular, at given temperature it would mainly depend on the difference between the Ac and EO activation energies, $\Delta E^* = E^*_{\mathrm{Ac}} - E^*_{\mathrm{EO}}$ (assuming the two corresponding prefactors are similar). The selectivity, S, can be approximately expressed as:

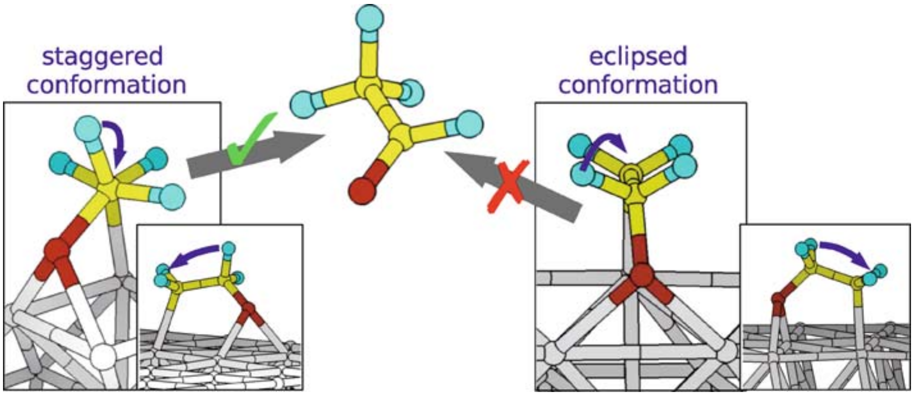
$$S = \frac{r_{\mathrm{EO}}}{r_{\mathrm{EO}} + r_{\mathrm{Ac}}} \approx \frac{\mathrm{e}^{-E^*_{\mathrm{EO}}/kT}}{\mathrm{e}^{-E^*_{\mathrm{EO}}/kT} + \mathrm{e}^{-E^*\mathrm{Ac}/kT}} = \frac{1}{1 + \mathrm{e}^{-\Delta E^*/kT}}, \qquad (20)$$

where r_{EO} and r_{Ac} are the rates of EO and Ac formation, respectively. Therefore, the larger is the ΔE^* the larger is the selectivity. The ability to stabilize the transition state for EO formation with respect to that for Ac formation would therefore lead to increased selectivity.

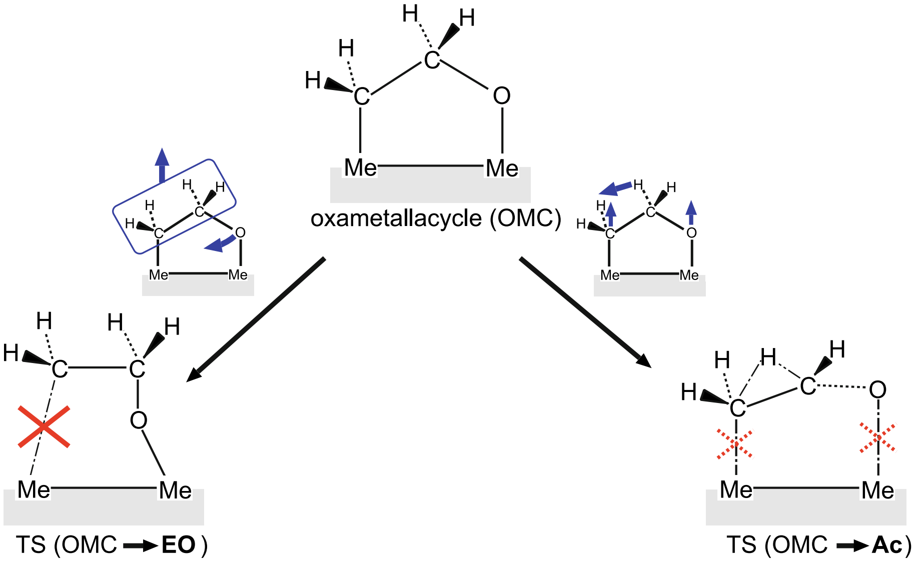
Bocquet et al. [4] realized on the basis of DFT calculations that the selectivity is linked to the conformational arrangement of the OMC intermediate. By considering the trajectories of atoms during the EO and Ac formation and the structure of the two transition states, they suggested that staggered OMC conformation favors the formation of Ac, whereas eclipsed OMC conformation favors the EO formation. This is due to the fact that the Ac formation involves the shift of the H-atom from the first C atom to the second C atom, which is assisted if the two CH_2 groups of the OMC intermediate are in staggered conformation. This is schematically shown in Fig. 28. Hence, if on a given surface both staggered and eclipsed forms exist, the difference between the two activation energies would be related to the relative energy stability of the two conformational OMC forms.

Kokalj et al. [51] provided on the basis of DFT calculations an alternative explanation as to what determines the selectivity. These authors applied the BEP relationship for the EO and Ac formation on a number of different circumstances, including (111) and (100) surfaces of Rh, Cu, Ag, and Au. This analysis revealed that the formation of Ac follows remarkably well the BEP relation, whereas the formation of EO does so to a lesser extent. By knowing that the EO and Ac bind weakly to the surface [51,64], the main message from this BEP analysis was that the stronger is the OMC–surface interaction the larger are the two barriers [51], which demonstrates that for a facile OMC to EO and Ac transformation a less reactive catalyst is required, such as, for example, silver.

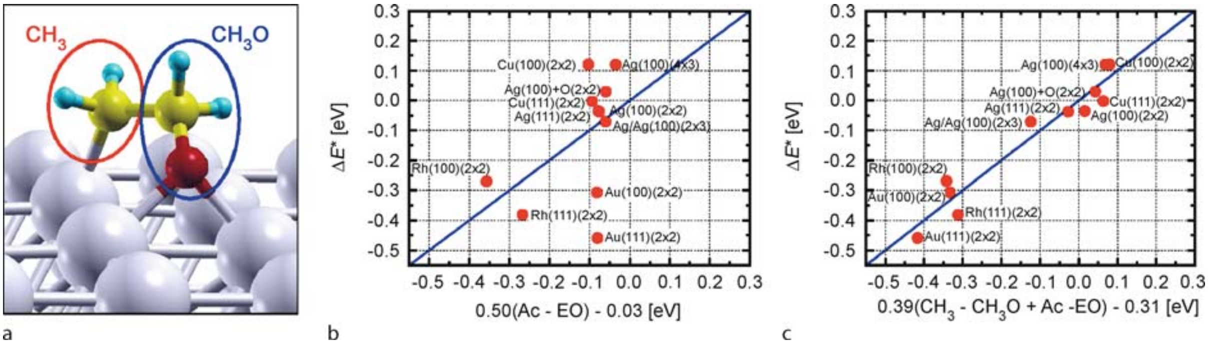
Although an approximate magnitude of the activation barriers for the EO and Ac formation can be determined from the BEP principle, the BEP principle alone is not accurate enough to estimate the selectivity of a catalyst toward the EO formation (see Fig. 30b). The identification of other important factors that determine the catalyst's selectivity toward the formation of EO with respect to Ac formation was made possible by the structural analysis of the TS's for both reactions. Figure 29 schematically depicts the TS's for the EO and Ac formation from OMC on Ag(100). The main difference between the two is that, in TS for EO formation the C–surface bond is fully broken, whereas in TS for Ac formation both C– and O–surface bonds are only partially broken [51]. This may suggest that the two bonds contribute differently to the two activation energies. In particular the stronger is the OMC's O–surface bond with respect to the C–surface bond, the more selective the substrate will be toward EO formation. This argument was quantified by decomposing the OMC–surface interaction into C– and O–surface contributions, which were approximated by the interaction of methyl ($CH_3\cdot$) and methoxy



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 28
The formation of acetaldehyde (Ac) proceeds through the staggered conformation of the two CH_2 groups in oxametallacycle (OMC). Bocquet et al. [4] realized on the basis of DFT calculations that the selectivity in the OMC branching reaction toward either EO or Ac is related to the relative stabilities of eclipsed and staggered conformations of the OMC



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 29
Schematic presentation of an OMC intermediate and transition states for EO (*left*) and Ac (*right*) formation. In the TS leading to EO the C–surface bond is broken, whereas in the TS leading to Ac both C– and O–surface bonds are elongated. During the EO formation the O–atom moves beneath the ethylene fragment, which is concomitantly shifted upward (indicated by *blue arrows*) and as a consequence the C–metal bond is ruptured, whereas in the formation of Ac the whole molecule is upshifted during the 1,2 hydrogen shift, and consequently the C– and O–surface bonds are partially broken. Based on [51]



Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 30
a The OMC–surface interaction can be decomposed into C– and O–surface terms, which can be approximated by the CH_3– and CH_3O–surface interaction, respectively. This leads to an indicator embodied in Eq. (21) that predicts the $\Delta E^* = E^*_{Ac} - E^*_{EO}$ remarkably better than the standard BEP relation. b Standard BEP and c indicator of Eq. (21) in predicting the ΔE^*. The rms and the maximum errors are 0.19 and 0.38 eV, respectively. Reprinted from [51]

($CH_3O\cdot$) radicals with the surface, respectively, as shown in Fig. 30a. By combining this observation with the BEP relation and after some trivial algebra, an equation for the ΔE^* was derived. This equation was than fitted to calculated data, and that resulted in the following expression [51]:

$$\Delta E^* \simeq 0.39\left[E^{CH_3}_{ads} - E^{CH_3O}_{ads} + E^{Ac}_{ads} - E^{EO}_{ads}\right] - 0.31\,, \quad (21)$$

where E_{ads} stands for adsorption energy of corresponding species. The quality of the fit given by Eq. (21) is compared to that obtained from the simple BEP relation in Figs. 30b and 30c. The indicator as embodied in Eq. (21) is able to estimate ΔE^* with an accuracy better than 0.1 eV (the rms and maximum errors are 0.05 and 0.07 eV, respectively), and it suggests that ΔE^* is mainly determined by two contributions: (i) *differential bonding affinity* of the catalyst toward the O- and C-atoms of the OMC, and (ii) the difference between the adsorption energies of the two final states, Ac and EO.

Recently, Linic et al. [58] showed on the basis of DFT computational screening, that Cu/Ag alloy should display a greater selectivity toward EO compared to pure

Ag, which was later confirmed by surface science experiments [40,41]. However, no explanation was given of why this is so. This finding can be straightforwardly explained by the model presented above [51]. The reasons are that Cu shows larger bonding affinity toward the O-atom with respect to the C-atom than silver (this also explains why Cu is intrinsically more selective than Ag for OMC transformation into EO [105,106]). Because Cu is also more reactive than Ag, this implies that on diluted $Ag_{1-x}Cu_x$ alloys ($x \ll 1$, assuming no Cu-Cu nearest neighbors), the most stable OMC binds its O-atom to Cu and its C-atom to Ag (for instance, such OMC orientation is preferred by 0.16 eV on Cu/Ag(100)), enhancing the O–metal bond strength and making the C–metal bond "relatively" weaker. According to the *differential bonding affinity* term of Eq. (21), this makes the formation of EO more selective. In particular, on Cu/Ag(100) alloy ΔE^* is larger than on Ag by 0.17 eV, where 0.04 eV is due to BEP contribution, and 0.10 eV is due to *differential bonding affinity* [51].

Future Directions

Electronic structure calculations, such as DFT, have reached the point, where they can provide understanding of the atomic-scale details of the elementary steps and the mechanisms underlying the reactivity of a catalyst toward specific chemical reactions. This knowledge can then be exploited in the search for better catalysts, and surpasses the *traditional* error-and-trial procedure. In particular, applications are emerging, where new materials are screened computationally. Ever increasing computer power will make it possible to screen large numbers of new materials in the future, and this has great potential in the search for new materials with tailored catalytic properties.

Bibliography

1. Barteau MA (2006) Surface science and the advancement of direct olefin epoxidation. Surf Sci 600:5021
2. Besenbacher F, Chorkendorff I, Clausen BS, Hammer B, Molenbroek AM, Nørskov JK, Stensgaard I (1998) Design of a surface alloy catalyst for steam reforming. Science 279:1913
3. Bligaard T, Nørskov JK, Dahl S, Matthiesen J, Christensen CH, Sehested J (2004) The Brønsted–Evans-Polanyi relation and the volcano curve in heterogeneous catalysis. J Catal 224:206
4. Bocquet M-L, Loffreda D (2005) Ethene epoxidation selectivity inhibited by twisted oxametallacycle: A DFT study on Ag surface-oxide. J Am Chem Soc 127:17207–17215
5. Brainard RL, Madix RJ (1989) Oxidation of tert-butyl alcohol to isobutylene oxide on a silver(110) surface: the role of unactivated carbon-hydrogen bonds in product selectivity. J Am Chem Soc 111:3826
6. Brivio GP, Trioni MI (1999) The adiabatic molecule–metal surface interaction: Theoretical approaches. Rev Mod Phys 71(1):231–265
7. Brönsted JN (1928) Acid and basic catalysis. Chem Rev 5:231–338
8. Campbell CT, Paffett MT (1984) The role of chlorine promoters in catalytic ethylene epoxidation over the Ag(110) surface. Appl Surf Sci 19:28–42
9. Campbell RA, Rodriguez JA, Goodman DW (1992) Chemical and electronic properties of ultrathin metal films: The Pd/Re(0001) and Pd/Ru(0001) systems. Phys Rev B 46(11):7077–7087
10. Chou S-C, Yeh C-T, Chang T-H (1997) Adsorption of hydrogen on dispersed copper-rhodium bimetallic crystallites. J Phys Chem B 101:5828
11. Ciobica IM, Frechard F, van Santen RA, Kleyn AW, Hafner J (2000) A DFT study of transition states for C–H activation on the Ru(0001) surface. J Phys Chem B 104:3364
12. Crabtree RH (1995) Aspects of methane chemistry. Chem Rev 95:897
13. Crawford P, Hu P (2007) Trends in C–O and C–N bond formations over transition metal surfaces: An insight into kinetic sensitivity in catalytic reactions. J Chem Phys 126:194706
14. Dahl S, Logadottir A, Egeberg RC, Larsen JH, Chorkendorff I, Törnqvist E, Nørskov JK (1999) Role of steps in N_2 activation on Ru(0001). Phys Rev Lett 83:1814–1817
15. Dumesic JA (1999) Analyses of reaction schemes using De Donder relations. J Catal 185:496–505
16. Evans MG, Polanyi M (1935) Some applications of the transition state method to the calculation of reaction velocities, especially in solution. Trans Faraday Soc 31:875
17. Evans MG, Polanyi M (1936) Further considerations on the thermodynamics of chemical equilibria and reaction rates. Trans Faraday Soc 32:1333–1360
18. Eyring H (1935) The activated complex in chemical reactions. J Chem Phys 3(2):107–115
19. Fichthorn KA, Weinberg WH (1991) Theoretical foundations of dynamical Monte Carlo simulations. J Chem Phys 95(2):1090–1096
20. Gesser HD, Hunter NR (1998) A review of C-1 conversion chemistry. Catal Today 42:183
21. Grant RB, Lambert RM (1985) A single crystal study of the silver-catalysed selective oxidation and total oxidation of ethylene. J Catal 92:364
22. Greeley J, Mavrikakis M (2004) Alloy catalysts designed from first principles. Nature Matter 3:810
23. Greeley J, Thomas Jaramillo F, Bonde J, Chorkendorff I, Nørskov JK (2006) Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. Nature Matter 5:909
24. Grimley TB, Pisani C (1974) Chemisorption theory in the Hartree–Fock approximation. J Phys C: Solid State Phys 7:2831–2848
25. Hall C, Perutz RN (1996) Transition metal alkane complexes. Chem Rev 96:3125
26. Hammer B (1999) Bond activation at monatomic steps: NO dissociation at corrugated Ru(0001). Phys Rev Lett 83(18):3681–3684
27. Hammer B (2006) Special sites at noble and late transition metal catalysts. Top Catal 37:3
28. Hammer B, Nørskov JK (1995) Electronic factors determining the reactivity of metal surfaces. Surf Sci 343:211
29. Hammer B, Nørskov JK (1995) Why gold is the noblest of all the metals. Nature 376:238

30. Hammer B, Nørskov JK (2000) Theoretical surface science and catalysis - calculations and concepts. Adv Catal 45:71–129
31. Haruta M, Yamada N, Kobayashi T, Iijima S (1989) Gold catalysts prepared by coprecipitation for low-temperature oxidation of hydrogen and of carbon monoxide. J Catal 115:301
32. Henkelman G, Jonsson H (2000) Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. J Chem Phys 113:9978
33. Henkelman G, Uberuaga BP, Jonsson H (2000) A climbing image nudged elastic band method for finding saddle points and minimum energy paths. J Chem Phys 113:9901
34. Hermann K, Bagus PS, Nelin CJ (1987) Size dependence of surface cluster models: CO adsorbed on Cu(100). Phys Rev B 35:9467
35. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. Phys Rev 136(3B):B864–B871
36. Holmblad PM, Larsen HJ, Chorkendorff I (1996) Modication of Ni(111) reactivity toward CH_4, CO, and D_2 by two-dimensional alloying. J Chem Phys 104:7289
37. Hvolbæk B, Janssens TVW, Clausen BS, Falsig H, Christensen CH, Nørskov JK (2007) Catalytic activity of Au nanoparticles. Nano Today 2:14
38. Jacobsen CJH (2000) Novel class of ammonia synthesis catalysts. Chem Commun 12:1057
39. Jacobsen CJH, Dahl S, Clausen BS, Bahn S, Logadottir A, Nørskov JK (2001) Catalyst design by interpolation in the periodic table: Bimetallic ammonia synthesis catalyst. J Am Chem Soc 123:8404
40. Jankowiak JT, Barteau AM (2005) Ethylene epoxidation over silver and copper-silver bimetallic catalysts: II. Cs and Cl promotion. J Catal 236:379
41. Jankowiak JT, Barteau MA (2005) Ethylene epoxidation over silver and copper-silver bimetallic catalysts: I. kinetics and selectivity. J Catal 236:366
42. Jones GS, Mavrikakis M, Mark Barteau A, John Vohs M (1998) First synthesis, experimental and theoretical vibrational spectra of an oxametallacycle on a metal surface. J Am Chem Soc 120:2196–3204
43. Klust A, Madix RJ (2006) Partial oxidation of higher olefins on Ag(111): Conversion of styrene to styrene oxide, benzene, and benzoic acid. Surf Sci 600:2025
44. Knudsen J, Nilekar AU, Vang RT, Schnadt J, Kunkes EL, Dumesic JA, Mavrikakis M, Besenbacher F (2007) A Cu/Pt near-surface alloy for water-gas shift catalysis. J Am Chem Soc 129:6485–6490
45. Kochloefl K (1997) Steam reforming. In: Ertl G, Knözinger H, Weitkamp J (eds) Handbook of Heterogeneous Catalysis, vol 4, chapter 3.2. Wiley, Weinheim, p 1819–1831
46. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140(4A):A1133–A1138
47. Kojima R, Aika K-I (2000) Cobalt molybdenum bimetallic nitride catalysts for ammonia synthesis. Chem Lett 29:514
48. Kokalj A, Bonini N, Sbraccia C, de Gironcoli S, Baroni S (2004) Engineering the reactivity of metal catalysts: a model study of methane dehydrogenation on Rh(111). J Am Chem Soc 126:16732
49. Kokalj A, Bonini N, de Gironcoli S, Sbraccia C, Fratesi G, Baroni S (2006) Methane dehydrogenation on Rh@Cu(111): A first-principles study of a model catalyst. J Am Chem Soc 128:12448
50. Kokalj A, Gava P, de Gironcoli S, Baroni S (2008) Activated adsorption of ethylene on atomic-oxygen covered Ag(100) and Ag(210): formation of an oxametallacycle. J Phys Chem C 112:1019–1027. doi:10.1021/jp0747961
51. Kokalj A, Gava P, de Gironcoli S, Baroni S (2008) What determines the catalyst's selectivity in the ethylene epoxidation reaction. J Catal 254:304–309
52. Kratzer P, Hammer B, Nørskov JK (1996) A theoretical study of CH_4 dissociation on pure and gold-alloyed Ni(111) surface. J Chem Phys 105:5595
53. Lambert RM, Williams FJ, Cropley RL, Palermo A (2005) Heterogeneous alkene epoxidation: past, present and future. J Mol Catal A 228:27
54. Li W-X, Stampfl C, Scheffler M (2002) Oxygen adsorption on Ag(111): A density-functional investigation. Phys Rev B 65:075407
55. Linic S, Barteau MA (2002) Formation of a stable surface oxametallacycle that produces ethylene oxide. J Am Chem Soc 124:310
56. Linic S, Barteau MA (2003) Construction of a reaction coordinate and a microkinetic model for ethylene epoxidation on silver from DFT calculations and surface science experiments. J Catal 214:200–212
57. Linic S, Barteau MA (2003) Control of ethylene epoxidation selectivity by surface oxametallacycles. J Am Chem Soc 125:4034
58. Linic S, Jankowiak J, Barteau MA (2004) Selectivity driven design of bimetallic ethylene epoxidation catalysts from first principles. J Catal 224:489–493
59. Linic S, Piao H, Adib K, Barteau MA (2004) Ethylene epoxidation on Ag: Identification of the crucial surface intermediate by experimental and theoretical investigation of its electronic structure. Angew Chem Int Ed 43:2918
60. Liu Z-P, Gong X-Q, Kohanoff J, Sanchez C, Hu P (2003) Catalytic role of metal oxides in gold-based catalysts: A first principles study of CO oxidation on TiO_2 supported Au. Phys Rev Lett 91(26):266102
61. Lynggaard H, Andreasen A, Stegelmann C, Stoltze P (2004) Analysis of simple kinetic models in heterogeneous catalysis. Prog Surf Sci 77:71–137
62. Masel RI (1996) Principles of adsorption and reaction on solid surfaces. Wiley, New York
63. Massalski TB, Okamoto H, Subramanian PR, Kacprezak L (1990) Binary Alloy Phase Diagrams, vol 2E. ASM, Materials Parks, OH
64. Mavrikakis M, Doren DJ, Barteau MA (1998) Density functional theory calculations for simple oxametallacycles: Trends across the periodic table. J Phys Chem B 102:394–399
65. Mavrikakis M, Hammer B, Nørskov JK (1998) Effect of strain on the reactivity of metal surfaces. Phys Rev Lett 81(13):2819–2822 Sep
66. Mavrikakis M, Stoltze P, Nørskov JK (2000) Making gold less noble. Catal Lett 64:101
67. Medlin JW, Barteau MA, Vohs JM (2000) Oxametallacycle formation via ring-opening of 1-epoxy-3-butene on Ag(110): a combined experimental/theoretical approach. J Mol Catal A 163:129
68. Michaelides A, Hu P (1999) Methyl chemisorption on Ni(111) and C–H–M multicentre bonding: a density functional theory study. Surf Sci 437:362

69. Michaelides A, Hu P (2001) Softened C–H modes of adsorbed methyl and their implications for dehydrogenation: An ab initio study. J Chem Phys 114:2523
70. Michaelides A, Hu P (2000) A first principles study of CH_3 dehydrogenation on Ni(111). J Chem Phys 112:8120
71. Michaelides A, Liu Z-P, Zhang CJ, Alavi A, King DA, Hu P (2003) Identification of general linear relationships between activation energies and enthalpy changes for dissociation reactions at surfaces. J Am Chem Soc 125:3704
72. Molina LM, Hammer B (2003) Active role of oxide support during CO oxidation at Au/MgO. Phys Rev Lett 90(20):206102
73. Molina LM, Hammer B (2004) Theoretical study of CO oxidation on Au nanoparticles supported by MgO(100). Physical Review B (Condensed Matter and Materials Physics) 69(15):155424
74. Molina LM, Rasmussen MD, Hammer B (2004) Adsorption of O_2 and oxidation of CO at Au nanoparticles supported by TiO_2(110). J Chem Phys 120:7673
75. Newns DM (1969) Self-consistent model of hydrogen chemisorption. Phys Rev 178(3):1123–1135
76. Nørskov JK, Bligaard T, Logadottir A, Bahn S, Hansen LB, Bollinger M, Bengaard H, Hammer B, Sljivancanin Z, Mavrikakis M, Xu Y, Dahl S, Jacobsen CJH (2002) Universality in heterogeneous catalysis. J Catal 209:275
77. Pallassana V, Neurock M (2000) Electronic factors governing ethylene hydrogenation and dehydrogenation activity of pseudomorphic Pd_{ML}/Re(0001), Pd_{ML}/Ru(0001), Pd(111) and Pd_{ML}/Au(111) surfaces. J Catal 191:301–317
78. Papoian G, Nørskov JK, Hoffmann R (2000) A comparative theoretical study of hydrogen, methyl, and ethyl chemisorption on the Pt(111) surface. J Am Chem Soc 122:4129
79. Parr RG, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, New York, USA
80. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77(18):3865–3868
81. Perdew JP, Wang Y (1992) Accurate and simple analytic representation of the electron-gas correlation energy. Phys Rev B 45:13244–13249
82. Petersen MA, Jenkins SJ, King DA (2004) Theory of methane dehydrogenation on Pt{110}(1 × 2). Part II: Microscopic reaction pathways for $CH_x \rightarrow CH_{x-1}$ ($x = 1$–3). J Phys Chem B 108:5920
83. Ramsier RD, Gao Q, Waltenburg NH, JT Yates Jr. (1994) Thermal dissociation of NO on Pd surfaces: The influence of step sites. J Chem Phys 100:6837–6845
84. Remediaki IN, Lopez N, Nørskov JK (2005) CO oxidation on rutile-supported Au nanoparticles. Angew Chem Int Ed 44:1824
85. Reuter K, Scheffler M (2006) First-principles kinetic Monte Carlo simulations for heterogeneous catalysis: Application to the CO oxidation at RuO_2(110). Phys Rev B 73(4):045433
86. Ricart JM, Clotet A, Illas F, Rubio J (1994) The analysis of the chemisorption bond from uncorrelated and correlated cluster model wave functions. J Chem Phys 100:1988
87. Rodriguez JA, Goodman DW (1992) The nature of metal-metal bond in bimetallic surfaces. Science 257:897
88. Ruban A, Hammer B, Stoltze P, Skriver HL, Nørskov JK (1997) Surface electronic structure and reactivity of transition and noble metals. J Mol Catal A 115:421
89. Ruban AV, Skriver HL, Nørskov JK (1999) Surface segregation energies in transition-metal alloys. Phys Rev B 59(24):15990–16000
90. van Santen RA (1997) Epoxidation catalysis using heterogeneous catalysts. In: Ertl G, Knözinger H, Weitkamp J (eds) Handbook of Heterogeneous Catalysis, vol 5, chap 4. Wiley, Weinheim, pp 2244–2252
91. van Santen RA, de Groot CPM (1986) The mechanism of ethylene epoxidation. J Catal 98:530–539
92. van Santen RA (1997) Epoxidation catalysis using heterogeneous catalysts In: Handbook of Heterogeneous Catalysis, vol 5, chap. 4.6.1. Wiley, Weinheim, pp 2244–2252
93. Savio L, Vattuone L, Rocca M (2001) The role of steps and terrace width in gas-surface interaction: O_2/Ag(410). Phys Rev Lett 87:276101
94. Schlögl R (1997) Ammonia synthesis. In: Ertl G, Knözinger H, Weitkamp J (eds) Handbook of Heterogeneous Catalysis, vol 4, chap 2.1. Wiley, Weinheim, pp 1697–1748
95. Sellidj A, Koel BE (1994) Electronic and CO chemisorption properties of ultrathin Pd films vapor deposited on Au(111). Phys Rev B 49(12):8367–8376 Mar
96. Serafin JG, Liu AC, Seyedmonir SR (1998) Surface science and the silver-catalyzed epoxidation of ethylene: an industrial perspective. J Mol Catal A 131:157
97. Sinfelt JH (1983) Bimetallic Catalysts. Willey, New York
98. Somorjai GA (1981) Chemistry in two dimensions: Surfaces. Cornell University Press, Ithaca
99. Somorjai GA (1994) Introduction to surface chemistry and catalysis. Willey, Inc., New York
100. Stacchiola D, Wu G, Kaltchev M, Tysoe WT (2000) Spectroscopic studies of ethylene adsorption on oxygen-modified Ag(111) at high pressures. Surf Sci 463:81
101. Stegelmann C, Schiødt NC, Campbell CT, Stoltze P (2004) Microkinetic modeling of ethylene oxidation over silver. J Catal 221:630
102. Stoltze P (2000) Microkinetic simulation of catalytic reactions. Prog Surf Sci 65:65–150
103. Svensson M, Blomberg MRA, Siegbahn PEM (1991) Reaction of second-row transition-metal atoms with methane. J Am Chem Soc 113:7076
104. Szabo A, Ostlund NS (1996) Modern quantum chemistry: introduction to advanced electronic structure theory. Dover publications, Inc., Mineola
105. Torres D, Lopez N, Illas F (2006) A theoretical study of coverage effects for ethylene epoxidation on Cu(111) under low oxygen pressure. J Catal 243:404
106. Torres D, Lopez N, Illas F, Lambert RM (2005) Why copper is intrinsically more selective than silver in alkene epoxidation: Ethylene oxidation on Cu(111) versus Ag(111). J Am Chem Soc 127:10774
107. Valden M, Lai X, Goodman DW (1998) Onset of catalytic activity of gold clusters on titania with the appearance of nonmetallic properties. Science 281:1647
108. Watwe RM, Bengaard HS, Rostrup-Nielsen JR, Dumesic JA, Nørskov JK (2000) Theoretical studies of stability and reactivity of CH_x species on Ni(111). J Catal 189:16
109. Whitten JL, Yang H (1996) Theory of chemisorption and reactions on metal surfaces. Surf Sci Rep 24:55
110. Wu Y, Schmitt JD, Car R (2004) Mapping potential energy surfaces. J Chem Phys 121:1193–1200
111. Zaera F (1995) An organometallic guide to the chemistry of hydrocarbons moieties on transition metal surfaces. Chem Rev 95:2651

112. Zambelli T, Wintterlin J, Trost J, Ertl G (1996) Identification of the active sites of a surface-catalyzed reaction. Science 273:1688–1690
113. Zhang CJ, Hu P (2002) Methane transformation to carbon and hydrogen on Pd(100): Pathways and energetics from density functional theory calculations. J Chem Phys 116:322
114. Zhang CJ, Hu P (2002) The possibility of single C–H bond activation in CH_4 on MoO_3-supported Pt catalyst: A density functional theory study. J Chem Phys 116:4281
115. Zhdanov VP (2002) Monte Carlo simulations of oscillations, chaos and pattern formation in heterogeneous catalytic reactions. Surf Sci Rep 45:231–326
116. Zhdanov VP, Kasemo B (1994) Kinetic phase transitions in simple reactions on solid surfaces. Surf Sci Rep 20:113–189

Computer Graphics and Games, Agent Based Modeling in

BRIAN MAC NAMEE
School of Computing, Dublin Institute of Technology,
Dublin, Ireland

Article Outline

Glossary

Computer generated imagery (CGI) The use of computer generated images for special effects purposes in film production.

Intelligent agent A hardware or (more usually) software-based computer system that enjoys the properties autonomy, social ability, reactivity and pro-activeness.

Non-player character (NPC) A computer controlled character in a computer game – as opposed to a player controlled character.

Virtual character A computer generated character that populates a virtual world.

Virtual world A computer generated world in which places, objects and people are represented as graphical (typically three dimensional) models.

Definition of the Subject

As the graphics technology used to create virtual worlds has improved in recent years, more and more importance has been placed on the behavior of virtual characters in applications such as games, movies and simulations set in these virtual worlds simulations. The behavior of these virtual characters should be believable in order to create the illusion that virtual worlds are populated with living characters. This has led to the application of agent-based modeling to the control of virtual characters. There are a number of advantages of using agent-based modeling techniques which include the fact that they remove the requirement for hand controlling all agents in a virtual environment, and allow agents in games to respond to unexpected actions by players or users.

Introduction

Advances in computer graphics technology in recent years have allowed the creation of realistic and believable virtual worlds. However, as such virtual worlds have been developed for applications spanning games, education and movies it has become apparent that in order to achieve real believability, virtual worlds must be populated with life-like virtual characters. This is where the application of agent-based modeling has found a niche in the areas of computer graphics and, in a huge way, computer games. Agent-based modeling is a perfect solution to the problem of controlling the behaviors of the virtual characters that populate a virtual world. In fact, because virtual characters are embodied and autonomous these applications require an even stronger notion of agency than many other areas in which agent-based modeling is employed.

Before proceeding any further, and because there are so many competing alternatives, it is worth explicitly stating the definition of an intelligent agent that will inform the remainder of this article. Taken from [83] an intelligent agent is defined as "… *a hardware or (more usually) software-based computer system that enjoys the following properties:*

- ***autonomy:*** *agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state;*
- ***social ability:*** *agents interact with other agents (and possibly humans) via some kind of agent-communication language;*
- ***reactivity:*** *agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or*

perhaps all of these combined), and respond in a timely fashion to changes that occur in it;

- ***pro-activeness:*** *agents do not simply act in response to their environment, they are able to exhibit goal-directed behavior by taking the initiative."*

Virtual characters implemented using agent-based modeling techniques satisfy all of these properties. The characters that populate virtual worlds should be fully autonomous and drive their own behaviors (albeit sometimes following the orders of a director or player). Virtual characters should be able to interact believably with other characters and human participants. This property is particularly strong in the case of virtual characters used in games which by their nature are particularly interactive. It is also imperative that virtual characters appear to perceive their environments and react to events that occur in that environment, especially the actions of other characters or human participants. Finally virtual characters should be pro-active in their behaviors and not always require prompting from a human participant in order to take action.

The remainder of this article will proceed as follows. Firstly, a broad overview of the use of agent-based modeling in computer graphics will be given, focusing in particular on the genesis of the field. Following on from this, the focus will switch to the use of agent-based modeling techniques in two particular application areas: computer generated imagery (CGI) for movies, and computer games. CGI has been used to astounding effect in movies for decades, and in recent times has become heavily reliant on agent-based modeling techniques in order to generate CGI scenes containing large numbers of computer generated extras. Computer games developers have also been using agent-based modeling techniques effectively for some time now for the control of non-player characters (NPCs) in games. There is a particularly fine match between the requirements of computer games and agent-based modeling due to the high levels of interactivity required.

Finally, the article will conclude with some suggestions for the future directions in which agent-based modeling technology in computer graphics and games is expected to move.

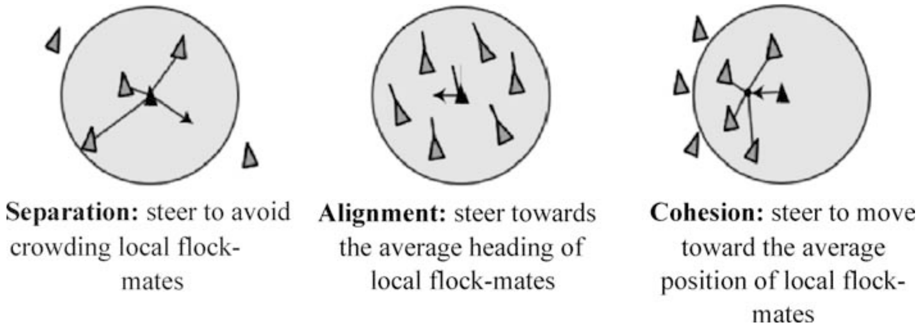
Agent-Based Modelling in Computer Graphics

The serious use of agent-based modeling in computer graphics first arose in the creation of autonomous groups and crowds – for example, crowds of people in a town square or hotel foyer, or flocks of birds in an outdoor scene. While initially this work was driven by visually unappealing simulation applications such as fire safety testing for buildings [75], focus soon turned to the creation of visually realistic and believable crowds for applications such as movies, games and architectural walkthroughs. Computer graphics researchers realized that creating scenes featuring large virtual crowds by hand (a task that was becoming important for the applications already mentioned) was laborious and time-consuming and that agent-based modeling techniques could remove some of the animator's burden. Rather than requiring that animators hand-craft all of the movements of a crowd, agent-based systems could be created in which each character in a crowd (or flock, or swarm) could drive its own behavior. In this way the behavior of a crowd would emerge from the individual actions of the members of that crowd.

Two of the earliest, and seminal, examples of such systems are Craig Reynolds' Boids system [64] and Tu & Terzopoulos' animations of virtual fish [76]. The Boids system simulates the flocking behaviors exhibited in nature by schools of fish, or flocks of birds. The system was first presented at the prestigious SIGGRAPH conference (www.siggraph.org) in 1987 and was accompanied by the short movie "Stanley and Stella in: Breaking the Ice". Taking influence from the area of artificial life (or aLife) [52], Reynolds postulated that the individual members of a flock would not be capable of complex reasoning, and so flocking behavior must emerge from simple decisions made by individual flock members. This notion of emergent behavior is one of the key characteristics of aLife systems.

In the original Boids system, each virtual agent (represented as a simple particle and known as a *boid*) used just three rules to control its movement. These were separation, alignment and cohesion and are illustrated in Fig. 1. Based on just these three simple rules extremely realistic flocking behaviors emerged. This freed animators from the laborious task of hand-scripting the behavior of each creature within the flock and perfectly demonstrates the advantage offered by agent-based modeling techniques for this kind of application.

The system created by Tu and Terzopoulos took a more complex approach in that they created complex models of biological fish. Their models took into account fish physiology, with a complex model of fish muscular structure, along with a perceptual model of fish vision. Using these they created sophisticated simulations in which properties such as schooling and predator avoidance were displayed. The advantage of this approach was that it was possible to create unique, unscripted, realistic simulations without the intervention of human animators. Terzopoulos has since gone on to apply similar techniques to the control of virtual humans [68].



Separation: steer to avoid crowding local flock-mates

Alignment: steer towards the average heading of local flock-mates

Cohesion: steer to move toward the average position of local flock-mates

Computer Graphics and Games, Agent Based Modeling in, Figure 1
The three rules used by Reynolds' original Boids system to simulate flocking behaviors

Moving from animals to crowds of virtual humans, the Virtual Reality Lab at the Ecole Polytechnique Fédérale de Lausanne in Switzerland (vrlab.epfl.ch) led by Daniel Thalmann has been at the forefront of this work for many years. They group currently has a highly evolved system, VICrowd, for the animation of virtual crowds [62] which they model as a hierarchy which moves from individuals to groups to crowds. This hierarchy is used to avoid some of the complications which arise from trying to model large crowds in real time – one of the key gaols of ViCrowd.

Each of the levels in the ViCrowd hierarchy can be modeled as an agent and this is done based on *beliefs*, *desires* and *intentions*. The beliefs of an agent represent the information that the agent possesses about the world, including information about places, objects and other agents. An agent's desires represent the motivations of the agent regarding objectives it would like to achieve. Finally, the intentions of an agent represent the actions that an agent has chosen to pursue. The belief-desire-intention (BDI) model of agency was proposed by Rao and Georgeff [61] and has been used in many other application areas of agent-based modeling.

ViCrowd has been used in ambitious applications including the simulation of a virtual city comprised of, amongst other things, a train station a park and a theater [22]. In all of these environments the system was capable of driving the believable behaviors of large groups of characters in real-time.

It should be apparent to readers from the examples given thus far that the use of agent-based modeling techniques to control virtual characters gives rise to a range of unique requirements when compared to the use of agent based modeling in other application areas. The key to understanding these is to realize that the goal in designing agents for the control of virtual characters is typically not to design the most efficient or effective agent, but rather to design the most interesting or believable character. Outside of very practical applications such as evacuation simulations, when creating virtual characters, designers are concerned with maintaining what Disney, experts in this field, refer to as the *illusion of life* [36].

This refers to the fact that the user of a system must believe that virtual characters are living, breathing creatures with goals, beliefs, desires, and, essentially, lives of their own. Thus, it is not so important for a virtual human to always choose the most efficient or cost effective option available to it, but rather to always choose reasonable actions and respond realistically to the success or failure of these actions. With this in mind, and following a similar discussion given in [32], some of the foremost researchers in virtual character research have the following to say about the requirements of agents as virtual characters.

Loyall writes [46] that "*Believable agents are personality-rich autonomous agents with the powerful properties of characters from the arts.*" Coming from a dramatic background it is not surprising that Loyall's requirements reflect this. Agents should have strong personality and be capable of showing emotion and engaging in meaningful social relationships.

According to Blumberg [11], "*… an autonomous animated creature is an animated object capable of goal-directed and time-varying behavior*". The work of Blumberg and his group is very much concerned with virtual creatures, rather than humans in particular, and his requirements reflect this. Creatures must appear to make choices which improve their situation and display sophisticated and individualistic movements.

Hayes–Roth and Doyle focus on the differences between "*animate characters*" and traditional agents [27]. With this in mind they indicate that agents' behaviors must be "*variable rather than reliable*", "*idiosyncratic instead of predictable*", "*appropriate rather than correct*", "*effective instead of complete*", "*interesting rather than effi-*

cient", and "*distinctively individual as opposed to optimal*".

Perlin and Goldberg [59] concern themselves with building believable characters "*that respond to users and to each other in real-time, with consistent personalities, properly changing moods and without mechanical repetition, while always maintaining an author's goals and intentions*".

Finally, in characterizing believable agents, Bates [7] is quite forgiving requiring "*only that they not be clearly stupid or unreal*". Such broad, shallow agents must "*exhibit some signs of internal goals, reactivity, emotion, natural language ability, and knowledge of agents … as well as of the … micro-world*".

Considering these definitions, Isbister and Doyle [32] identify the fact that the consistent themes which run through all of the requirements given above match the general goals of agency – virtual humans must display autonomy, reactivity, goal driven behavior and social ability – and again support the use of agent-based modeling to drive the behavior of virtual characters.

The Spectrum of Agents

The differences between the systems mentioned in the previous discussion are captured particularly well on the *spectrum of agents* presented by Aylett and Luck [5]. This positions agent systems on a spectrum based on their capabilities, and serves as a useful tool in differentiating between the various systems available. One end of this spectrum focuses on *physical agents* which are mainly concerned with simulation of believable physical behavior, (including sophisticated physiological models of muscle and skeleton systems), and of sensory systems. Interesting work at this end of the spectrum includes Terzopoulos' highly realistic simulation of fish [76] and his virtual stuntman project [21] which creates virtual actors capable of realistically synthesizing a broad repertoire of lifelike motor skills.

Cognitive agents inhabit the other end of the agent spectrum and are mainly concerned with issues such as reasoning, decision making, planning and learning. Systems at this end of the spectrum include Funge's cognitive modeling approach [26] which uses the situation calculus to control the behavior of virtual characters, and Nareyek's work on planning agents for simulation [55], both of which will be described later in this article.

While the systems mentioned so far sit comfortably at either end of the agent spectrum, many of the most effective inhabit the middle ground. Amongst these are *c*4 [13], used to great effect to simulate a virtual sheep dog with the ability to learn new behaviors, *Improv* [59] which augments sophisticated physical human animation with scripted behaviors and the *ViCrowd* system [62] which sits on top of a realistic virtual human animation system and uses planning to control agents' behavior.

Virtual Fidelity

The fact that so many different agent-based modeling systems, for the control of virtual humans exist gives rise to the question why? The answer to this lies in the notion of *virtual fidelity*, as described by Badler [6]. Virtual fidelity refers to the fact that virtual reality systems need only remain true to actual reality in so much as this is required by, and improves, the system.

In [47] the point is illustrated extremely effectively. The article explains that when game designers are architecting the environments in which games are set, the scale to which these environments are created is not kept true to reality. Rather, to ease players' movement in these worlds, areas are designed to a much larger scale, compared to character sizes, than in the real world. However, game players do not notice this digression from reality, and in fact have a negative response to environments that are designed to be more true to life finding them cramped. This is a perfect example of how, although designers stay true to reality for many aspects of environment design, the particular blend of virtual fidelity required by an application can dictate certain real world restrictions can be ignored in virtual worlds.

With regard to virtual characters, virtual fidelity dictates that the set of capabilities which these characters should display is determined by the application which they are to inhabit. So, the requirements of an agent-based modeling system for CGI in movies would be very different to those of a agent-based modeling system for controlling the behaviors of game characters.

Agent-Based Modelling in CGI for Movies

With the success of agent-based modeling techniques in graphics firmly established there was something of a search for application areas to which they could be applied. Fortunately, the success of agent-based modeling techniques in computer graphics was paralleled with an increase in the use of CGI in the movie industry, which offered the perfect opportunity. In many cases CGI techniques were being used to replace traditional methods for creating expensive, or difficult to film scenes. In particular, scenes involving large numbers of people or animals were deemed no longer financially viable when set in the real world. Creating these scenes using CGI involved painstaking hand animation of each character within a scene, which again was not financially viable.

The solution that agent-based modeling offers is to make each character within a scene an intelligent agent that drives its own behavior. In this way, as long as the initial situation is set up correctly scenes will play out without the intervention of animators. The facts that animating for movies does not need to be performed in real-time, and is in no way interactive (there are no human users involved in the scene), make the use of agent-based modeling a particularly fine match for this application area.

Craig Reynolds' Boids system [64] which simulates the flocking behaviors exhibited in nature by schools of fish, or flocks of birds and was discussed previously is one of the seminal examples of agent-based modeling techniques being used in movie CGI. Reynold's approach was first used for CGI in the 1999 film "Batman Returns" [14] to simulate colonies of bats. Reynold's technologies have been used in "The Lion King" [4] and "From Dusk 'Till Dawn" [65] amongst other films. Reynolds' approach was so successful, in fact, that he was awarded an Academy Award for his work in 1998.

Similar techniques to those utilized in the Boids system have been used in many other films to animate such diverse characters as ants, people and stampeding wildebeest. Two productions which were released in the same year, "Antz" [17] by Dreamworks and "A Bug's Life" [44] by Pixar took great steps in using CGI effects to animate large crowds for. For "Antz" systems were developed which allowed animators easily create scenes containing large numbers of virtual characters modeling each as an intelligent agent capable of obstacle avoidance, flocking and other behaviors. Similarly, the creators of "A Bug's Life" created tools which allowed animators easily combine pre-defined motions (known as alibis) to create behaviors which could easily be applied to individual agents in scenes composed of hundreds of virtual characters.

However, the largest jump in the use of agent-based modeling in movie CGI was made in the recent Lord of the Rings trilogy [33,34,35]. In these films the bar was raised markedly in terms of the sophistication of the virtual characters displayed and the sheer number of characters populating each scene. To achieve the special effects shots required by the makers of these films, the Massive software system was developed by Massive Software (www.massivesoftware.com). This system [2,39] uses agent-based modeling techniques, again inspired by aLife, to create virtual extras that control their own behaviors. This system was put to particularly good use in the large scale battle sequences that feature in all three of the Lord of the Rings films. Some of the sequences in the final film of the trilogy, the Return of the King, contain over 200,000 digital characters.

In order to create a large battle scene using the Massive software, each virtual extra is represented as an intelligent agent, making its own decisions about which actions it will perform based on its perceptions of the world around it. Agent control is achieved through the use of fuzzy logic based controllers in which the state of an agent's brain is represented as a series of motivations, and knowledge it has about the world – such as the state of the terrain it finds itself on, what kinds of other agents are around it and what these other agents are doing. This knowledge about the world is perceived through simple simulated visual, auditory and tactile senses. Based on the information they perceive agents decide on a best course of action. Designing the brains of these agents is made easier that it might seem at first by the fact that agents are developed for short sequences, and so a small range of possible tasks. So for example, separate agent models would be used for a fighting scene and a celebration scene.

In order to create a large crowd scene using Massive animators initially set up an environment populating it with an appropriate cast of virtual characters where the brains of each character are slight variations (based on physical and personality attributes) of a small number of archetypes. The scene will then play itself out with each character making it's own decisions. Therefore there is no need for any hand animation of virtual characters. However, directors can view the created scenes and by tweaking the parameters of the brains of the virtual characters have a scene play out in the exact way that they require.

Since being used to such impressive effect in the Lord of the Rings trilogy (the developers of the Massive system were awarded an academy award for their work), the Massive software system has been used in numerous other films such as "I, Robot" [60], "The Chronicles of Narnia: The Lion, the Witch and the Wardrobe" [1] and "Ratatouille" [10] along with numerous television commercials and music videos.

While the achievements of using agent-based modeling for movie CGI are extremely impressive, it is worth noting that none of these systems run in real-time. Rather, scenes are rendered by banks of high powered computers, a process that can take hours for relatively simple scenes. For example, the famous Prologue battle sequence in the "Lord of the Rings: The Fellowship of the Ring" took a week to render. When agent-based modeling is applied to the real-time world of computer games, things are very different.

Agent-Based Modelling in Games

Even more so than in movies, agent-based modeling techniques have been used to drive the behaviors of virtual characters in computer games. As games have become graphically more realistic (and in recent years they have become extremely so) game-players have come to expect that games are set in hugely realistic and believable virtual worlds. This is particularly evident in the widespread use of realistic physics modeling which is now commonplace in games [67]. In games that make strong use of physics modeling, objects in the game world topple over when pushed, float realistically when dropped in water and generally respond as one would expect them to. Players expect the same to be true of the virtual characters that populate virtual game worlds. This can be best achieved by modeling virtual characters as embodied virtual agents. However, there are a number of constraints which have a major influence on the use of agent-based modeling techniques in games.

The first of these constraints stems from the fact that modern games are so highly interactive. Players expect to be able to interact with all of the characters they encounter within a game world. These interactions can be as simple as having something to shoot at or having someone to race against; or involve much more sophisticated interactions in which a player is expected to converse with a virtual character to find out specific information or to cooperate with a virtual character in order to accomplish some task that is key to the plot of a game. Interactivity raises a massive challenge for practitioners as there is very little restriction in terms of what the player might do. Virtual characters should respond in a believable way at all times regardless of how bizarre and unexpected the actions of the player might be.

The second challenge comes from the fact that the vast majority of video games should run in real time. This means that the computational complexity must be kept to a reasonable level as there are only a finite number of processor cycles available for AI processing. This problem is magnified by the fact that an enormous amount of CPU power it usually dedicated to graphics processing. When compared to the techniques that can be used for controlling virtual characters in films some of the techniques used in games are rudimentary due to this real-time constraint.

Finally, modern games resemble films in the fact that creators go to great lengths to include intricate storylines and control the building of tension in much the way that film script writers do. This means that games are tested heavily in order to ensure that the game proceeds smoothly and that the level of difficulty is finely tuned so as to always hold the interest of a player. In fact, this testing of games has become something of a science in itself [77]. Using autonomous agents gives game characters the ability to do things that are unexpected by the game designers and so upset their well laid plans. This can often be a barrier to the use of sophisticated techniques such as learning.

Unfortunately there is also a barrier to the discussion of agent-based modeling techniques used in commercial games. Because of the very competitive nature of the games industry, game development houses often consider the details of how their games work as valuable trade secrets to be kept well guarded. This can make it difficult to uncover the details of how particularly interesting features of a game are implemented. While this situation is improving – more commercial game developers are speaking at games conferences about how their games are developed and the release of game systems development kits for the development of game modifications (or *mods*) allows researchers to plumb the depths of game code – it is still often impossible to find out the implementation details of very new games.

Game Genres

Before discussing the use of agent-based modeling in games any further, it is worth making a short clarification on the kinds of computer games that this article refers to. When discussing modern computer games, or video games, this article does not refer to computer implementations of traditional games such as chess, backgammon or card games such as solitaire. Although these games are of considerable research interest (chess in particular has been the subject of extremely successful research [23]) they are typically not approached using agent-based modeling techniques. Typically, artificial intelligence approaches to games such as these rely largely on sophisticated searching techniques which allow the computer player to search through a multitude of possible future situations dictated by the moves it will make and the moves it expects its opponent to make in response. Based on this search, and some clever heuristics that indicate what constitutes a good game position for the computer player, the best sequence of moves can be chosen. This searching technique relies on the fact that there are usually a relatively small number of moves that a player can make at any one time in a game. However, the fact that the ancient Chinese game of Go-Moku has not, to date, been mastered by computer players [80] illustrates the restrictions of such techniques.

The common thread linking together the kinds of games that this article focuses on is that they all contain computer controlled virtual characters that possess

Computer Graphics and Games, Agent Based Modeling in, Figure 2
A screenshot of the upcoming action game Rogue Warrior from Bethesda Softworks (image courtesy of Bethesda Softworks)

a strong notion of agency. Efforts are often made to separate the many different kinds of modern video games that are the focus of this article into a small set of descriptive genres. Unfortunately, much like in music, film and literature, no categorization can hope to perfectly capture the nuances of all of the available titles. However, a brief mention of some of the more important game genres is worth while (a more detailed description of game genres, and artificial intelligence requirements of each is given in [41]).

The most popular game genre is without doubt the *action game* in which the player must defeat waves of demented foes, typically (for increasingly bizarre motivations) bent upon global destruction. Illustrative examples of the genre include *Half-Life 2* (www.half-life2.com) and the *Halo* series (www.halo3.com). A screenshot of the upcoming action game *Rogue Warrior* (www.bethsoft.com) is shown in Fig. 2.

Strategy games allow players to control large armies in battle with other people, or computer controlled opponents. Players do not have direct control over their armies, but rather issue orders which are carried out by agent-based artificial soldiers. Well regarded examples of the genre include the *Age of Empires* (www.ageofempires.com) and *Command & Conquer* (www.commandandconquer.com) series.

Role playing games (such as the *Elder Scrolls* (www.elderscrolls.com) series) place game players in expansive virtual worlds across which they must embark on fantastical quests which typically involve a mixture of solving puzzles, fighting opponents and interacting with non-player characters in order to gain information. Figure 3 shows a screenshot of the aforementioned role-playing game *The Elder Scrolls IV: Oblivion*.

Almost every sport imaginable has at this stage been turned into a computer based *sports game*. The challenges in developing these games are creating computer controlled opponents and team mates that play the games at a level suitable to the human player. Interesting examples include *FIFA Soccer 08* (www.fifa08.ea.com) and *Forza Motorsport 2* (www.forzamotorsport.net).

Finally, many people expected that the rise of massively multi-player online games (MMOGs), in which hundreds of human players can play together in an online world, would sound the death knell for the use of virtual non-player characters in games. Examples of MMOGs include *World of Warcraft* (www.worldofwarcraft.com) and *Battlefield 2142* (www.battlefield.ea.com). However, this has not turned out to be the case as there are still large numbers of single player games being produced and even MMOGs need computer controlled characters for roles that players do not wish to play.

Of course there are many games that simply do not fit into any of these categorizations, but that are still relevant for a discussion of the use of agent-based techniques – for example *The Sims* (www.thesims.ea.com) and the *Microsoft Flight Simulator* series (www.microsoft.com/games/flightsimulatorx). However the categorization still serves to introduce those unfamiliar with the subject to the kinds of games up for discussion.

© 2004 Bethesda Softworks LLC, a ZeniMax Media company. All rights reserved.

Computer Graphics and Games, Agent Based Modeling in, Figure 3
A screenshot from Bethesda Softwork's role playing game The Elder Scrolls IV: Oblivion (image courtesy of Bethesda Softworks)

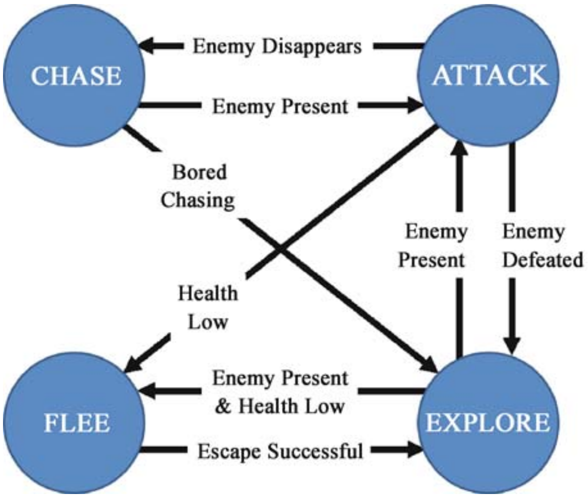
Implementing Agent-Based Modelling Techniques in Games

One of the earliest examples of using agent-based modeling techniques in video games was its application to path planning. The ability of non-player characters (NPCs) to manoeuvre around a game world is one of the most basic competencies required in games. While in very early games it was sufficient to have NPCs move along prescripted paths, this soon become unacceptable. Games programmers soon began to turn to AI techniques which might be applied to solve some of the problems that were arising. The A* path planning algorithm [74] was the first example of such a technique to find wide-spread use in the games industry. Using the A* algorithm NPCs can be given the ability to find their own way around an environment. This was put to particularly fine effect early on in real-time strategy games where the units controlled by players are semi-autonomous and are given orders rather than directly controlled. In order to use the A* algorithm a game world must be divided into a series of cells each of which is given a rating in terms of the effort that must be expended to cross it. The A* algorithm then performs a search across these cells in order to find the shortest path that will take a game agent from a start position to a goal.

Since becoming widely understood amongst the game development community many interesting additions have been made to the basic A* algorithm. It was not long before three dimensional versions of the algorithm became commonly used [71]. The basic notion of storing the energy required to cross a cell within a game world has also been extended to augment cells with a wide range of other useful information (such as the level of danger in crossing a cell) that can be used in the search process [63].

The next advance in the kind of techniques being used to achieve agent-based modeling in games was the finite state machine (FSM) [30]. An FSM is a simple system in which a finite number of *states* are connected in a directed graph by *transitions* between these states. When used for the control of NPCs, the nodes of an FSM indicate the possible actions within a game world that an agent can perform. Transitions indicate how changes in the state of the game world or the character's own attributes (such as health, tiredness etc) can move the agent from one state to another.

Figure 4 shows a sample FSM for the control of an NPC in a typical action game. In this example the behaviors of the character are determined by just four states – CHASE, ATTACK, FLEE and EXPLORE. Each of these states provides an action that the agent should take. For exam-



Computer Graphics and Games, Agent Based Modeling in, Figure 4

A simple finite state machine for a soldier NPC in an action game

ple, when in the EXPLORE state the character should wander randomly around the world, or while in the FLEE state the character should determine a direction to move in that will take it away from its current enemy and move in that direction. The links between the states show how the behaviors of the character should move between the various available states. So, for example, if while in the ATTACK state the agent's health measure becomes low, they will move to the FLEE state and run away from their enemy.

FSMs are widely used because they are so simple, well understood and extremely efficient both in terms of processing cycles required and memory usage. There have also been a number of highly successful augmentations to the basic state machine model to make them more effective, such as the introduction of layers of parallel state machines [3], the use of fuzzy logic in finite state machines [19] and the implementation of cooperative group behaviors through state machines [72].

The action game *Halo 2* is recognized as having a particularly good implementation of state machine based NPC control [79]. At any time an agent could be in any one of the four states *Idle*, *Guard/Patrol*, *Attack/Defend*, and *Retreat*. Within each of these states a set of rules was used in order to select from a small set of appropriate actions for that state – for example a number of different ways to attack the player. The decisions made by NPCs were influenced by a number of character attributes including strength, speed and cowardliness. Transition between states was triggered by perceptions made by characters simulated senses of vision and hearing and by internal attributes such as health. The system implemented also allowed for group behaviors allowing NPCs to hold conversations and cooperate to drive vehicles.

However, FSMs are not without their drawbacks. When designing FSMs developers must envisage every possible situation that might confront an NPC over the course of a game. While this is quite possible for many games, if NPCs are required to move between many different situations this task can become overwhelming. Similarly, as more and more states are added to an FSM designing the links between these states can become a mammoth undertaking.

From [31] the definition of rule based systems states that they are "*… comprised of a database of associated rules. Rules are conditional program statements with consequent actions that are performed if the specified conditions are satisfied*". Rule based systems have been applied extensively to control NPCs in games [16], in particular for the control of NPCs in role-playing games. NPCs behaviors are scripted using a set of rules which typically indicate how an NPC should respond to particular events within the game world. Borrowed from [82], the listing below shows a snippet of the rules used to control a warrior character in the RPG *Baldur's Gate* (www.bioware.com).

```
IF
  // If my nearest enemy is not within 3
  !Range(NearestEnemyOf(Myself),3)
  // and is within 8
  Range(NearestEnemyOf(Myself),8)
THEN
  // 1/3 of the time
  RESPONSE #40
    // Equip my best melee weapon
    EquipMostDamagingMelee()
    // and attack my nearest enemy, checking every 60
    // ticks to make sure he is still the nearest
    AttackReevalutate(NearestEnemyOf (Myself),60)
  // 2/3 of the time
  RESPONSE #60
    // Equip a ranged weapon
    EquipRanged()
    // and attack my nearest enemy, checking every 30
    // ticks to make sure he is still the nearest
    AttackReevalutate(NearestEnemyOf (Myself), 30)
```

The implementation of an NPC using a rule-based system would consist of a large set of such rules, a small set of which would fire based on the conditions in the world at any given time. Rule based systems are favored by game developers as they are relatively simple to use and can be exhaustively tested. Rule based systems also have the advantage that rule sets can be written using simple proprietary scripting systems [9], rather than full programming languages, making them easy to implement. Development companies have also gone so far as to make these scripting languages available to the general public, enabling them to author there own rule sets.

Rule based systems, however, are not without their drawbacks. Authoring extensive rule sets is not a trivial task, and they are usually restricted to simple situations. Also, rule based systems can be restrictive in that they don't allow sophisticated interplay between NPCs motivations, and require that rule set authors foresee every situation that the NPC might find itself in.

Some of the disadvantages of simple rule based systems can be alleviated by using more sophisticated inference engines. One example uses Dempster Schafer theory [43] which allows rules to be evaluated by combining multiple sources of (often incomplete) evidence to determine actions. This goes some way towards supporting the use of rule based systems in situations where complete knowledge is not available.

ALife techniques have also been applied extensively in the control of game NPCs, as much as a philosophy as any particular techniques. The outstanding example of this is *The Sims* (thesims.ea.com) a surprise hit of 2000 which has gone on to become the best selling PC game of all time. Created by games guru Will Wright the Sims puts the player in control of the lives of a virtual family in their virtual home. Inspired by aLife, the characters in the game have a set of motivations, such as hunger, fatigue and boredom and seek out items within the game world that can satisfy these desires. Virtual characters also develop sophisticated social relationships with each other based on common interest, attraction and the amount of time spent together. The original system in the Sims has gone on to be improved in the sequel *The Sims 2* and a series of expansion packs.

Some of the more interesting work in developing techniques for the control of game characters (particularly in action games) has been focused on developing interesting sensing and memory models for game characters. Players expect when playing action games that computer controlled opponents should suffer from the same problems that players do when perceiving the world. So, for example, computer controlled characters should not be able to see through walls or from one floor to the next. Similarly, though, players expect computer controlled characters to be capable of perceiving events that occur in a world and so NPCs should respond appropriately to sound events or on seeing the player.

One particularly fine example of a sensing model was in the game *Thief: The Dark Project* where players are required to sneak around an environment without alerting guards to their presence [45]. The developers produced a relatively sophisticated sensing model that was used by non-player characters which modeled visual effects such as not being able to see the player if they were in shadows, and moving some way towards modeling acoustics so that non-player characters could respond reasonably to sound events.

2004's *Fable* (fable.lionhead.com) took the idea of adding memory to a game to new heights. In this adventure game the player took on the role of a hero from boyhood to manhood. However, every action the player took had an impact on the way in which the game world's population would react to him or her as they would remember every action the next time they met the player. This notion of long-term consequences added an extra layer of believability to the game-playing experience.

Serious Games & Academia

It will probably have become apparent to most readers of the previous section that much of the work done in implementing agent-based techniques for the control of NPCs in commercial games is relatively simplistic when compared to the application of these techniques in other areas of more academic focus, such as robotics [54]. The reasons for this have been discussed already and briefly relate to the lack of available processing resources and the requirements of commercial quality control. However, a large amount of very interesting work is taking place in the application of agent-based technologies in academic research, and in particular the field of serious games. This section will begin by introducing the area of serious games and then go on to discuss interesting academic projects looking at agent-based technologies in games.

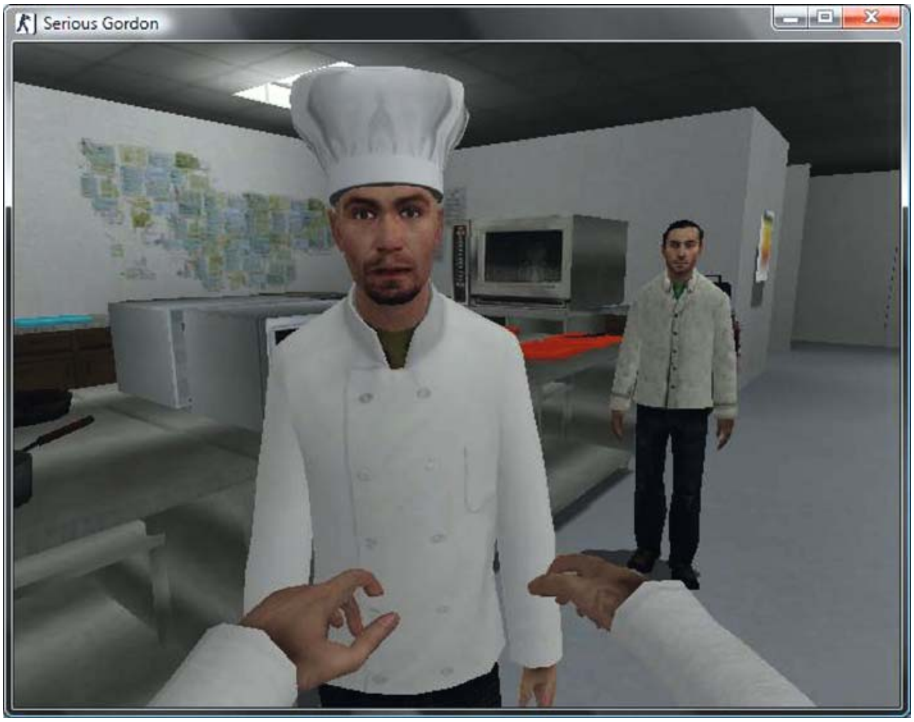
The term serious games [53] refers to games designed to do more than just entertain. Rather, serious games, while having many features in common with conventional games, have ulterior motives such as teaching, training, and marketing. Although games have been used for ends apart from entertainment, in particular education, for a long time, the modern serious games movement is set apart from these by the level of sophistication of the games it creates. The current generation of serious games is comparable with main-stream games in terms of the quality of production and sophistication of their design. Serious games offer particularly interesting opportunities for the use of agent-based modeling techniques due to the facts that they often do not have to live up to the rigorous testing of commercial games, can have the requirement of specialized hardware rather than being restricted to commercial games hardware and often, by the nature of their application domains, require more in-depth interactions between players and NPCs.

The modern serious games movement can be said to have begun with the release of *America's Army* (www.americasarmy.com) in 2002 [57]. Inspired by the realism of commercial games such as the *Rainbow 6* series (www.rainbow6.com), the United States military developed America's Army and released it free of charge in order to give potential recruits a flavor of army life. The game was hugely successful and is still being used today as both a recruitment tool and as an internal army training tool.

Spurred on by the success of America's Army the serious games movement began to grow, particularly within academia. A number of conferences sprung up and notably the Serious Games Summit became a part of the influential Game Developer's Conference (www.gdconf.com) in 2004.

Some other notable offerings in the serious games field include *Food Force* (www.food-force.com) [18], a game developed by the United Nations World Food Programme in order to promote awareness of the issues surrounding emergency food aid; *Hazmat Hotzone* [15], a game developed by the Entertainment Technology Centre at Carnegie Mellon University to train fire-fighters to deal with chemical and hazardous materials emergencies; *Yourself!Fitness* (www.yourselffitness.com) [53] an interactive virtual personal trainer developed for modern games consoles; and *Serious Gordon* (www.seriousgames.ie) [50] a game developed to aid in teaching food safety in kitchens. A screen shot of Serious Gordon is shown in Fig. 5.

Over the past decade, interest in academic research that is directly focused on artificial intelligence, and in particular agent-based modelling techniques and their application to games (as opposed to the general virtual character/computer graphics work discussed previously) has grown dramatically. One of the first major academic research projects into the area of Game-AI was led by John Laird at the University of Michigan, in the United States. The SOAR architecture was developed in the early nineteen eighties in an attempt to *"develop and apply a unified theory of human and artificial intelligence"* [66]. SOAR is essentially a rule based inference system which takes the current state of a problem and matches this to production rules which lead to actions.



Computer Graphics and Games, Agent Based Modeling in, Figure 5
A screenshot of Serious Gordon, a serious game developed to aid in the teaching of food safety in kitchens

After initial applications into the kind of simple puzzle worlds which characterized early AI research [42], the SOAR architecture was applied to the task of controlling computer generated forces [37]. This work lead to an obvious transfer to the new research area of game-AI [40].

Initially the work of Laird's group focused on applying the SOAR architecture to the task of controlling NPC opponents in the action game *Quake* (www.idsoftware.com) [40]. This proved quite successful leading to opponents which could successfully play against human players, and even begin to plan based on anticipation of what the player was about to do. More recently Laird's group have focused on the development of a game which requires more involved interactions between the player and the NPCs. Named *Haunt 2*, this game casts the player in the role of a ghost that must attempt to influence the actions of a group of computer controlled characters inhabiting the ghost's haunted house [51]. The main issue that arises with the use the SOAR architecture is that it is enormously resource hungry, with the NPC controllers running on a separate machine to the actual game.

At Trinity College in Dublin in Ireland, the author of this article worked on an intelligent agent architecture, the Proactive Persistent Agent (PPA) architecture, for the control of background characters (or support characters) in character-centric games (games that focus on character interactions rather than action, e. g. role-playing games) [48,49]. The key contributions of this work were that it made possible the creation of NPCs that were capable of behaving believably in a wide range of situations and allowed for the creation of game environments which it appeared had an existence beyond their interactions with players. Agent behaviors in this work were based on models of personality, emotion, relationships to other characters and behavioral models that changed according to the current role of an agent. This system was used to develop a stand alone game and as part of a simulation of areas within Trinity College. A screenshot of this second application is shown in Fig. 6.

At Northwestern University in Chicago the Interactive Entertainment group has also applied approaches from more traditional research areas to the problems facing game-AI. Ian Horswill has led a team that are attempting to use architectures traditionally associated with robotics for the control of NPCs. In [29] Horswill and Zubek consider how perfectly matched the behavior based architectures often used in robotics are with the requirements of NPC control architectures. The group have demonstrated some of their ideas in a test-bed environment built on top

Computer Graphics and Games, Agent Based Modeling in, Figure 6
Screenshots of the PPA system simulating parts of a college

of the game Half-Life [38]. The group also looks at issues around character interaction [85] and the many psychological issues associated with creating virtual characters asking how we can create virtual game agents that display all of the foibles that make us relate to characters in human stories [28].

Within the same research group a team led by Ken Forbus have extended research previously undertaken in conjunction with the military [24] and applied it to the problem of terrain analysis in computer strategy games [25]. Their goal is to create strategic opponents which are capable of performing sophisticated reasoning about the terrain in a game world and using this knowledge to identify complex features such as ambush points. This kind of high level reasoning would allow AI opponents play a much more realistic game, and even surprise human players from time to time, something that is sorely missing from current strategy games.

As well as this work which has spring-boarded from existing applications, a number of projects began expressly to tackle problems in game-AI. Two which particularly stand out are the Excalibur Project, led by Alexander Nareyek [55] and work by John Funge [26]. Both of these projects have attempted to applying sophisticated planning techniques to the control of game characters.

Nareyek uses constraint based planning to allow game agents reason about their world. By using techniques such as local search Nareyek has attempted to allow these sophisticated agents perform resource intensive planning within the constraints of a typical computer game environment. Following on from this work, the term *anytime agent* was coined to describe the process by which agents actively refine original plans based on changing world conditions. In [56] Narayek describes the directions in which he intends to take this work in future.

Funge uses the situational calculus to allow agents reason about their world. Similarly to Nareyek he has addressed the problems of a dynamic, ever changing world, plan refining and incomplete information. Funge's work uses an extension to the situational calculus which allows the expression of uncertainty. Since completing this work Funge has gone on to be one of the founders of AiLive (www.ailive.net), a middleware company specializing in AI for games.

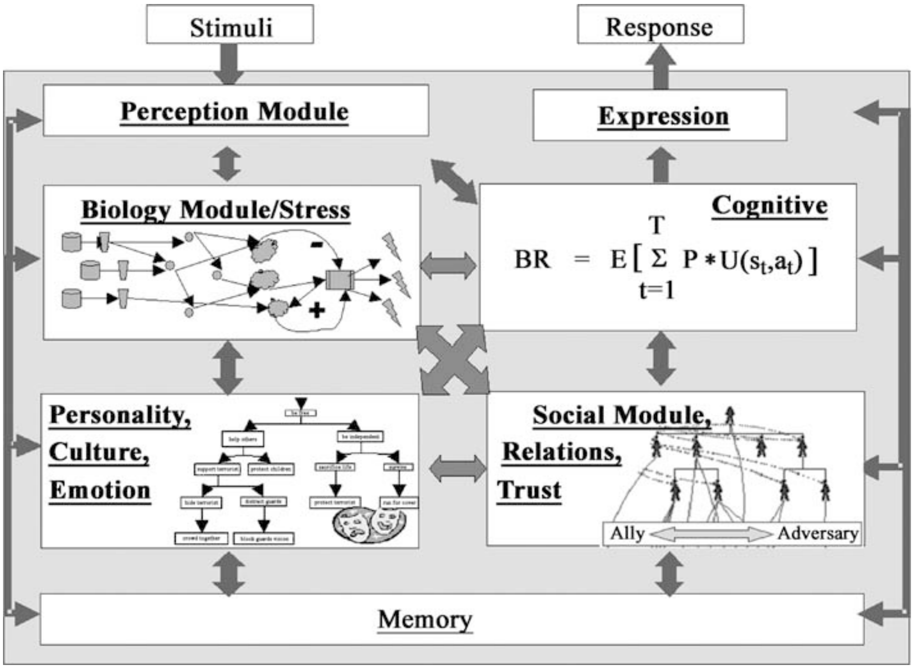
While the approaches of both of these projects have shown promise within the constrained environments to which they have been applied during research, (and work continues on them) it remains to be seen whether such techniques can be successfully applied to a commercial game environment and all of the resource constraints that such an environment entails.

One of the most interesting recent examples of agent-based work in the field of serious games is that undertaken by Barry Silverman and his group at the University of Pennsylvania in the United States [69,70]. Silverman models the protagonists in military simulations for use in training programmes and has taken a very interesting approach in that his agent models are based on established cognitive science and behavioral science research. While Silverman admits that many of the models described in the cognitive science and behavioral science literature are not well quantified enough to be directly implemented, he has adapted a number of well respected models for his purposes. Silverman's work is an excellent example of the capabilities that can be explored in a serious games setting rather than a commercial game setting, and as such merits an in depth discussion. A high-level schematic diagram of Silverman's approach is shown in Fig. 7 and shows the agent architecture used by Silverman's system, PMFserv.

The first important component of the PMFserv system is the biology module which controls biological needs using a metaphor based on the flow of water through a system. Biological concepts such as hunger and fatigue are simulated using a series of reservoirs, tanks and valves which model the way in which resources are consumed by the system. This biological model is used in part to model stress which has an important impact on the way in which agents make decisions. To model the way in which agent performance changes under pressure Silverman uses *performance moderator functions* (PMFs). An example of one of the earliest PMFs used is the Yerkes–Dodson "inverted-u" curve [84] which illustrates that as mental arousal is increased performance initially improves, peaks and then trails off again. In PMFserv a range of PMFs are used to model the way in which behavior should change depending on stress levels and biological conditions.

The second important module of PMFserv attempts to model how personality, culture and emotion affect the behavior of an agent. In keeping with the rest of their system PMFserv uses models inspired by cognitive science to model emotions. In this case the well known OCC model [58], which has been used in agent-based applications before [8], is used. The OCC model provides for 11 pairs of opposite emotions such as pride and shame, and hope and fear. The emotional state of an agent with regard to past, current and future actions heavily influences the decisions that the agent makes.

The second portion of the Personality, Culture, Emotion module uses a value tree in order to capture the values of an agent. These values are divided into a Preference Tree which captures long term desired states for the world, a Standards Tree which relates to the actions that an agent



Computer Graphics and Games, Agent Based Modeling in, Figure 7
A schematic diagram of the main components of the PMFserv system (with kind permission of Barry Silverman)

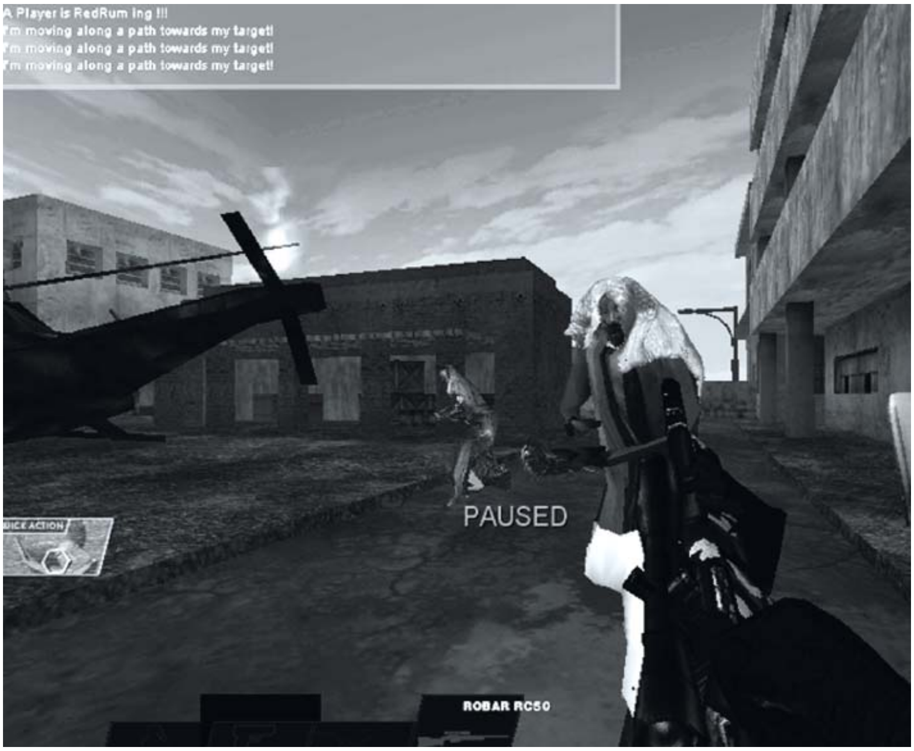
believes it can or cannot follow in order to achieve these desired states and a Goal Tree which captures short term goals.

PMFserv also models the relationships between agents (Social Model, Relations, Trust in Fig. 7). The relationship of one agent to another is modeled in terms of three axes. The first is the degree to which the other agent is thought of as a human rather than an inanimate object – locals tend to view American soldiers as objects rather than people. The second axis is the cognitive grouping (ally, foe etc) to which the other agent belongs and whether this is also a group to which the first agent has an affinity. Finally, the valence, or strength, of the relationship is stored. Relationships continually change based on actions that occur within the game world. Like the other modules of the system this model is also based on psychological research [58].

The final important module of the PMFserv architecture is the Cognitive module which is used to decide on particular actions that agents will undertake. This module uses inputs from all of the other modules to make these decisions and so the behavior of PMFserv agents is driven by their stress levels, relationships to other agents and objects within the game world, personality, culture and emotions. The details of the PMFserv cognitive process are beyond the scope of this article, so it will suffice to say that action selection is based on a calculation of the utility of a particular action to an agent, with this calculation modified by the factors listed above.

The most highly developed example using the PMFserv model is a simulation of the 1993 event in Mogadishu, Somalia in which a United States military Black Hawk helicopter crashed, as made famous by the book and film "Black Hawk Down" [12]. In this example, which was developed as a military training aid as part of a larger project looking at agent implementations within such systems [78,81] the player took on the role of a US army ranger on a mission to secure the helicopter wreck in a modification (or "mod") of the game *Unreal Tournament* (www.unreal.com). A screenshot of this simulation is shown in Fig. 8.

The PMFserv system was used to control the behaviors of characters within the game world such as Somali militia, and Somali civilians. These characters were imbued with physical attributes, a value system and relationships with other characters and objects within the game environment. The sophistication of PMFserv was apparent in many of the behaviors of the simulations NPCs. One particularly good example was the fact that Somali women would offer themselves as human shields for militia fighters. This behavior was never directly programmed into the agents make-up, but rather emerged as a result of their values and assessment of their situation. PMFserv remains one of the most sophisticated current agent implementa-


A Player is RedRum ing !!!
I'm moving along a path towards my target!
I'm moving along a path towards my target!
I'm moving along a path towards my target!

Computer Graphics and Games, Agent Based Modeling in, Figure 8
A screenshot of the PMFserv system being used to simulate the Black Hawk Down scenario (with kind permission of Barry Silverman)

tions and shows the possibilities when the shackles of commercial game constraints are thrown off.

Future Directions

There is no doubt that with the increase in the amount of work being focused on the use of agent-based modeling in computer graphics and games there will be major developments in the near future. This final section will attempt to predict what some of these might be.

The main development that might be expected in all of the areas that have been discussed in this article is an increase in the depth of simulation. The primary driver of this increase in depth will be the development of more sophisticated agent models which can be used to drive ever more sophisticated agent behavior. The PMFserv system described earlier is one example of the kinds of deeper systems that are currently being developed. In general computer graphics applications this will allow for the creation of more interesting simulations including previously prohibitive features such as automatic realistic facial expressions and other physical expressions of agents' internal states. This would be particularly use in CGI for movies in which, although agent based modeling techniques are commonly used for crowd scenes and background characters, main characters are still animated almost entirely by hand.

In the area of computer games it can be expected that many of the techniques being used in movie CGI will filter over to real-time game applications as the processing power of game hardware increases – this is a pattern that has been evident for the past number of years. In terms of depth that might be added to the control of game characters one feature that has mainly been conspicuous by its absence in modern games is genuine learning by game agents. 2000's *Black & White* and its sequel *Black & White 2* (www.lionhead.com) featured some learning by one of the game's main characters that the player could teach in a reinforcement manner [20]. While this was particularly successful in the game, such techniques have not been more widely applied. One interesting academic project in this area is the NERO project (www.nerogame.org) which allows a player to train an evolving army of soldiers and have them battle the armies of other players [73]. It is expected that these kinds of capabilities will become more and more common in commercial games.

One new feature of the field of virtual character control in games is the emergence of specialized middleware. Middleware has had a massive impact in other areas of game development including character modeling (for example Maya available from www.autodesk.com) and physics modeling (for example Havok available from www.havok.com). AI focused middleware for games is now becoming more common with notable offerings including AI-Implant (www.ai-implant.com) and Kynogon (www.kynogon.com) which perform path finding and state machine based control of characters. It is expected that more sophisticated techniques will over time find their way into such software.

To conclude the great hope for the future is that more and more sophisticated agent-based modeling techniques from other application areas and other branches of AI will find their way into the control of virtual characters.

Bibliography

Primary Literature

1. Adamson A (Director) (2005) The Chronicles of Narnia: The Lion, the Witch and the Wardrobe. Motion Picture. http://adisney.go.com/disneypictures/narnia/lb_main.html
2. Aitken M, Butler G, Lemmon D, Saindon E, Peters D, Williams G (2004) The Lord of the Rings: the visual effects that brought middle earth to the screen. International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Course Notes
3. Alexander T (2003) Parallel-State Machines for Believable Characters. In: Massively Multiplayer Game Development. Charles River Media
4. Allers R, Minkoff R (Directors) (1994) The Lion King. Motion Picture. http://disney.go.com/disneyvideos/animatedfilms/lionking/
5. Aylett R, Luck M (2000) Applying Artificial Intelligence to Virtual Reality: Intelligent Virtual Environments. Appl Artif Intell 14(1):3–32
6. Badler N, Bindiganavale R, Bourne J, Allbeck J, Shi J, Palmer M (1999) Real Time Virtual Humans. In: Proceedings of the International conference on Digital Media Futures.
7. Bates J (1992) The Nature of Characters in Interactive Worlds and the Oz Project. Technical Report CMU-CS-92-200. School of Computer Science, Carnegie Melon University
8. Bates J (1992) Virtual reality, art, and entertainment. Presence: J Teleoper Virtual Environ 1(1):133–138
9. Berger L (2002) Scripting: Overview and Code Generation. In: Rabin S (ed) AI Game Programming wisdom. Charles River Media
10. Bird B, Pinkava J (Directors) (2007) Ratatouille. Motion Picture. http://disney.go.com/disneyvideos/animatedfilms/ratatouille/
11. Blumberg B (1996) Old Tricks, New Dogs: Ethology and Interactive Creatures. Ph D Thesis, Media Lab, Massachusetts Institute of Technology
12. Bowden M (2000) Black Hawk Down. Corgi Adult
13. Burke R, Isla D, Downie M, Ivanov Y, Blumberg B (2002) Creature Smarts: The Art and Architecture of a Virtual Brain. In: Proceedings of Game-On 2002: the 3rd International Conference on Intelligent Games and Simulation, pp 89–93
14. Burton T (Director) (1992) Batman Returns. Motion Picture. http://www.warnervideo.com/batmanmoviesondvd/
15. Carless S (2005) Postcard From SGS 2005: Hazmat: Hotzone – First-Person First Responder Gaming. Retrieved October 2007, from Gamasutra: www.gamasutra.com/features/20051102/carless_01b.shtml
16. Christian M (2002) A Simple Inference Engine for a Rule Based Architecture. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
17. Darnell E, Johnson T (Directors) (1998) Antz. Motion Picture. http://www.dreamworksanimation.com/
18. DeMaria R (2005) Postcard from the Serious Games Summit: How the United Nations Fights Hunger with Food Force. Retrieved October 2007, from Gamasutra: www.gamasutra.com/features/20051104/demaria_01.shtml
19. Dybsand E (2001) A Generic Fuzzy State Machine in C++. In: Rabin S (ed) Game Programming Gems 2. Charles River Media
20. Evans R (2002) Varieties of Learning. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
21. Faloutsos P, van de Panne M, Terzopoulos D (2001) The Virtual Stuntman: Dynamic Characters with a Repetoire of Autonomous Motor Skills. Comput Graph 25(6):933–953
22. Farenc N, Musse S, Schweiss E, Kallmann M, Aune O, Boulic R et al (2000) A Paradigm for Controlling Virtual Humans in Urban Environment Simulations. Appl Artif Intell J Special Issue Intell Virtual Environ 14(1):69–91
23. Feng-Hsiung H (2002) Behind Deep Blue: Building the Computer that Defeated the World Chess Champion. Princeton University Press
24. Forbus K, Nielsen P, Faltings B (1991) Qualitative Spatial Reasoning: The CLOCK Project. Artif Intell 51:1–3
25. Forbus K, Mahoney J, Dill K (2001) How Qualitative Spatial Reasoning Can Improve Strategy Game AIs. In: Proceedings of the AAAI Spring Symposium on AI and Interactive Entertainment
26. Funge J (1999) AI for Games and Animation: A Cognitive Modeling Approach. A.K. Peters
27. Hayes-Roth B, Doyle P (1998) Animate Characters. Auton Agents Multi-Agent Syst 1(2):195–230
28. Horswill I (2007) Psychopathology, narrative, and cognitive architecture (or: why NPCs should be just as screwed up as we are). In: Proceedings of AAAI Fall Symposium on Intelligent Narrative Technologies
29. Horswill I, Zubek R (1999) Robot Architectures for Believable Game Agents. In: Proceedings of the 1999 AAAI Spring Symposium on Artificial Intelligence and Computer Games
30. Houlette R, Fu D (2003) The Ultimate Guide to FSMs in Games. In: Rabin S (ed) AI Game Programming Wisdom 2. Charles River Media
31. IGDA (2003) Working Group on Rule-Based Systems Report. International Games Development Association
32. Isbister K, Doyle P (2002) Design and Evaluation of Embodied Conversational Agents: A Proposed Taxonomy. In: Proceedings of the AA- MAS02 Workshop on Embodied Conversational Agents: Lets Specify and Compare Them! Bologna, Italy
33. Jackson P (Director) (2001) The Lord of the Rings: The Fellowship of the Ring. Motion Picture. http://www.lordoftherings.net/
34. Jackson P (Director) (2002) The Lord of the Rings: The Two Towers. Motion Picture. http://www.lordoftherings.net/
35. Jackson P (Director) (2003) The Lord of the Rings: The Return of the King. Motion Picture. http://www.lordoftherings.net/

36. Johnston O, Thomas F (1995) The Illusion of Life: Disney Animation. Disney Editions
37. Jones R, Laird J, Neilsen P, Coulter K, Kenny P, Koss F (1999) Automated Intelligent Pilots for Combat Flight Simulation. AI Mag 20(1):27–42
38. Khoo A, Zubek R (2002) Applying Inexpensive AI Techniques to Computer Games. IEE Intell Syst Spec Issue Interact Entertain 17(4):48–53
39. Koeppel D (2002) Massive Attack. http://www.popsci.com/popsci/science/d726359b9fa84010vgnvcm1000004eecbccdrcrd.html. Accessed Oct 2007
40. Laird J (2000) An Exploration into Computer Games and Computer Generated Forces. The 8th Conference on Computer Generated Forces and Behavior Representation
41. Laird J, van Lent M (2000) Human-Level AI's Killer Application: Interactive Computer Games. In: Proceedings of the 17th National Conference on Artificial Intelligence
42. Laird J, Rosenbloom P, Newell A (1984) Towards Chunking as a General Learning Mechanism. The 1984 National Conference on Artificial Intelligence (AAAI), pp 188–192
43. Laramée F (2002) A Rule Based Architecture Using Dempster-Schafer theory. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
44. Lasseter J, Stanton A (Directors) (1998) A Bug's Life; Motion Picture. http://www.pixar.com/featurefilms/abl/
45. Leonard T (2003) Building an AI Sensory System: Examining the Deign of Thief: The Dark Project. In: Proceedings of the 2003 Game Developers' Conference, San Jose
46. Loyall B (1997) Believable Agents: Building Interactive Personalities. Ph D Thesis, Carnegie Melon University
47. Määta A (2002) Realistic Level Design for Max Payne. In: Proceedings of the 2002 Game Developer's conference, GDC 2002
48. Mac Namee B, Cunningham P (2003) Creating Socially Interactive Non Player Characters: The μ-SIC System. Int J Intell Games Simul 2(1)
49. Mac Namee B, Dobbyn S, Cunningham P, O'Sullivan C (2003) Simulating Virtual Humans Across Diverse Situations. In: Proceedings of Intelligent Virtual Agents '03, pp 159–163
50. Mac Namee B, Rooney P, Lindstrom P, Ritchie A, Boylan F, Burke G (2006) Serious Gordon: Using Serious Games to Teach Food Safety in the Kitchen. The 9th International Conference on Computer Games: AI, Animation, Mobile, Educational & Serious Games CGAMES06, Dublin
51. Magerko B, Laird JE, Assanie M, Kerfoot A, Stokes D (2004) AI Characters and Directors for Interactive Computer Games. The 2004 Innovative Applications of Artificial Intelligence Conference. AAAI Press, San Jose
52. Thalmann MN, Thalmann D (1994) Artificial Life and Virtual Reality. Wiley
53. Michael D, Chen S (2005) Serious Games: Games That Educate, Train, and Inform. Course Technology PTR
54. Muller J (1996) The Design of Intelligent Agents: A Layered Approach. Springer
55. Nareyek A (2001) Constraint Based Agents. Springer
56. Nareyek A (2007) Game AI is Dead. Long Live Game AI! IEEE Intell Syst 22(1):9–11
57. Nieborg D (2004) America's Army: More Than a Game. Bridging the Gap;Transforming Knowledge into Action through Gaming and Simulation. Proceedings of the 35th Conference of the International Simulation and Gaming Association (ISAGA), Munich
58. Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
59. Perlin K, Goldberg A (1996) Improv: A System for Scripting Interactive Actors in Virtual Worlds. In: Proceedings of the ACM Computer Graphics Annual Conference, pp 205–216
60. Proyas A (Director) (2004) I, Robot. Motion Picture. http://www.irobotmovie.com
61. Rao AS, Georgeff MP (1991) Modeling rational agents within a BDI-architecture. In: Proceedings of Knowledge Representation and Reasoning (KR&R-91). Morgan Kaufmann, pp 473–484
62. Musse RS, Thalmann D (2001) A Behavioral Model for Real Time Simulation of Virtual Human Crowds. IEEE Trans Vis Comput Graph 7(2):152–164
63. Reed C, Geisler B (2003) Jumping, Climbing, and Tactical Reasoning: How to Get More Out of a Navigation System. In: Rabin S (ed) AI Game Programming Wisdom 2. Charles River Media
64. Reynolds C (1987) Flocks, Herds and Schools: A Distributed Behavioral Model. Comput Graph 21(4):25–34
65. Rodriguez R (Director) (1996) From Dusk 'Till Dawn. Motion Picture
66. Rosenbloom P, Laird J, Newell A (1993) The SOAR Papers: Readings on Integrated Intelligence. MIT Press
67. Sánchez-Crespo D (2006) GDC: Physical Gameplay in Half-Life 2. Retrieved October 2007, from gamasutra.com: http://www.gamasutra.com/features/20060329/sanchez_01.shtml
68. Shao W, Terzopoulos D (2005) Autonomous Pedestrians. In: Proceedings of SIGGRAPH/EG Symposium on Computer Animation, SCA'05, pp 19–28
69. Silverman BG, Bharathy G, O'Brien K, Cornwell J (2006) Human Behavior Models for Agents in Simulators and Games: Part II: Gamebot Engineering with PMFserv. Presence Teleoper Virtual Worlds 15(2):163–185
70. Silverman BG, Johns M, Cornwell J, O'Brien K (2006) Human Behavior Models for Agents in Simulators and Games: Part I: Enabling Science with PMFserv. Presence Teleoper Virtual Environ 15(2):139–162
71. Smith P (2002) Polygon Soup for the Programmer's Soul: 3D Path Finding. In: Proceedings of the Game Developer's Conference 2002, GDC2002
72. Snavely P (2002) Agent Cooperation in FSMs for Baseball. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
73. Stanley KO, Bryant BD, Karpov I, Miikkulainen R (2006) Real-Time Evolution of Neural Networks in the NERO Video Game. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, AAAI-2006. AAAI Press, pp 1671–1674
74. Stout B (1996) Smart Moves: Intelligent Path-Finding. Game Dev Mag Oct
75. Takahashi TS (1992) Behavior Simulation by Network Model. Memoirs of Kougakuin University 73, pp 213–220
76. Terzopoulos D, Tu X, Grzeszczuk R (1994) Artificial Fishes with Autonomous Locomotion, Perception, Behavior and Learning, in a Physical World. In: Proceedings of the Artificial Life IV Workshop. MIT Press
77. Thompson C (2007) Halo 3: How Microsoft Labs Invented a New Science of Play. Retrieved October 2007, from wired.com: http://www.wired.com/gaming/virtualworlds/magazine/15-09/ff_halo
78. Toth J, Graham N, van Lent M (2003) Leveraging gaming in DOD modelling and simulation: Integrating performance and behavior moderator functions into a general cognitive archi-

tecture of playing and non-playing characters. Twelfth Conference on Behavior Representation in Modeling and Simulation (BRIMS, formerly CGF), Scotsdale, Arizona

79. Valdes R (2004) In the Mind of the Enemy: The Artificial Intelligence of Halo 2. Retrieved October 2007, from HowStuffWorks.com: http://entertainment.howstuffworks.com/halo2-ai.htm
80. van der Werf E, Uiterwijk J, van den Herik J (2002) Programming a Computer to Play and Solve Ponnuki-Go. In: Proceedings of Game-On 2002: The 3rd International Conference on Intelligent Games and Simulation, pp 173–177
81. van Lent M, McAlinden R, Brobst P (2004) Enhancing the behavioral fidelity of synthetic entities with human behavior models. Thirteenth Conference on Behavior Representation in Modeling and Simulation (BRIMS)
82. Woodcock S (2000) AI Roundtable Moderator's Report. In: Proceedings of the Game Developer's Conference 2000 (GDC2000)
83. Wooldridge M, Jennings N (1995) Intelligent Agents: Theory and Practice. Know Eng Rev 10(2):115–152
84. Yerkes RW, Dodson JD (1908) The relation of strength of stimulus to rapidity of habit formation. J Comp Neurol Psychol 18:459–482
85. Zubek R, Horswill I (2005) Hierarchical Parallel Markov Models of Interaction. In: Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference, AIIDE 2005

Books and Reviews

DeLoura M (ed) (2000) Game Programming Gems. Charles River Media

DeLoura M (ed) (2001) Game Programming Gems 2. Charles River Media

Dickheiser M (ed) (2006) Game Programming Gems 6. Charles River Media

Kirmse A (ed) (2004) Game Programming Gems 4. Charles River Media

Pallister K (ed) (2005) Game Programming Gems 5. Charles River Media

Rabin S (ed) (2002) Game AI Wisdom. Charles River Media

Rabin S (ed) (2003) Game AI Wisdom 2. Charles River Media

Rabin S (ed) (2006) Game AI Wisdom 3. Charles River Media

Russell S, Norvig P (2002) Artificial Intelligence: A Modern Approach. Prentice Hall

Treglia D (ed) (2002) Game Programming Gems 3. Charles River Media

Computing in Geometrical Constrained Excitable Chemical Systems

JERZY GORECKI[1,2], JOANNA NATALIA GORECKA[3]

[1] Institute of Physical Chemistry,
Polish Academy of Science, Warsaw, Poland

[2] Faculty of Mathematics and Natural Sciences,
Cardinal Stefan Wyszynski University, Warsaw, Poland

[3] Institute of Physics, Polish Academy of Science,
Warsaw, Poland

Article Outline

Glossary

Some of the terms used in our article are the same as in the article of Adamatzky in this volume, and they are explained in the glossary of ► Reaction-Diffusion Computing.

Activator A substance that increases the rate of reaction.

Excitability Here we call a dynamical system excitable if it has a single stable state (the rest state) with the following properties: if the rest state is slightly perturbed then the perturbation uniformly decreases as the system evolves towards it. However, if the perturbation is sufficiently strong it may grow by orders of magnitude before the system approaches the rest state. The increase in variables characterizing the system (usually rapid if compared with the time necessary to reach the rest state) is called an excitation. A forest is a classical example of excitable medium and a wildfire that burns it is an excitation. A dynamical system is non-excitable if applied perturbations do not grow up and finally decay.

Excitability level The measure of how strong a perturbation has to be applied to excite the system. For example, for the Ru-catalyzed Belousov–Zhabotinsky reaction, increasing illumination makes the medium less excitable. The decrease in the excitability level can be observed in reduced amplitudes of spikes and in decreased velocities of autowaves.

Firing number The ratio between the number of generated excitations to the number of applied external stimulations. In most of the cases we define the firing number as the ratio between the number of output spikes to the number of arriving pulses.

Inhibitor A substance that decreases the rate of reaction or even prevents it.

Medium In the article we consider a chemical medium, fully characterized by local concentrations of reagents and external conditions like temperature or illumination level. The time evolution of concentrations is governed by a set of reaction-diffusion equations, where the reaction term is an algebraic function of variables characterizing the system and the non-local coupling is described by the diffusion operator. We are mainly concerned with a two dimensional medium (i. e., a membrane with a solution of reagents used in experiments), but the presented ideas can be also applied to one-dimensional or three-dimensional media.

Refractory period A period of time during which an excitable system is incapable of repeating its response to an applied, strong perturbation. After the refractory period the excitable medium is ready to produce an excitation as an answer to the stimulus.

Spike, autowave In a spatially distributed excitable medium, a local excitation can spread around as a pulse of excitation. Usually a propagating pulse of excitation converges to a stationary shape characteristic for the medium, which is not dependent on initialization and propagates with a constant velocity – thus it is called an autowave.

Subexcitability A system is called subexcitable if the amplitude and size of the initiated pulse of excitation decreases in time. However, if the decay time is comparable with the characteristic time for the system, defined as the ratio between system size and pulse velocity, then pulses in a subexcitable system travel for a sufficiently long distance to carry information. Subexcitable media can be used to control the amplitude of excitations. Subexcitability is usually related to system dynamics, but it may also appear as the result of geometrical constraints. For example, a narrow excitable channel surrounded by a non-excitable medium may behave as a subexcitable system because a propagating pulse dies out due to the diffusion of the activator in the neighborhood.

Definition of the Subject

It has been shown in the article of Adamatzky ▶ Reaction-Diffusion Computing that an excitable system can be used as an information processing medium. In such a medium, information is coded in pulses of excitation; the presence of a single excitation or of a group of excitations forms a message. Information processing discussed by Adamatzky is based on a homogeneous excitable medium and the interaction between pulses in such medium. Here we focus our attention on a quite specific type of excitable medium that has an intentionally introduced structure of regions characterized by different excitability levels. As the simplest case we consider a medium composed of excitable regions where autowaves can propagate and non-excitable ones where excitations rapidly die. Using such two types of medium one can, for example, construct signal channels: stripes of excitable medium where pulses can propagate surrounded by non-excitable areas thick enough to cancel potential interactions with pulses propagating in neighboring channels. Therefore, the propagation of a pulse along a selected line in an excitable system can be realized in two ways. In a homogeneous excitable medium, it can be done by a continuous control of pulse propagation and the local feedback with activating and inhibiting factors (▶ Reaction-Diffusion Computing). In a structured excitable medium, the same result can be achieved by creating a proper pattern of excitable and non-excitable regions. The first method gives more flexibility, the second is just simpler and does not require a permanent watch. As we show in this article, a number of devices that perform simple information processing operations, including the basic logic functions, can be easily constructed with structured excitable medium. Combining these devices as building blocks we can perform complex signal processing operations. Such an approach seems similar to the development of electronic computing where early computers were built of simple integrated circuits.

The research on information processing with structured excitable media has been motivated by a few important problems. First, we would like to investigate how the properties of a medium can be efficiently used to construct devices performing given functions, and what tasks are the most suitable for chemical computing. There is also a question of generic designs valid for any excitable medium and specific ones that use unique features of a particular system (for example, a one-dimensional generator of excitation pulses that can be built with an excitable surface reaction [23]). In information processing with a structured excitable medium, the geometry of the medium is as important as its dynamics, and it seems interesting to know what type of structures are related to specific functions. In the article we present a number of such structures characteristic for particular information processing operations.

Another important motivation for research comes from biology. Even the simplest biological organisms can process information and make decisions important for their life without CPU-s, clocks or sequences of commands as it is in the standard von Neumann computer architecture [15]. In biological organisms, even at a very basic cellular level, excitable chemical reactions are responsible for information processing. The cell body considered as an information processing medium is highly structured. We believe that analogs of geometrical structures used for certain types of information processing operations in structured excitable media will be recognized in biological systems, so we will better understand their role in living organisms. At a higher level, the analogies between information processing with chemical media and signal processing in the brain seems to be even closer because the excitable dynamics of calcium in neural tissue is responsible for signal propagation in nerve system [30]. Excitable chemical channels that transmit signals between processing elements look similar to dendrites and axons. As we show in the article, the biological neuron has its chemical analog, and this allows for the construction of artificial neural networks using chemical processes. Having in mind that neural networks are less vulnerable to random errors than classical algorithms one can go back from biology to man-made computing and adopt the concepts in a fast excitable medium, for example especially prepared semiconductors (▶ Unconventional Computing, Novel Hardware for).

The article is organized in the following way. In the next section we discuss the basic properties of a structured chemical medium that seem useful for information processing. Next we consider the binary information coded in propagating pulses of concentration and demonstrate how logic gates can be built. In the following chapter we show that a structured excitable medium can acquire information about distances and directions of incoming stimuli. Next we present a simple realization of read-write memory cell, discuss its applications in chemical counting devices, and show its importance for programming with pulses of excitation. In the following section we present a chemical realization of artificial neurons that perform multiargument operation on sets of input pulses. Finally we discuss the perspectives of the field, in particular more efficient methods of information coding and some ideas of self-organization that can produce structured media capable of information processing.

Introduction

Excitability is the wide spread behavior of far-from-equilibrium systems [35,39] observed, for example, in chemical reactions (Bielousov–Zhabotynsky BZ-reaction [45], CO oxidation on Pt [37], combustion of gases [26]) as well as in many other physical (laser action) and biochemical (signaling in neural systems, contraction of cardiovascular tissues) processes [51]. All types of excitable systems share a common property that they have a stable stationary state (the rest state) they reside in when they are not perturbed. A small perturbation of the rest state results only in a small-amplitude linear response of the system that uniformly decays in time. However, if a perturbation is sufficiently large then the system can evolve far away from the rest state before finally returning to it. This response is strongly nonlinear and it is accompanied by a large excursion of the variables through phase space, which corresponds to an excitation peak (a spike). The system is refractory after a spike, which means that it takes a certain recovery time before another excitation can take place. The excitability is closely related with relaxation oscillations and the phenomena differ by one bifurcation only [56].

Properties of excitable systems have an important impact on their ability to process information. If an excitable medium is spatially distributed then an excitation at one point of the medium (usually seen as an area with a high concentration of a certain reagent), may introduce a sufficiently large perturbation to excite the neighboring points as the result of diffusion or energy transport. Therefore, an excitation can propagate in space in the form of a pulse. Unlike mechanical waves that dissipate the initial energy and finally decay, traveling spikes use the energy of the medium to propagate and dissipate it. In a typical excitable medium, after a sufficiently long time, an excitation pulse converges to the stationary shape, independent of the initial condition what justifies to call it an autowave. Undamped propagation of signals is especially important if the distances between the emitting and receiving devices are large. The medium's energy comes from the nonequilibrium conditions at which the system is kept. In the case of a batch reactor, the energy of initial composition of reagents allows for the propagation of pulses even for days [41], but a typical time of an experiment in such conditions is less than an hour. In a continuously fed reactor pulses can run as long as the reactants are delivered [44].

If the refractory period of the medium is sufficiently long then the region behind a pulse cannot be excited again for a long time. As a consequence, colliding pulses annihilate. This type of behavior is quite common in excitable systems. Another important feature is the dispersion relation for a train of excitations. Typically, the first pulse is the fastest, and the subsequent ones are slower, which is related to the fact that the medium behind the first pulse has not relaxed completely [63]. For example, this phenomenon is responsible for stabilization of positions in a train of pulses rotating on a ring-shaped excitable area. However, in some systems [43] the anomalous dispersion relation is observed and there are selected stable distances between subsequent spikes. The excitable systems characterized by the anomalous dispersion can play an important role in information processing, because packages of pulses are stable and thus the information coded in such packages can propagate without dispersion.

The mathematical description of excitable chemical media is based on differential equations of reaction-diffusion type, sometimes supplemented by additional equations that describe the evolution of the other important properties of the medium, for example the orientation of the surface in the case of CO oxidation on a Pt surface [12,37]. Numerical simulations of pulse propagation in an excitable medium which are presented in many papers [48,49] use the FitzHugh–Nagumo model describing the time evolution of electrical potentials in nerve channels [17,18,54]. The models for systems with the Belousov–Zhabotinsky (BZ) reaction, for example the Rovinsky and Zhabotinsky model [61,62] for the ferroin catalyzed BZ reaction with the immobilized catalyst, can be derived from "realistic" reaction schemes [16] via different techniques of variable reduction. Experiments with the Ru-catalyzed, photosensitive BZ reaction have become standard in experiments with structured excitable media because the level of excitation can be easily controlled by illumination; see for example [7,27,33]. Light catalyzes the production of bromine that inhibits the reaction, so non illuminated regions are excitable and those strongly illuminated are not. The pattern of excitable (dark) and non-excitable (transparent) fields is just projected on a membrane filled with the reagents. For example, the labyrinth shown in Fig. 1 has been obtained by illuminating a membrane through a proper mask. The presence of a membrane is important because it stops convection in the solution and reduces the speed and size of spikes, so studied systems can be smaller. The other methods of forming excitable channels based on immobilizing a catalyst by imprinting it on a membrane [68] or attaching it by lithography [21,71,72] have been also used, but they seem to be more difficult.

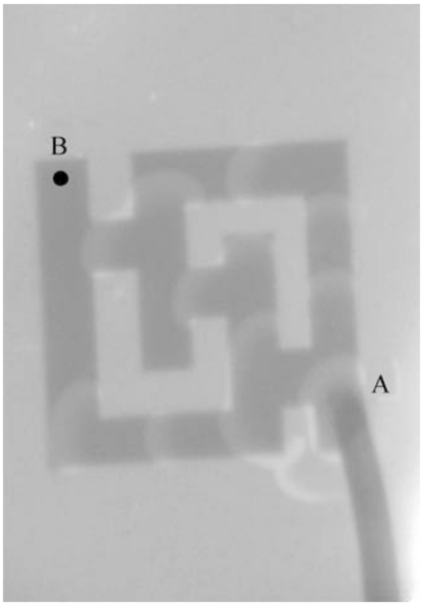
Numerical simulations of the Ru-catalyzed BZ reaction can be performed with different variants of the Oregonator model [9,19,20,38]. For example, the three-variable model uses the following equations:

$$\varepsilon_1 \frac{\partial u}{\partial t} = u(1-u) - w(u-q) + D_u \nabla^2 u \qquad (1)$$

$$\frac{\partial v}{\partial t} = u - v \qquad (2)$$

$$\varepsilon_2 \frac{\partial w}{\partial t} = \phi + f v - w(u+q) + D_w \nabla^2 w \qquad (3)$$

where u, v and w denote dimensionless concentrations of the following reagents: $HBrO_2$, Ru(4,4′-dm-bpy)$_3^{3+}$, and Br^-, respectively. In the considered system of equations, u is an activator and v is an inhibitor. The set of Oregonator equations given above reduces to the two-variable model cited in the article of Adamatzky ▶ Reaction-Diffusion Computing if the processes responsible for bromide production are very fast if compared to the other reactions ($\varepsilon_2 \ll \varepsilon_1 \ll 1$). In such cases, the local value of w can be calculated assuming that it corresponds to the stationary solution of the third equation. If such a w is substituted into the first equation, one obtains the two-variable Oregonator model. In the equations given above, the units of space and time are dimensionless and they have been chosen to scale the reaction rates. Here we also neglected the diffusion of ruthenium catalytic complex because usually it is much smaller than those of the other reagents. The reaction-diffusion equations describing the time evolution of the system can be solved with the standard numerical techniques [57].



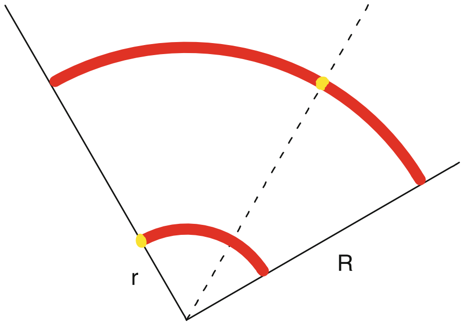
Computing in Geometrical Constrained Excitable Chemical Systems, Figure 1
Pulses of excitation propagating in a labyrinth observed in an experiment with a Ru-catalyzed BZ-reaction. The excitable areas are *dark*, the non-excitable *light*. The source of a train of pulses (a tip of a silver wire) is placed at the point A

The parameter ϕ represents the rate of bromide production caused by illumination and it is proportional to the applied light intensity. Therefore, by adjusting the local illumination (or choosing the proper ϕ as a function of space variables in simulations) we create regions with the required level of excitability, like for example excitable stripes insulated by a non-excitable neighborhood. Of course, the reagents can freely diffuse between the regions characterized by different illuminations.

The structure of the equations describing the system's evolution in time and space gives the name "reaction diffusion computing" [4] to information processing with an excitable medium. Simulations play an important role in tests of potential information processing devices, because, unlike in experiment, the conditions and parameters of studied systems can be easily adjusted with the required precision and kept forever. Most of the information processing devices discussed below were first tested in simulations and next verified experimentally. One of the problems is related to the short time of a typical experiment in a batch condition (a membrane filled with reagents) that does not exceed one hour. In such systems the period in which the conditions can be regarded as stable is usually shorter than 30 minutes and within this time an experimentalist should prepare the medium, introduce the required illumination and perform observations. The problem can be solved when one uses a continuously fed reactor [44], but experimental setups are more complex. On the other hand, simulations often indicate that the range of parameters in which a given phenomenon appears is very narrow. Fortunately experiments seem to be more robust than simulations and the expected effects can be observed despite of inevitable randomness in reagent preparation. Having in mind the relatively short time of experiments in batch conditions and the low stability of the medium, we believe that applications of a liquid chemical excitable medium like the Ru-catalyzed BZ reaction as information processors (wetware) are rather academic and they are mainly oriented on the verification of ideas. Practical applications of reaction-diffusion computers will probably be based on an other type of medium, like structured semiconductors (▶ Unconventional Computing, Novel Hardware for).

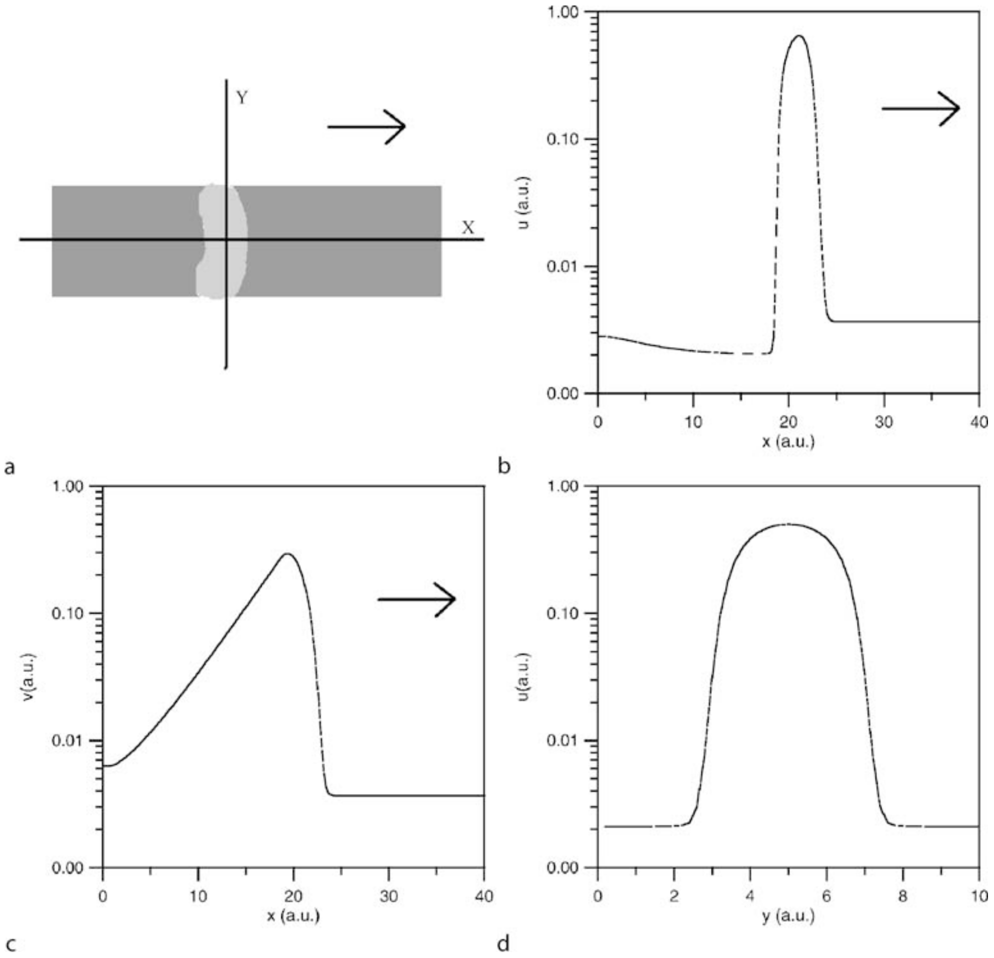
In an excitable reaction-diffusion medium, spikes propagate along the minimum time path. Historically, one of the first applications of a structured chemical medium used in information processing was the solution of the problem of finding the shortest path in a labyrinth [67]. The idea is illustrated in Fig. 1. The labyrinth is build of excitable channels (dark) separated by non-excitable medium (light) that does not allow for interactions between pulses propagating in different channels. Let us assume that we are interested in the distance between the right bottom corner (point A) and the left upper corner (point B) of the labyrinth shown. The algorithm that scans all the possible paths between these points and selects the shortest one is automatically executed if the paths in labyrinth are build of an excitable chemical medium. To see how it works, let us excite the medium at the point A. The excitation spreads out through the labyrinth, separates at the junctions and spikes enter all possible paths. During the time evolution, pulses of excitation can collide and annihilate, but the one that propagates along the shortest path has always unexcited medium in front. Knowing the time difference between the moment when the pulse is initiated at the point A and the moment when it arrives at the point B and assuming that the speed of a pulse is constant, we can estimate the length of shortest path linking both points. The algorithm described above is called the "prairie fire" algorithm and it is automatically executed by an excitable medium. It finds the shortest path in a highly parallel manner scanning all possible routes at the same time. It is quite remarkable that the time required for finding the shortest path does not depend on the the complexity of labyrinth structure, but only on the distance between the considered points.

Although the estimation of the minimum distance separating two points in a labyrinth is relatively easy (within the assumption that corners do not significantly change pulse speed) it is more difficult to tell what is the shortest path. To do it one can trace the pulse that arrived first and plot a line tangential to its velocity. This idea was discussed in [6] in the context of finding the length of the shortest path in a nonhomogeneous chemical medium with obstacles. The method, although relatively easy, requires the presence of an external observer who follows the propagation of pulses. An alternative technique of extracting the shortest path based on the image processing was demonstrated in [59]. One can also locate the shortest path connecting two points in a labyrinth in a purely chemical way using the coincidence of excitations generated at the endpoints of the path. Such a method, described in [29], allows one to find the midpoint of a trajectory, and thus locate the shortest path point by point.



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 2

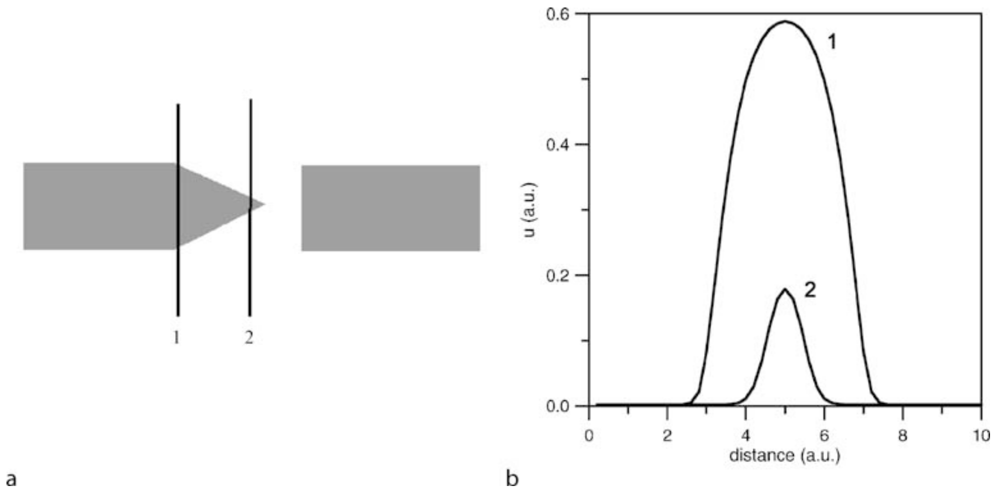
The idea of angle trisection. Angle arms are linked with *red* arc shaped excitable channels with the radius ratio $r : R = 1 : 3$. The channels have been excited at the same time at the right arm of the angle. At the moment when the excitation pulse (*a yellow dot*) on a smaller arch reaches the second arm the excitation on the other arch shows a point on a line trisecting the angle (*the dashed line*)



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 3
The shape of the excitation pulse in a stripe of excitable medium. The *arrow* indicates the direction of propagation. Calculations were done for the Oregonator model (Eqs. (1)–(3)) and the following values of parameters were applied: $f = 1.12, q = 0.002, \varepsilon_1 = 0.08, \varepsilon_2 = 0.00097, \phi_{\text{excitable}} = 0.007, \phi_{\text{non-excitable}} = 0.075$. a The position of a spike on the stripe, b, c concentration of activator and inhibitor along the *x*-axis, d concentration of activator along the *y*-axis

The approximately constant speed of excitation pulses in a reaction-diffusion medium allows one to solve some geometrically oriented problems. For example one can "measure" the number π by comparing the time of pulse propagation around a non-excitable circle of the radius d with the time of propagation around a square with the same side [69]. Similarly, a constant speed of propagating pulses can be used to obtain a given fraction of an angle. For example, a trisection of an angle can be done if the angle arms are linked with two arc-shaped excitable channels as shown in Fig. 2. The ratio of channel radii should be equal to 3. Both channels are excited at the same time at points on the same arm. The position of excitation at the larger arch at the moment when the excitation propagating on the shorter arch reaches the other arm belongs to a line that trisects the angle.

The examples given above are based on two properties of structured excitable medium: the fact that the non-excitability of the neighborhood can restrict the motion of an excitation pulse to a channel, and that the speed of propagation depends on the excitability level of the medium but not on channel shape. This is true when channels are wide and the curvature is small. There are also other properties of an excitable medium useful for information processing. A typical shape of a pulse propagating in a stripe of excitable medium is illustrated in Fig. 3. The pulse moves from the left to the right as the arrow indicates. The profiles on Fig. 3b,c show cross sections of the concentrations



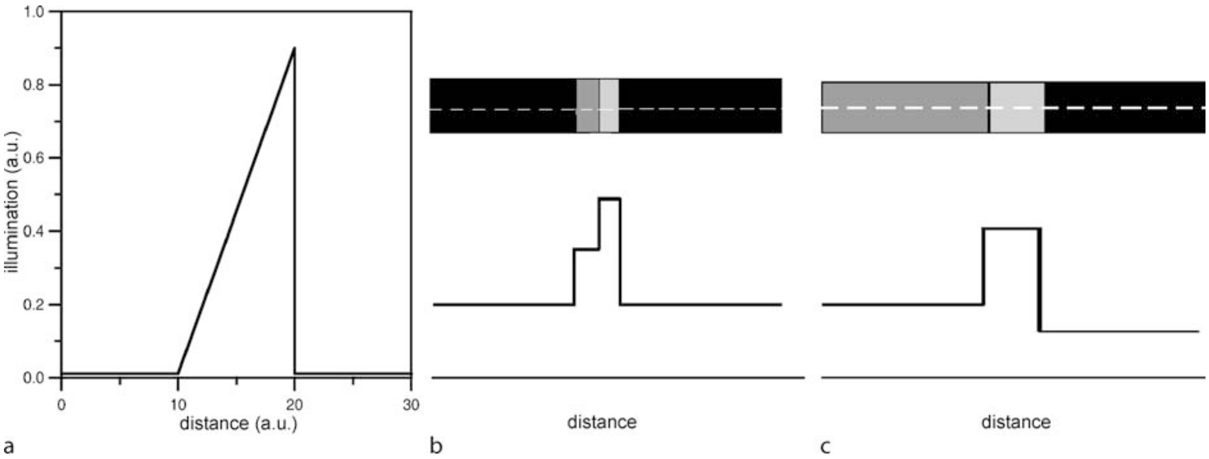
Computing in Geometrical Constrained Excitable Chemical Systems, Figure 4
The shape of activator concentration in an excitable medium of a triangular shape. a The structure of excitable (*dark*) and non-excitable (*light*) regions, b concentration of activator along the lines marked in a. The profiles correspond to times when the activator reaches its maximum at a given line

of the activator $u(x, y)$ and inhibitor $v(x, y)$ along the horizontal axis of the stripe at a selected moment of time. The peak of the inhibitor follows activator maximum and is responsible for the refractory character of the region behind the pulse. Figure 3d illustrates the profile of an activator along the line perpendicular to the stripe axis. The concentration of the activator reaches its maximum in the stripe center and rapidly decreases at the boundary between the excitable and non-excitable areas. Therefore, the width of the excitable channel can be used as a parameter that controls the maximum concentration of an activator in a propagating pulse.

Figure 4 shows profiles of a concentration in an excitable channel with a triangular shape. The two curves on Fig. 4b illustrate the profile of u along the lines 1 and 2, respectively measured at the time when the concentration of u on a given line reaches its maximum. It can be seen that the maximum concentration of the activator decreases when a pulse propagates towards the tip of the triangle. This effect can be used to build a chemical signal diode. Let us consider two pieces of excitable medium, one of triangular shape and another rectangular, separated by a non-excitable gap as shown on Fig. 4a. It is expected that a perturbation of the rectangular area by a pulse propagating towards the tip of the triangular one is much smaller than the perturbation of the triangular area by a pulse propagating towards the end of rectangular channel. Using this effect, a chemical signal diode that transmits pulses only in one direction can be constructed just by selecting the right width of non-excitable medium separating two pieces of excitable medium: a triangular one and a rectangular one [5,40].

The idea of a chemical signal diode presented in Fig. 4a was, in some sense, generalized by Davydov et al. [46], who considered pulses of excitation propagating on a 2-dimensional surface in 3-dimensional space. It has been shown that the propagation of spikes on surfaces with rapidly changing curvature can be unidirectional. For example, such an effect occurs when an excitation propagates on the surface of a tube with a variable diameter. In such a case, a spike moving from a segment characterized by a small diameter towards a larger one is stopped.

For both of the chemical signal diodes mentioned above, the excitability of the medium is a non trivial function of some space variables. However, a signal diode can be also constructed when the excitability of the medium changes in one direction only, so in a properly selected coordinate system it is a function of a single space variable. For example, the diode behavior in a system where the excitability level is a triangular function of a single space variable has been confirmed in numerical simulations based on the Oregonator model of a photosensitive, Ru-catalyzed BZ reaction (Eqs. (1)–(3)) [76]. If the catalyst is immobilized, then pulses that enter the region of inhomogeneous illumination from the strongly illuminated site are not transmitted, whereas the pulses propagating in the other direction can pass through (see Fig. 5a). A similar diode-like behavior resulting from a triangular profile of medium excitability can be expected for oxidation of CO on a Pt surface. Calculations demonstrate that a triangular

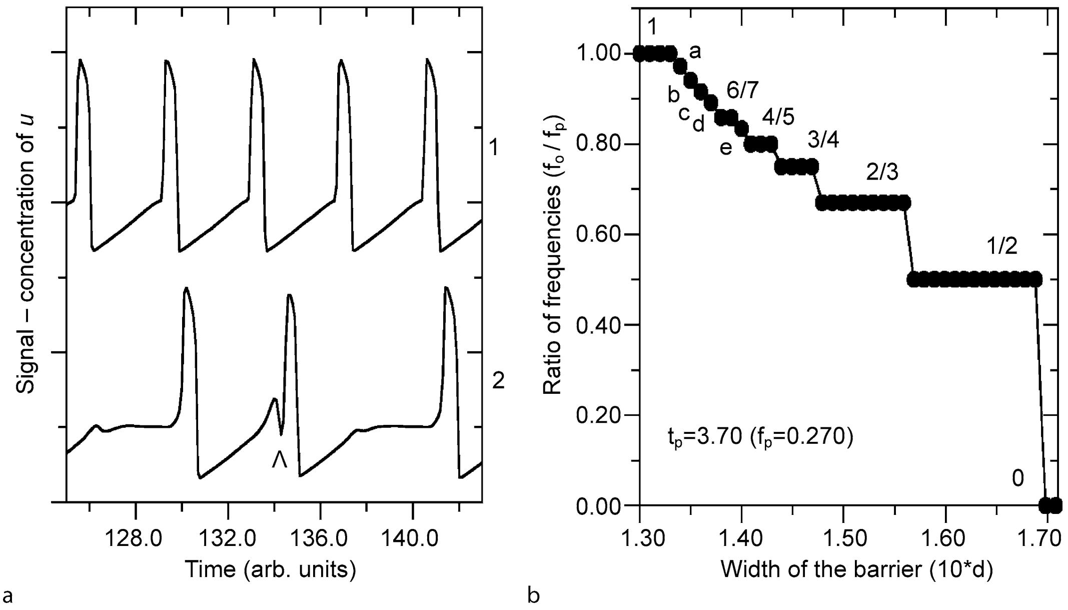


Computing in Geometrical Constrained Excitable Chemical Systems, Figure 5
The excitability as a function of a space variable in different one-dimensional realizations of a signal diode with BZ- reaction inhibited by light: a a triangular profile of illumination, b the illumination profile in a diode composed of two non-excitable barriers, c the illumination profile in a single barrier diode with nonsymmetrical excitable inputs. The *upper graphs* in b and c illustrate illumination on a membrane used in experiments [25]

profile of temperature (in this case, reduced with respect to that which characterizes the excitable medium) allows for the unidirectional transmission of spikes characterized by a high surface oxygen concentration [23]. However, a realization of the chemical signal diode can be simplified. If the properties of excitable channels on both sites of a diode are the same, then the diode can be constructed with just two stripes of non-excitable medium characterized by different excitabilities as illustrated in Fig. 5b. If the symmetry is broken at the level of input channels then the construction of a signal diode can be yet simpler, and it reduces to a single narrow non-excitable gap with a much lower excitability than that of the neighboring channels (cf. Fig. 5c). In both cases the numerical simulations based on the Oregonator model have shown that a diode works. The predictions of simulations have been qualitatively confirmed by experimental results [25]. Different realizations of a signal diode show that even for very simple signal processing devices the corresponding geometrical structure of excitable and non-excitable regions is not unique. Alternative constructions of chemical information processing devices seem important because they tell us on the minimum conditions necessary to build a device that performs a given function. In this respect a diode built with a single non-excitable region looks interesting, because such situation may occur at a cellular level, where the conditions inside the cell are different from those around. The diode behavior in the geometry shown on Fig. 5c indicates that the unidirectional propagation of spikes can be forced by a channel in a membrane, transparent to molecules or ions responsible for signal transmission.

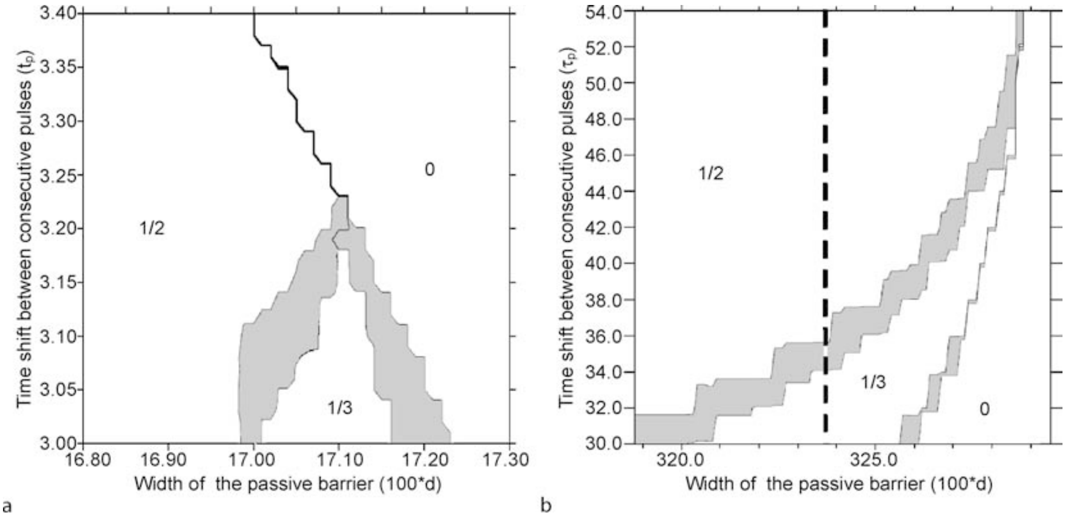
Wave-number-dependent transmission through a non-excitable barrier is another feature of a chemical excitable medium important for information processing [48]. Let us consider two excitable areas separated by a stripe of non-excitable medium and a pulse of excitation propagating in one of those areas. The perturbation of the area behind the stripe introduced by an arriving pulse depends on the direction of its propagation. If the pulse wavevector is parallel to the stripe then the perturbation is smaller than in the case when it arrives perpendicularly [48]. Therefore, the width of the stripe can be selected such that pulses propagating perpendicularly to the stripe can cross it, whereas a pulse propagating along the stripe do not excite the area on the other side. This feature is frequently used to arrange the geometry of excitable channels such that pulses arriving from one channel do not excite the other.

Non-excitable barriers in a structured medium can play a more complex role than that described above. The problem of barrier crossing by a periodic train of pulses can be seen as an excitation via a periodic perturbation of the medium. It has been studied in detail in [13]. The answer of the medium is quite characteristic in the form of a devil-staircase-like firing number as a function of perturbation strength. In the case of barrier crossing, the strength of the excitation behind a barrier generated by an arriving pulse depends on the character of the non-excitable medium, the barrier width, and the frequency of the incoming signal (usually, due to an uncompleted relaxation of the medium, the amplitude of spikes decreases with frequency). A typical complex frequency transformation



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 6

Frequency transformation on a barrier – the results for the FitzHugh–Nagumo model [64]. a The comparison of an arriving train of pulses (1) and the transmitted signal (2). b A typical dependence of the firing number as a function of barrier width. The plateaus are labeled with corresponding values of firing number (1, 6/7, 4/5, 3/4, 2/3, 1/2 and 0). At points labeled a–e the following values have been observed: a – 35/36; b – 14/15; c – 10/11; d – 8/9 and e – 5/6



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 7

The firing number as a function of the barrier width *d* and the interval of time between consecutive pulses (t_p). Labels in the *white* areas give the firing number, and the *gray* color marks values of parameters where more complicated transformations of frequency occur. a Results calculated for the FitzHugh–Nagumo model, b the Rovinsky–Zhabotinsky model. Both space and time are in dimensionless units

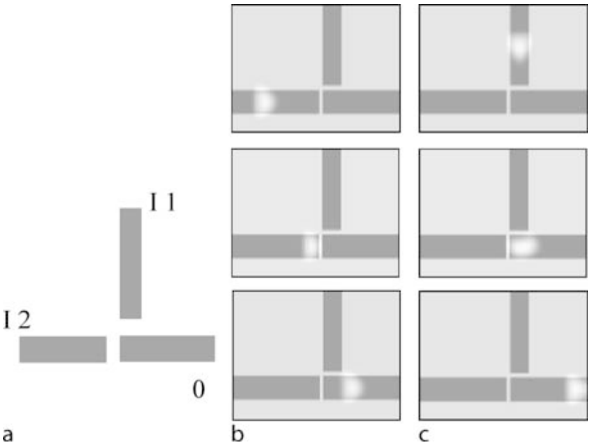
after barrier crossing is illustrated in Fig. 6. Experimental and numerical studies on firing number of a transmitted signal have been published [10,64,72,74]. It is interesting that the shape of regions characterized by the same firing number in the space of two parameters, barrier width and signal frequency, is not generic and depends on the type of the excitable medium. Figure 7 compares the firing numbers obtained for FitzHugh–Nagumo and Rovinsky–

Zhabotinsky models. In the first case, trains of pulses with small periods can cross wider barriers than trains characterized by low frequency; for the second model the dependence is reversed.

Logic Gates, Coincidence Detectors and Signal Filters

The simplest application of excitable media in information processing is based on the assumption that the logical FALSE and TRUE variables are represented by the rest state and by the presence of an excitation pulse at a given point of the system respectively. Within such interpretation a pulse represents a bit of information propagating in space. When the system remains in its rest state, no information is recorded or processed, which looks plausible for biological interpretation. Information coded in excitation pulses is processed in regions of space where pulses interact (via collision and subsequent annihilation or via transient local change in the properties of the medium). In this section we demonstrate that the geometry of excitable channels and non-excitable gaps can be tortured (the authors are grateful to prof. S. Stepney for this expression) to the level at which the system starts to perform the simplest logic operations on pulses. The binary chemical logic gates can be used as building blocks for devices performing more complex signal processing operations. Information processing with structured excitable media is "unconventional" because it is performed without a clock that sequences the operations, as it is in the standard von Neumann type computer architecture. On the other hand, in the signal processing devices described below, the proper timing of signals is important and this is achieved by selecting the right length and geometry of channels. In some cases, for example for the operations on trains of pulses, the presence of a reference signal, which plays a role similar to a clock, would significantly help to process information [70]. Historically, the logical gates were the first devices that have been realized with a structured chemical medium [1,2,14,65,73,75].

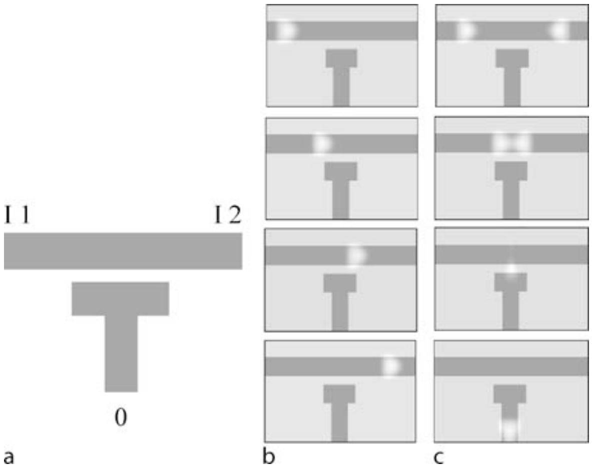
The setup of channels that execute the logic sum (OR) operation is illustrated in Fig. 8 [48]. The gate is composed of three excitable stripes (marked gray) surrounded by a non-excitable medium. The gaps separating the stripes have been selected such that a pulse that arrives perpendicularly to the gap can excite the area on the other side and a pulse that propagates parallel to the gap does not generate sufficient perturbation to excite the medium behind the gap. If there is no input pulse, the OR gate remains in the rest state and does not produce any output. A pulse in any of the input channels I1 and I2 can cross the gap separating these channels from the output O and an output spike appears. The excitation of the output channel generated by a pulse from one of the input channels propagates parallel to the gap separating the other input channel so it does not interfere with the other input as seen in Fig. 8b and c. The frequency at which the described OR gate operates is limited by the refractory period of the output medium. If the signals from both input channels arrive, but the time difference between pulses is smaller than the refractory time, then only the first pulse will produce the output spike.



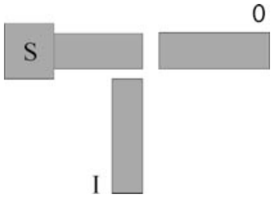
Computing in Geometrical Constrained Excitable Chemical Systems, Figure 8
The distribution of excitable channels (*dark*) that form the OR gate. b and c illustrate the time evolution of a single pulse arriving from inputs I1 and I2, respectively

The gate that returns the logic product (AND) of input signals is illustrated in Fig. 9. The width of the gap separating the output channel O from the input one (I1,I2) is selected such that a single excitation propagating in the input channel does not excite the output (Fig. 9b). However, if two counterpropagating pulses meet, then the resulting perturbation is sufficiently strong to generate an excitation in the output channel (Fig. 9c). Therefore, the output signal appears only when both sites of the input channel have been excited and pulses of excitation collided in front of the output channel. The width of the output channel defines the time difference between input pulses treated as simultaneous. Therefore, the structure shown in Fig. 9 can also be used as a detector of time coincidence between pulses.

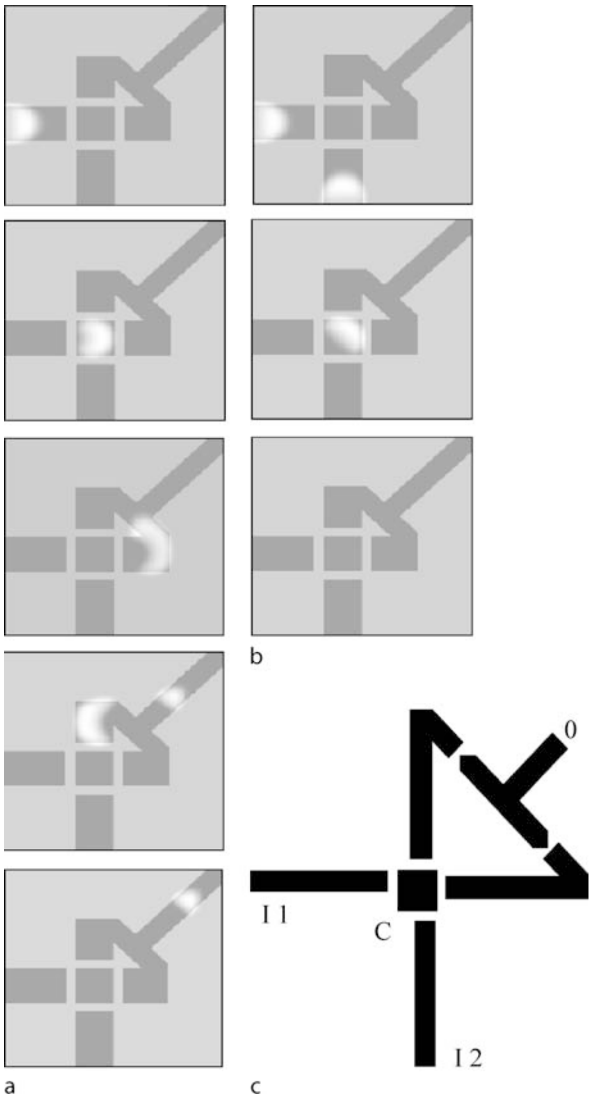
The design of the negation gate is shown in Fig. 10. The gaps separating the input channels, the source channel and the output channel can be crossed by a pulse that propagates perpendicularly, but they are non-penetrable for pulses propagating parallel to the gaps. The NOT gate



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 9
The distribution of excitable channels (*dark*) that form the AND gate. b and c illustrate the response of the gate to a single pulse arriving from input I1 and to a pair of pulses, respectively



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 10
The distribution of excitable channels (*dark*) that form the NOT gate



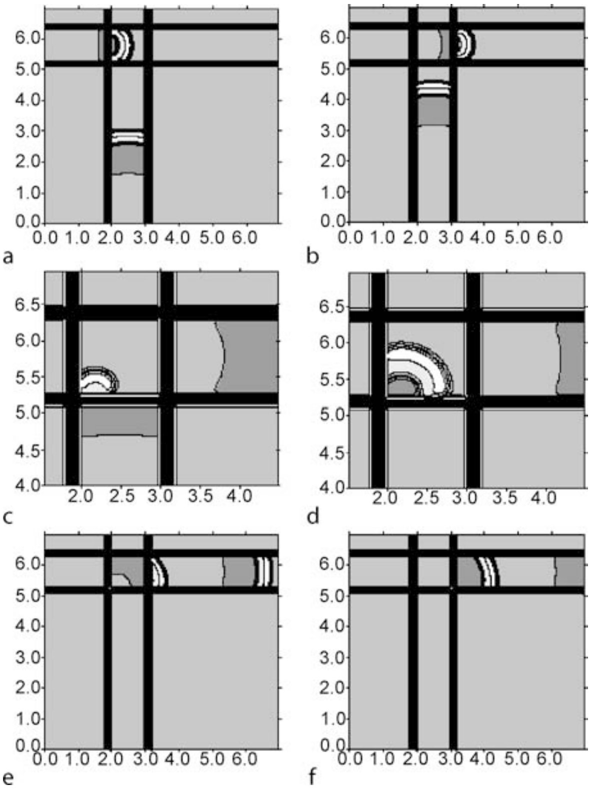
Computing in Geometrical Constrained Excitable Chemical Systems, Figure 11
c the distribution of excitable channels (*dark*) that form the XOR gate. a and b illustrate the response of the gate to a single pulse arriving from input I1 and to a pair of pulses, respectively

should deliver an output signal if the input is in the rest state. Therefore, it should contain a source of excitation pulses (marked S). If the input is in the rest state then pulses from the source propagate unperturbed and enter the output channel. If there is an excitation pulse in the input channel then it enters the channel linking the source with the output and annihilates with one of pulses generated by the source. As a result no output pulse appears. At the first look, the described NOT gate works fine, but if we assume that a single pulse is used in information coding than the input pulse should arrive at the right time to block the source. Therefore, additional synchronization of the source is required. If information is coded in trains of pulses then the frequency of the source should match with the one used for coding.

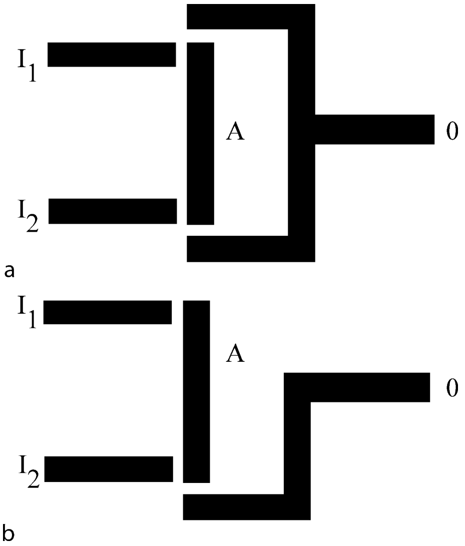
The structure of excitable channels for the exclusive OR (XOR) gate is illustrated in Fig. 11c [34]. Two input channels bring signals to the central area C. The output channels are linked together via diodes (Fig. 4) that stop possible backward propagation. As in the previous cases only pulses perpendicular to the gaps can pass through non-excitable gaps between the central area and both input and output channels. The shape of the central area has been designed such that an excitation generated by a single input pulse propagates parallel to one of the outputs and is perpendicular to another. As the result one of the output channels is excited (see Fig. 11a). However, if pulses from both input channels arrive at the same time then the wavevector of the excitation in the central part is always parallel to the boundaries (Fig. 11b). Therefore, no output signal appears. Of course, there is no output signal if none of the inputs are excited. It is worth noticing that for some geometries of the XOR gate the diodes in output channels are not necessary because the backward propagation does not produce a pulse with a wavevector perpendicular to a gap as seen in the fourth frame of Fig. 11a.

Another interesting example of behavior resulting from the interaction of pulses has been observed in a cross-shaped structure built of excitable regions, separated by gaps penetrable for perpendicular pulses [65,66] shown in Fig. 12. The answer of cross-shaped junction to a pair pulses arriving from two perpendicular directions has been studied as a function of the time difference between pulses. Of course, if the time difference is large, pulses propagate independently along their channels. If the time difference is small the cross-junction acts like the AND gate and the output excitation appears in one of the corner areas. However, for a certain time difference the first arriving pulse is able to redirect the other and force it to follow. The effect is related to uncompleted relaxation of the central area of the junction at the moment when the second pulse arrives. Pulse redirection seems to be an interesting effect from the point of programming with excitation pulses, but in practice it requires a high precision in selecting the right time difference.

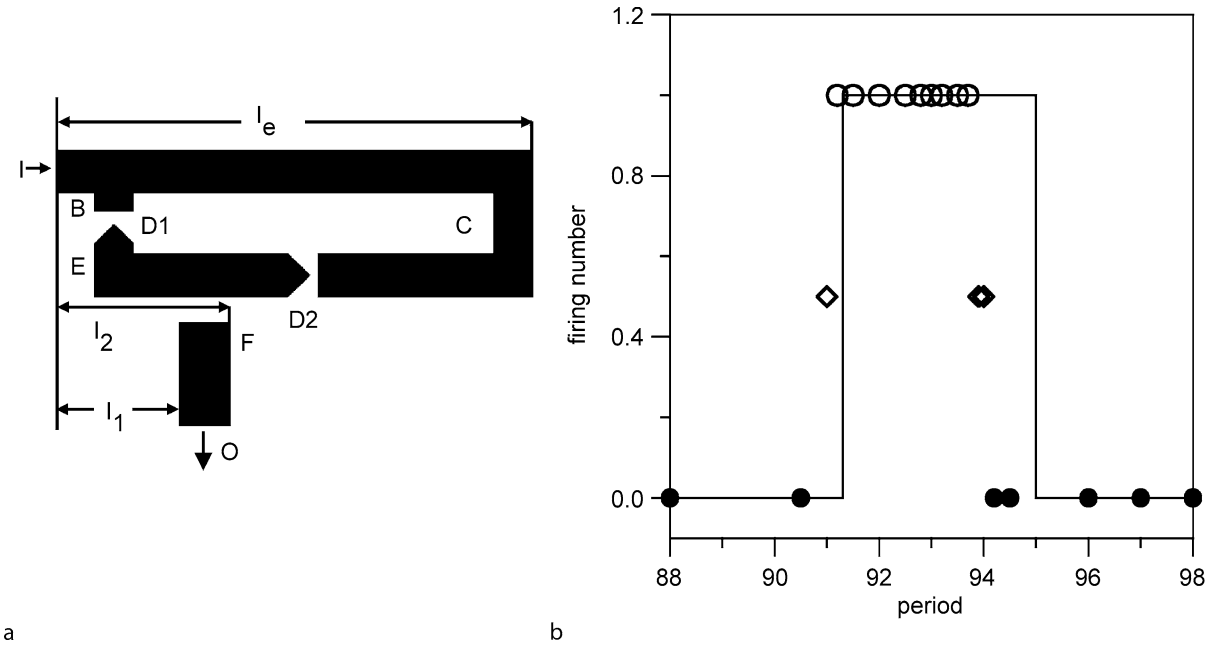
The logic gates described above can also be applied to transform signals composed of many spikes. For example, two trains of pulses can be added together if they pass through an OR gate. The AND gate creates a signal composed of coinciding pulses from both trains. It is also easy to use a structured excitable medium and generate a signal that does not contain coinciding pulses [50]. The structure of the corresponding device is illustrated in Fig. 13. All non-excitable gaps are penetrable for perpendicular pulses. If a pulse of excitation arrives from one of the input channels then it excites the segment A and propagates towards the other end of it. If there is no spike in the other input channel then this excitation propagates unperturbed and activates the output channel. However, if a pulse from the other channel arrives it annihilates with the original



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 12
The distribution of excitable and non-excitable regions in a cross-shaped junction. Here the excitable regions are *gray* and the non-excitable *black*. Consecutive figures illustrate an interesting type of time evolution caused by interaction of pulses. Two central figures are enlarged in order to show how uncompleted relaxation influences the shape of the next pulse



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 13
The distribution of excitable channels in devices that compare trains of pulses. The device shown in a applies the XOR operation to a pair of signals and makes a signal composed of spikes that do not coincide with the other signal. b generates a signal composed of spikes that arrive through I1 and do not coincide with excitation pulses coming from I2



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 14

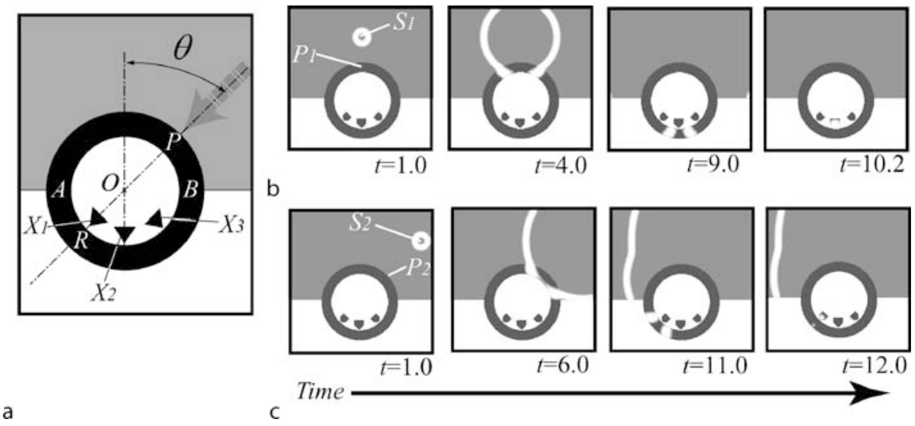
a The distribution of excitable and non-excitable regions in a band filter. b Typical characteristics of the band filter [22]. The *black dots* mark periods for which the signal is not transmitted, the *empty* ones indicate full transmission, the *empty diamonds* mark the periods of the arriving signal for which every second spike is transmitted

pulse and no output signal is generated. The structure illustrated on Fig. 13a can be easily transformed into a device that compares two trains of pulses such that the resulting signal is composed of spikes of the first signal that do not coincide with the pulses of the second train. Such a device can be constructed just by neglecting one of the output channels. Figure 13b illustrates the geometry of the device that produces the output signal composed of pulses in arriving at the input I1 that have no corresponding spikes in the signal arriving from I2.

The coincidence detector can be used as a frequency filter that transmits periodic signals within a certain frequency interval [22,50]. The idea of such a filter is shown in Fig. 14. The device tests if the time between subsequent spikes of the train remains in the assumed range. As illustrated the signal arriving from the input I separates and enters the segment E through diodes D1 and D2. The segments E and F form a coincidence detector (or an AND gate, c. f. Fig. 9). The excitation of the output appears when an excitation coming to E via the segment C and D2 is in the coincidence with the subsequent spike of the train that arrived directly via D1. The time shift for which coincidences are tested is decided by the difference in lengths of both paths. The time resolution depends on the width of the F channel (here $l_2 - l_1$). For periodic signals the presented structure works as a frequency filter and transmits signals within the frequency range $f_{\pm} = v/(\Delta r \pm (\Delta w)/2)$, where Δr is the difference of distances traveled by spikes calculated to the point in E placed above the center of F channels ($\sim 2 * l_e$), Δw is the width of the output channel, and v is the velocity of a spike in the medium. Typical characteristics of the filter are illustrated on Fig. 14b. The points represent results of numerical simulations and the line shows the filter characteristics calculated from the equation given above. It can be noticed that at the ends of the transmitted frequency band a change in output signal frequency is observed. This unwelcome effect can be easily avoided if a sequence of two identical filters is used. A filter tuned to a certain frequency will also pass any of its harmonics because the output signal can be generated by the coincidence of every second, third, etc. pulses in the train.

Chemical Sensors Built with Structured Excitable Media

Even the simplest organisms without specialized nerve systems or brains are able to search for the optimum living conditions and resources of food. We should not be astonished by this fact because even chemical systems can receive and process information arriving from their neighborhood. In this section we describe how simple struc-

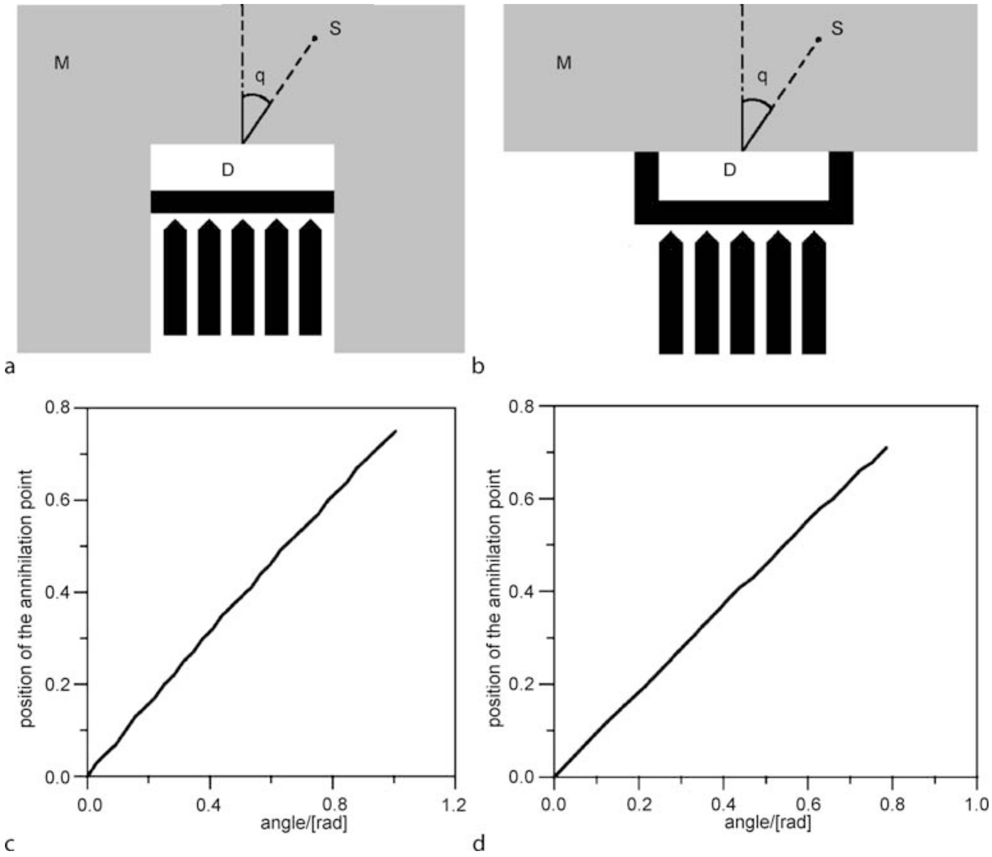


Computing in Geometrical Constrained Excitable Chemical Systems, Figure 15
The distribution of excitable areas (*gray* and *black*) and non-excitable regions (*white*) in the ring-shaped direction detector. *Triangles* marked X_1, X_2 and X_3 show the output channels. b and c illustrate time evolution of pulses generated from sources at different locations. The excited output channel depends on the direction of the source (copied from [53] with the permission of the authors)

tures built of excitable media can be used for direction and distance sensing. In order to simplify the conversion between an external stimulus and information processed by a sensor, we assume that the environment is represented by the same excitable medium as the sensor itself, so a stimulus can directly enter the sensor and be processed.

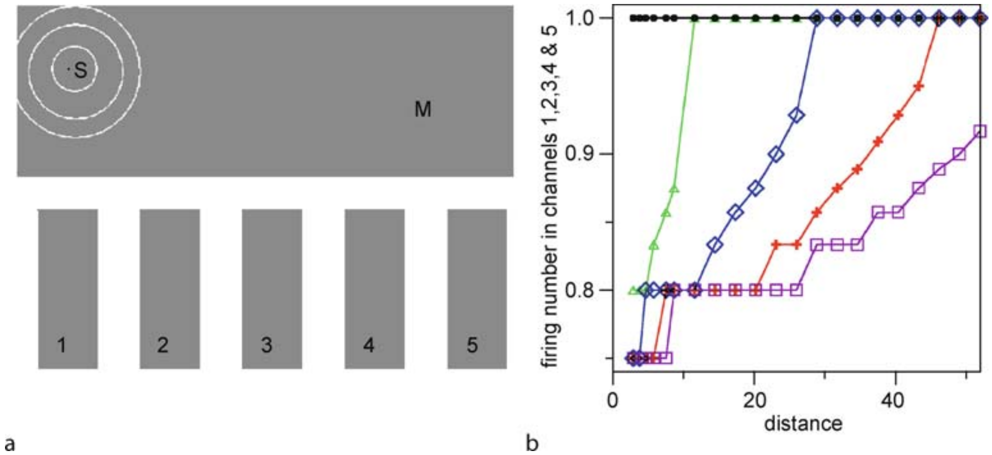
One possible strategy of sensing is based on a one-to-one relationship between the measured variable and the activated sensor channel [53]. An example of a sensor of that type is illustrated on Fig. 15. It is constructed with a highly excitable, black ring surrounded by a number of coincidence detectors, denoted as X_1, X_2 and X_3. The excitation of a detector appears if a pair of pulses collide on the ring in front of it. Let us consider a homogeneous excitable environment and a spherical pulse of excitation with the source at the point S_1. High excitability of the ring means that excitations on the ring propagate faster than those in the surrounding environment. At a certain moment the pulse arrives at the point P_1, which lies on a line connecting the S_1 and the center of the ring (see Fig. 15b). The arriving pulse creates an excitation on the ring originating from the point P_1. This excitation splits into a pair of pulses rotating in opposite directions and after propagating around the ring, they collide at the point, which is symmetric to P with respect to the center of the ring O. The point of collision can be located by an array of coincidence detectors and thus we have information on the wave vector of the arriving pulse. In this method the resolution depends on the number of detectors used because each of them corresponds to a certain range of the measured wave vectors. The fact that a pulse has appeared in a given output channel implies that no other channel of the sensor gets excited. It is interesting that the output information is reversed in space if compared with the input one; the left sensor channels are excited when an excitation arrives from the right and vice versa.

The geometry of the direction sensor that sends information as an excitation pulse in one of its detector channels can be yet simplified. Two such realizations are illustrated in Fig. 16. The excitable areas that form the sensor are marked black, the excitable surrounding medium where stimuli propagate is gray, and the non-excitable areas are white. Let us consider an excitation generated in the medium M by a source S. It is obvious that if the source is located just above the sensor then pulses of excitation are generated at the ends of the sensor channel D at the same time and they finally annihilate above the central coincidence channel and generate an output pulse in it. If the source is located at a certain angle with respect to the vertical line then the annihilation point is shifted off the center of D. We have performed a series of numerical simulations to find the relation between the position of the source and the position of annihilation point in D. In Fig. 16c and 16d the position of the annihilation point is plotted as a function of the angle between the source position and the vertical line. Although the constructions of sensors seem to be very similar, the working range of angles is larger for the sensor shown on Fig. 16b whereas the sensor shown in Fig. 16a offers slightly better resolution. Using information from two detectors of direction, the position of the source of excitation can be easily located because it is placed on the intersection of the lines representing the detected directions [77].



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 16

The geometry of excitable (*black*) and diffusive (*white*) areas in two realizations of simplified distance sensors. The *gray* color marks the excitable medium around the sensor, *S* marks the position of excitation source. c and d show the position of the annihilation point in the channel D as a function of the angle *q* between the source position and the vertical line. The half-length of D is used as the scale unit



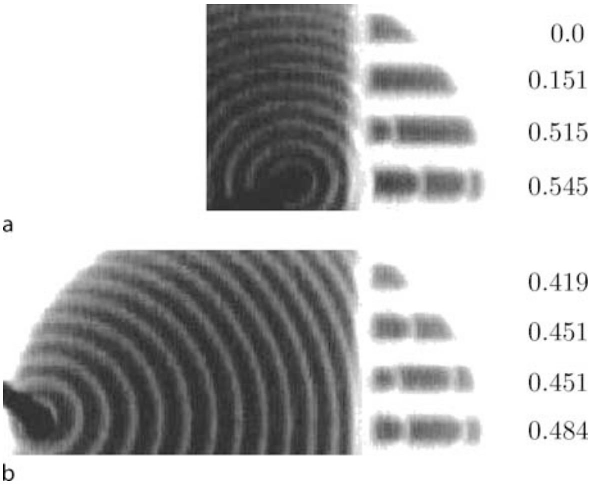
Computing in Geometrical Constrained Excitable Chemical Systems, Figure 17

The distance sensor based on frequency transformation. a The distribution of excitable channels, b the firing numbers in different channels as a function of distance; *black, green, blue red and violet curves* correspond to signals in channels 1–5, respectively

Another strategy of sensing is based on the observation that frequencies of pulses excited in a set of sensor channels by an external periodic perturbation contain information on the location of a stimulus. Within this strategy there is no direct relationship between the number of sensor channels and sensor resolution, and, as we show below, a sensor with a relatively small number of channels can quite precisely estimate the distance separating it from the excitation source. As we have mentioned in the Introduction, the frequency of a chemical signal can change after propagating through a barrier made of non-excitable medium. For a given frequency of arriving spikes the frequency of excitations behind the barrier depends on the barrier width and on the angle between the normal angle to the barrier and the wave vector of arriving pulses. This effect can be used for sensing. The geometrical arrangement of excitable and non-excitable areas in a distance sensor is shown in Fig. 17. The excitable signal channels (in Fig. 17a they are numbered 1–5) are wide enough to ensure stable propagation of spikes. They are separated from one another by parallel non-excitable gaps that do not allow for interference between pulses propagating in the neighboring channels. The sensor channels are separated from the excitable medium M by the non-excitable sensor gap G. The width of this gap is very important. If the gap is too wide then no excitation of the medium M can generate a pulse in the sensor channels. If the gap is narrow then any excitation in front of the sensor can pass G and create a spike in each sensor channel so the signals sent out by the sensor channels are identical. However, there is a range of gap widths such that the firing number depends on the wave vector characterizing a pulse at the gap in front of the channel. If the source S is close to the array of sensor channels, then the wave vectors characterizing excitations in front of various channels are significantly different. Thus the frequencies of excitations in various channels should differ too. On the other hand, if the source of excitations is far away from the gap G then the wave vectors in front of different channels should be almost identical and the frequencies of excitations should be the same. Therefore, the system illustrated in Fig. 17a can sense the distance separating it from the source of excitations. If this distance is small then the firing numbers in neighboring sensor channels are different and these differences decrease when the source of excitations moves away. A typical distance dependence of firing numbers observed in different channels is illustrated in Fig. 17b. This result has been obtained in numerical simulations based on the Oregonator model.

The range of sensed distances depends on the number of sensor channels. A similar sensor, but with 4 sensor channels, was studied in [28]. Comparing the results, we observe that the presence of 5th channel significantly improves the range of distances for which the sensor operates. On the other hand, the additional channel has almost no effect on sensor resolution at short distances. The sensor accuracy seems to be a complex function related to the width of the sensor gap and the properties of channels. The firing number of a single channel as a function of the distance between the sensor and the source has a devil-staircase-like form with long intervals where the function is constant corresponding to simple fractions as illustrated in Fig. 5b. In some range of distances the steps in firing numbers of different channels can coincide, so the resolution in this range of distances is poor. For the other regions, the firing numbers change rapidly and even small changes in distance can be easily detected.

The signal transformation on a barrier depends on the frequency of incoming pulses (c. f. Fig. 6) so the distance sensor that works well for one frequency may not work for another. In order to function properly for different stimuli the sensor has to be adapted to the conditions it operates. In practice such adaptation can be realized by the comparison of frequencies of signals in detector channels with the frequency in a control channel. For example the control channel can be separated from the medium M by a barrier so narrow that every excitation of the medium generates a spike. The comparison between the frequency of exci-



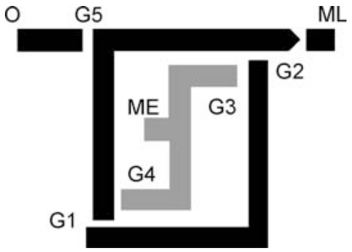
Computing in Geometrical Constrained Excitable Chemical Systems, Figure 18
Two snapshots from the experimental realization of the distance sensor with four channels. In the *upper* figure the source (1 mm thick silver wire) is placed 2 mm away from the sensor; in the *bottom* one the source is 12 mm away. The firing numbers are given next to the corresponding channels

tations in the control channel and in the sensor channels can be used to adjust the sensor gap G. If the frequency of excitations in the sensor channels is the same as in the control channels then the width of the gap should be increased or its excitability level decreased. On the other hand if the frequency in the sensor channel is much smaller than in the control channel (or null) then the gap should be more narrow or more excitable. Such an adaptation mechanism allows one to adjust the distance detector to any frequency of arriving excitations.

The fact that the distance detector described above actually works has been confirmed in experiments with a photosensitive Ru-catalyzed BZ reaction. Typical snapshots from two experiments performed for the source placed 2 and 12 mm away from the sensors are shown in Fig. 18. The firing numbers observed in different sensor channels confirm qualitatively the predictions of numerical simulations. If the source of excitations is close to the sensor gap, then the differences between firing numbers observed in neighboring channels are large. On the other hand, when the source of excitations is far away from the sensor, the frequencies in different channels become similar. The range of distances at which the sensor works is measured in centimeters so it is of the same order as the sensor size.

The Ring Memory and Its Applications

The devices discussed in the previous section can be classified as instant machines [58] capable of performing just the task they have been designed for. A memory where information coded in excitation pulses can be written-in, kept, read-out and, if necessary, erased, significantly increases the information processing potential of structured excitable media. Moreover, due to the fact that the state of memory can be changed by a spike, the memory allows for programming with excitation pulses. One possible realization of a chemical memory is based on the observation that a pulse of excitation can rotate on a ring-shaped excitable area as long as the reactants are supplied and the products removed [41,52,55]. Therefore, a ring with a number of spikes rotating on it can be regarded as a loaded memory cell. Such memory can be erased by counterpropagating pulses. The idea of memory with loading pulses rotating in one direction and erasing pulses in another has been discussed in [49]. If the ring is big then it can be used to memorize a large amount of information because it has many states corresponding to different numbers of rotating pulses. However, in such cases, loading the memory with subsequent pulses may not be reliable because the input can be blocked by the refractory tail left by one of already rotating pulses. The same effect can block the erasing pulses. Therefore, the memory capable of storing just a single bit seems to be more reliable and we consider it in this section. Such memory has two states: if there is a rotating pulse the ring is in the logical TRUE state (we call it loaded); if there is no pulse the state of memory corresponds to the logical FALSE and we call such memory erased.

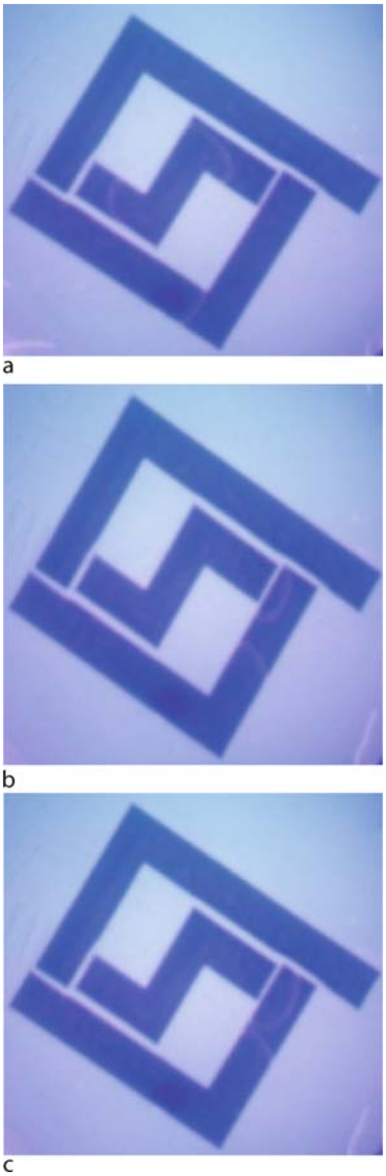


Computing in Geometrical Constrained Excitable Chemical Systems, Figure 19
The distribution of excitable channels (*dark*) that form a memory cell. The cell is composed of the loading channel ML, the memory ring, the erasing channel ME (marked *gray*) and the output channel O

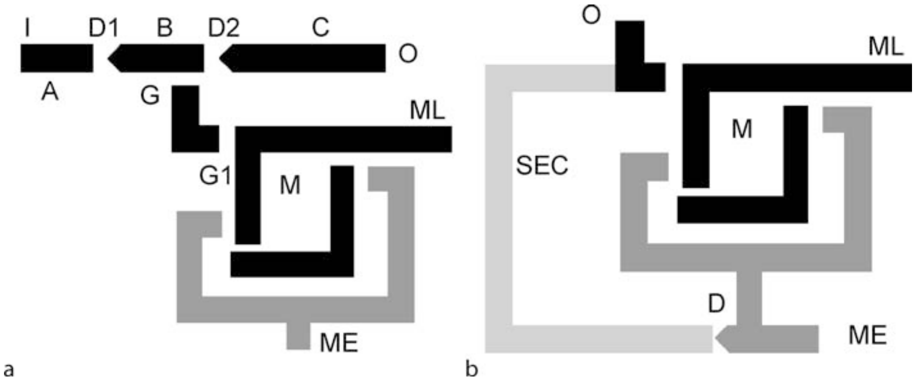
Let us consider the memory illustrated in Fig. 19. The black areas define the memory ring, the output channel O, and the loading channel ML. The memory ring is formed by two L-shaped excitable areas. The areas are separated by gaps and, as we show below, with a special choice of the gaps the symmetry of the ring is broken and unidirectional rotation ensured. The Z-shaped excitable area composed of gray segments inside the ring forms the erasing channel. The widths of all non-excitable gaps separating excitable areas are selected such that a pulse of excitation propagating perpendicularly to the gap excites the active area on the other site of the gap, but the gap is impenetrable for pulses propagating parallel to the gap. The memory cell can be loaded by a spike arriving from the ML channel. Such a spike crosses the gap and generates an excitation on the ring rotating counterclockwise. The information about loaded memory is periodically sent out as a series of spikes through the output channel O. The rotating pulse does not affect the erasing channel because it always propagates parallel to it. The erasing excitation is generated in the center of the Z-shaped area and it splits into two erasing pulses. These pulses can cross the gaps separating the erasing channel from the memory ring and create a pair of pulses rotating clockwise. A spike that propagates clockwise on the memory ring is not stable because it is not able to cross any of the gaps and dies. It also does not produce any output signal. Therefore, if the memory has not been loaded then an erasing excitation does not load it. On the other hand, if the memory is loaded then clockwise rotating pulses resulting from the excitation of the erasing channel annihilate with the loading pulse and the memory is erased. The idea of using two places where erasing pulses can enter the memory ring is used to ensure that at least one of those places is fully relaxed and so one of erasing pulses can always enter the ring.

In order to verify that such memory works, we have performed a number of simulations using the Rovinsky–Zhabotinsky model of a BZ reaction and have done the experiments. We considered a loaded memory and at a random time the excitation was generated in the middle of the erasing channel. In all cases, such an excitation erased the memory ring. A typical experimental result is illustrated in Fig. 20. Here the channels are 1.5 mm thick and the gaps are 0.1 mm wide. Figure 20a shows the loaded memory and initiated pair of pulses in the erasing channel. In Fig. 20b one of the pulses from the erasing channel enters the memory ring. The memory ring in front of the left part of erasing channel is still in the refractory state so the left erasing pulse has not produced an excitation on the ring. Finally in Fig. 20c we observe the annihilation of the loading pulse with one of erasing pulses. Therefore, the state of memory changed from loaded to unloaded. The experiment was repeated a few times and the results were in a qualitative agreement with the simulations: the loaded memory cell kept its information for a few minutes and it was erased after every excitation of the erasing channel.

Figure 21 illustrates two simple, yet interesting, applications of a memory cell. Figure 21a shows a switchable unidirectional channel that can be opened or closed depending on the state of memory [29]. The channel is constructed with three excitable segments A, B and C separated by signals diodes D1 and D2 (c. f. Fig. 4). The mid segment B is also linked with the output of the memory ring M. Here the erasing channels of M are placed outside the memory ring, but their function is exactly the same as in the memory illustrated in Fig. 19. The idea of switchable channel is similar to the construction of the NOT gate (Fig. 10). If the memory is not loaded then the spike propagation from input I to output O is unperturbed. However, if the memory is loaded then pulses of excitation periodically enter the segment B and annihilate with the transmitted signal. As a result, the channel is either open or blocked depending on the state of the memory. The excitations generated by the memory ring do not spread outside the B segment. On one end their propagation is stopped by the diode D1, on the other by the geometry of the junction between the B channel and the memory output. The state of memory is controlled by the excitation pulses coming from loading and erasing channels, so switchable channels can be used in devices that are programmable with excitation pulses. Figure 21b illustrates a self-erasing memory cell that changes its state from loaded to erased after a certain time. In such a memory cell, the output channel is connected with the erasing one. When the memory is loaded the output signal appears. After some time decided by the length of connecting channels an output pulse returns as an erasing pulse and switches the memory to

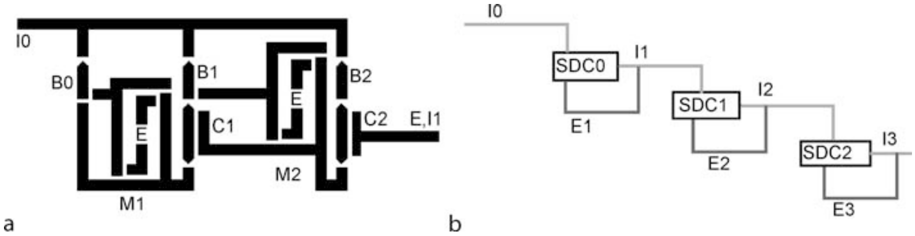


Computing in Geometrical Constrained Excitable Chemical Systems, Figure 20
Three snapshots from an experiment with memory erasing. The memory ring is formed by two L-shaped excitable channels, the Z-shaped erasing channel is inside the ring



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 21

Two simple applications of a memory cell. a A switchable unidirectional channel that stops or transmits signals depending on the state of memory. b A self-erasing memory cell that changes its state after a certain time. SEC marks the connection between the memory output and the erasing channel



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 22

The counter of excitation pulses that arrive at the input I0. a shows the geometry of excitable channels (*black*) in a single digit counter for the positional representation with the base 3. b is a schematic illustration of the cascade of single digit counters that provides a positional representation. The feedback signals from E1, E2 and E3 channels erase the memory of the single digit counters

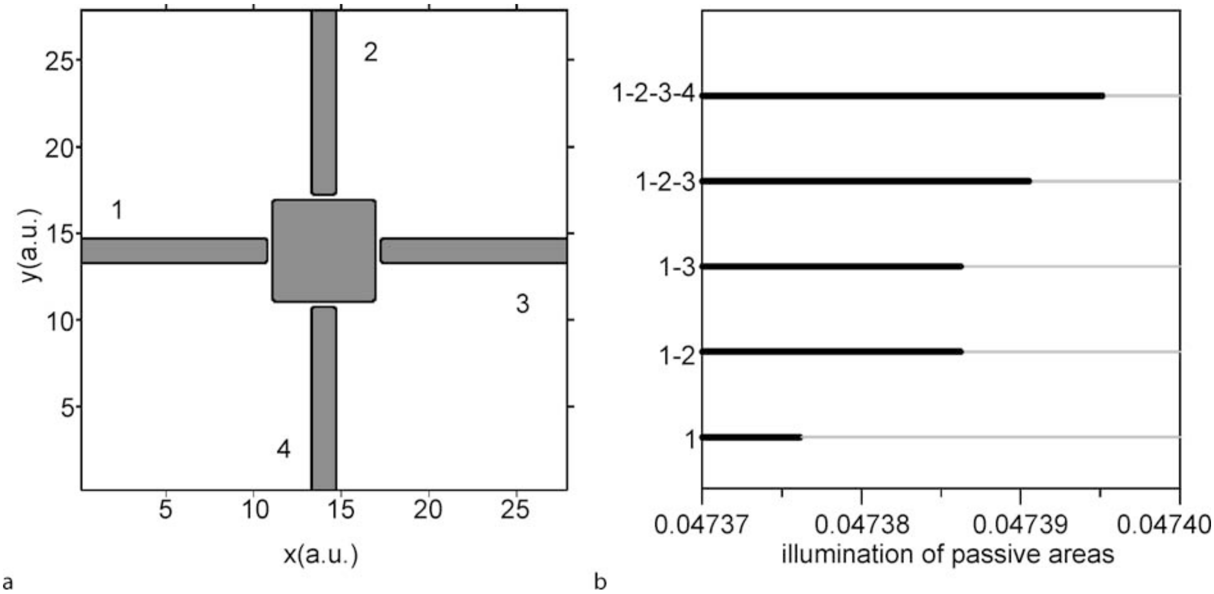
unloaded state. This behavior is an example of a simple feedback process common in self regulating information processing systems.

Using memory cells, signal diodes, and coincidence detectors, one can construct devices which perform more complex signal processing operations. As an example, we present a simple chemical realization of a device that counts arriving spikes and returns their number in any chosen positional representation [27]. Such a counter can be assembled from single digit counters. The construction of a single digit counter depends on the representation used. Here, as an example, we consider the positional representation with the base 3. The geometry of a single digit counter is schematically shown in Fig. 22. Its main elements are two memory cells M_1 and M_2 and two coincidence detectors C_1 and C_2. At the beginning let us assume that none of the memory cells are loaded. When the first pulse arrives through the input channel I_0, it splits at all junctions and excitations enter segments B_0, B_1 and B_2. The pulse that has propagated through B_0 loads the memory cell M_1. The pulses that have propagated through B_1 and B_2 die at the bottom diodes of segments C_1 and C_2 respectively. Thus, the first input pulse loads the memory M_1 and does not change the state of M_2. When M_1 is loaded, pulses of excitation are periodically sent to segments B_0 and C_1 via the bottom channel. Now let us consider what happen when the second pulse arrives. It does not pass through B_0 because it annihilates with the pulses arriving from the memory M_1. The excitations generated by the second pulse can enter B_1 and B_2. The excitation that propagated through B_2 dies at the bottom diode of the segment C_2. The pulse that has propagated through B_1 enters C_1, annihilates with a pulse from memory M_1 and activates the coincidence detector. The output pulse from the coincidence detector loads the memory M_2. Therefore, after the second input pulse both memories M_1 and M_2 are loaded. If the third pulse arrives the segments B_0 and B_1 are blocked by spikes sent from the memory rings. The generated excitation can enter channel B_2 and its collision with a pulse coming from the memory cell M_2 activates the output channel of C_2. The output signal is directed to the counter of responsible for the digit at next position (I1) and it is also used to erase all memory cells. Thus after the third pulse both memory cells M1 and M2 are erased. The counter shown in Fig. 22 returns a digit in a representation with the base 3: here 0 is represented by the $(M_1, M_2) = (0, 0)$, 1 by $(1, 0)$, 2 by $(1, 1)$ and the next pulse changes the state of memory cell into $(M_1, M_2) = (0, 0)$. Of course, using $n - 1$ memory cells in a single digit counter we can represent digits of the system with base n. A cascade of single digit counters (see Fig. 22b) gives a positional representation of the number of arriving pulses.

Artificial Chemical Neurons with Excitable Medium

In this section we discuss a simple realization of an artificial neuron with structured excitable medium and show how neuron networks can be used in programmable devices. We consider a chemical analogy of the McCulloch–Pitts neuron, i. e., a device that produces the output signal if the combined activation exceeds a critical value [32]. The geometry of a chemical neuron is inspired by a the structure of biological neuron [30]. One of its realizations with an excitable medium is illustrated on Fig. 23. Another geometry of a neuron has been discussed in [24]. In an artificial chemical neuron, like in real neurons, dendrites (input channels 1–4) transmit weak signals which are added together through the processes of spatial and temporal integration inside the cell body (part C). If the aggregate excitation is larger than the threshold value the cell body gets excited. This excitation is transmitted as an output signal down the axon (the output channel) and the amplitude of the output signal does not depend on the value of integrated inputs but only on the properties of the medium that makes the output channel. In Fig. 23a the axon is not shown and we assume that it is formed by an excitable channel located perpendicularly above the cell body. In the construction discussed in [24] both dendrites and the axon were on a single plane. We have studied the neuron using numerical simulations based on the Oregonator model and the reaction-diffusion equations have been solved on a square grid. The square shape of the neuron body and the input channels shown on Fig. 23a allows for precise definition of the boundary between the excitable and non-excitable parts. The idea behind the chemical neuron is similar to that of the AND gate: perturbations introduced by multiple inputs combine and generate a stronger excitation of the cell body than this resulting from a single input pulse. Therefore, it is intuitively clear that if we are able to adjust the amplitudes of excitations coming from individual input channels, then the output channel becomes excited only when the required number of excitations arrive from the inputs.

In the studied neuron, the amplitudes of spikes in input channels have been adjusted by sub-excitability of



Computing in Geometrical Constrained Excitable Chemical Systems, Figure 23
Artificial chemical neuron. a The geometry of excitable and non-excitable areas; b The response of the neuron to different types of excitations as a function of the illumination of non-excitable regions. The numbers given on the *left list* the excited channels. The values of ϕ_p for which the neuron body gets excited are marked by a *thick line*

these channels which can be controlled by the channel width or by the illumination of surrounding non-excitable medium. In our simulations we considered different values of ϕ_p for non-excitable areas, whereas $\phi_a = 0.007$ characterizing excitable regions has been fixed for dendrites and the neuron body. Simulation results shown on Fig. 23b indicate that the properties of chemical neurons are very sensitive with respect to changes in ϕ_p. The thick line marks the values of ϕ_p for which the output signal appears. For the parameters used when ϕ_p is slightly below 0.047377 any single excitation produces an output signal. On the other hand if $\phi_p > 0.047397$ then even the combined excitation of all inputs is not sufficient to excite the neuron. In between those two limiting values we observe all other thresholds; i. e., an output excitation as the result of two or three combined inputs. Therefore, by applying the proper illumination level the structure shown in Fig. 23a can work as a four input McCulloch–Pitts neuron with the required threshold. A similar high sensitivity of the neuron properties on ϕ_a is also observed. It means that neuron properties can be controlled with tiny changes in system parameters.

The chemical neuron illustrated in Fig. 23a can be used to program signal processing with pulses of excitation. If we set the illuminations such that any two input pulses produce the output and use one of the inputs to control the device, then if the control pulse is present the device performs the OR operation on the other inputs. If there is no control pulse, it calculates an alternative on conjunctions of all pairs of channels.

The geometry of a network constructed with neurons can be easily controlled if the switchable channels illustrated on Fig. 21b are used to establish or cut connections between processing elements. The state of the memory that controls the channel may depend on the state of the network through a feedback mechanism what allows for network training. In the chemical programming described above pulses of concentration of the same reagent are used to store and process information, and to program the medium. Therefore, the output signal may be directly used to change the network geometry. External programming factors like illumination or a temperature field [3] are difficult for applications in three dimensional structures. The pulse based programming seems to be easier if the proper geometry of switchable channels and of the feedbacks is introduced.

At the first look it seems that a practical realization of a network built of chemical neurons may be difficult. However, the geometry of a considered neuron looks quite similar to structures of phases formed in a multicomponent system. For example, the diamond structure in an oil-water-surfactant system, which spontaneously appears at certain thermodynamic conditions [11], has a form of centers linked with the four nearest neighbors. If the reactants corresponding for excitability are soluble in water, but not in oil then the water rich phase forms the structure of excitable channels and processing elements just as the result of thermodynamic conditions. Within a certain range of parameters, such a structure is thermodynamically stable. This means that the network has an auto-repair ability and robustness against unexpected destruction. The subexcitability of a channel is related to its diameter, so the required value can be obtained by selecting the right composition of the mixture and the conditions at which the phase transition occurs. Moreover, the structure is three dimensional, which allows for a higher density of processing elements than that obtained with the classical two-dimensional techniques, for example lithography.

Perspectives and Conclusions

Perspectives

In this article we have described a number of simple devices constructed with structured excitable chemical media which process information coded in excitation pulses. All the considered systems process information in an unconventional (non-von Neumann) way; i. e., without an external clock or synchronizing signal that controls the sequence of operations. On the other hand, in many cases the right timing of performed operation is hidden in the geometrical distribution and sizes of excitable regions. The described devices can be used as building blocks for more complex systems that process signals formed of excitation pulses. Some of the discussed devices can be controlled with spikes. Therefore, there is room for programming and learning. However, the further development of applications of structured excitable medium for information processing along this line seem to follow the evolution of classical electronic computing. In our opinion it would be more interesting to step away from this path and learn more about the potential offered by excitable media.

It would be interesting to study new types of excitable media suitable for information processing. Electrical analogs of reaction-diffusion systems (▶ Unconventional Computing, Novel Hardware for) seem promising, because they are more robust than the wetware based on liquid BZ media. In such media, spikes propagate much faster and the spatial scale can be much reduced if compared with typical chemical systems. They seem to be promising candidates for the hardware in geometrically oriented problems and direct image processing [4]. However, two interesting properties of chemical reaction-diffusion systems are lost in their electrical analogs. First, chemical information processing systems, unlike the electronic ones, integrate two functions: the chemical reactivity and the ability to process information. Having in mind the excitable oxidation of CO on a Pt surface and the potential application of this medium for information processing [23], we can think of catalysts that are able to monitor their activity and report it. Second, the media described in ▶ Unconventional Computing, Novel Hardware for are two-dimensional. Most of the systems discussed in this article can be realized in three dimensions. The use of two dimensional media in chemical experiments is mainly related with significant difficulties in observations of three dimensional chemical excitations [42]. Potential application of phase transitions for channel structure generation described in the previous section should increase the interest in information processing with three dimensional medium.

Studies on more effective methods of information coding are important for the further development of reaction-diffusion computing. The most simple translation of chemistry into the language of information science based on the equivalence between the presence of a spike and the TRUE logic value is certainly not the most efficient. It can be expected that information interpreted within multi-valued logic systems is more suitable for transformations with the use of structured media [47]. As an example, let us consider three-valued logic encoded by excitation pulses in the following way: the lack of a pulse represents a logical FALSE (F), two pulses separated by time δt correspond at the logical TRUE (T), and one pulse is interpreted as a "nonsense" (★). Let us assume that two signals coded as described above are directed to the inputs of the AND gate illustrated on Fig. 9 and that they are fully synchronized at the input. The gap between the input and output channels is adjusted such that an activator diffusing from a single traveling pulse is not sufficient to generate a new pulse in the output channel, but the excitation resulting from the collision of facing pulses at the junction point exceeds the threshold and the output pulse is generated. Moreover, let us assume that the system is described by the dynamics for which signals are transformed as illustrated in Fig. 7a (for example FitzHugh–Nagumo dynamics) and that the time difference between spikes δt is selected such that the second spike can pass the gap. The output signal appears when spikes from both input channels are in coincidence or when the second spike from the same input arrives. As a result the device performs the following function within 3-valued logic [47]:

$\boxdot_{\mathcal{F}_1}$	T	F	$\star$
T	T	$\star$	$\star$
F	$\star$	F	F
$\star$	$\star$	F	$\star$

The same operation, if performed on a classical computer would require a procedure that measures time between spikes. The structured excitable medium performs it naturally provided that the time δt is adjusted with the properties of the medium and the geometry of the gap. Of course, similar devices that process variables of n-valued logic can be also constructed with structured excitable media.

Conclusions

In the article we have presented a number of examples that should convince the reader that structured excitable media can be used for information processing. The future research will verify if this branch of computer science is fruitful. It would be important to find new algorithms that can be efficiently executed using a structured excitable medium. However, the ultimate test for the usefulness of ideas should come from biology. It is commonly accepted that excitable behavior is responsible for information processing and coding in living organisms [8,31,36,60]. We believe that studies on chemical information processing will help us to better understand these problems. And, although at the moment computing with a homogeneous excitable medium seems to offer more applications than that with the structured one, we believe the proportions will be reversed in the future. After all, our brains are not made of a single piece of a homogeneous excitable medium.

Acknowledgments

The research on information processing with structured excitable medium has been supported by the Polish State Committee for Scientific Research project 1 P03B 035 27.

Bibliography

Primary Literature

1. Adamatzky A, De Lacy Costello B (2002) Experimental logical gates in a reaction-diffusion medium: The XOR gate and beyond. Phys Rev E 66:046112
2. Adamatzky A (2004) Collision-based computing in Belousov–Zhabotinsky medium. Chaos Soliton Fractal 21(5):1259–1264
3. Adamatzky A (2005) Programming Reaction-Diffusion Processors. In: Banatre J-P, Fradet P, Giavitto J-L, Michel O (eds) LNCS, vol 3566. Springer, pp 47–55

4. Adamatzky A, De Lacy Costello B, Asai T (2005) Reaction-Diffusion Computers. Elsevier, UK
5. Agladze K, Aliev RR, Yamaguchi T, Yoshikawa K (1996) Chemical diode. J Phys Chem 100:13895–13897
6. Agladze K, Magome N, Aliev R, Yamaguchi T, Yoshikawa K (1997) Finding the optimal path with the aid of chemical wave. Phys D 106:247–254
7. Agladze K, Tóth Á, Ichino T, Yoshikawa K (2000) Propagation of Chemical Pulses at the Boundary of Excitable and Inhibitory Fields. J Phys Chem A 104:6677–6680
8. Agmon-Snir H, Carr CE, Rinzel J (1998) The role of dendrites in auditory coincidence detection. Nature 393:268–272
9. Amemiya T, Ohmori T, Yamaguchi T (2000) An Oregonator-class model for photoinduced Behavior in the $Ru(bpy)_3^{2+}$–Catalyzed Belousov–Zhabotinsky reaction. J Phys Chem A 104:336–344
10. Armstrong GR, Taylor AF, Scott SK, Gaspar V (2004) Modelling wave propagation across a series of gaps. Phys Chem Chem Phys 6:4677–4681
11. Babin V, Ciach A (2003) Response of the bicontinuous cubic D phase in amphiphilic systems to compression or expansion. J Chem Phys 119:6217–6231
12. Bertram M, Mikhailov AS (2003) Pattern formation on the edge of chaos: Mathematical modeling of CO oxidation on a Pt(110) surface under global delayed feedback. Phys Rev E 67:036207
13. Dolnik M, Finkeova I, Schreiber I, Marek M (1989) Dynamics of forced excitable and oscillatory chemical-reaction systems. J Phys Chem 93:2764–2774; Finkeova I, Dolnik M, Hrudka B, Marek M (1990) Excitable chemical reaction systems in a continuous stirred tank reactor. J Phys Chem 94:4110–4115; Dolnik M, Marek M (1991) Phase excitation curves in the model of forced excitable reaction system. J Phys Chem 95:7267–7272; Dolnik M, Marek M, Epstein IR (1992) Resonances in periodically forced excitable systems. J Phys Chem 96:3218–3224
14. Epstein IR, Showalter K (1996) Nonlinear Chemical Dynamics: Oscillations, Patterns, and Chaos. J Phys Chem 100:13132–13147
15. Feynman RP, Allen RW, Heywould T (2000) Feynman Lectures on Computation. Perseus Books, New York
16. Field RJ, Körös E, Noyes RM (1972) Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the bromate-cerium-malonic acid system. J Am Chem Soc 94:8649–8664
17. FitzHugh R (1960) Thresholds and plateaus in the Hodgkin-Huxley nerve equations. J Gen Physiol 43:867–896
18. FitzHugh R (1961) Impulses and physiological states in theoretical models of nerve membrane. Biophys J 1:445–466
19. Field RJ, Noyes RM (1974) Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. J Chem Phys 60:1877–1884
20. Gaspar V, Bazsa G, Beck MT (1983) The influence of visible light on the Belousov–Zhabotinskii oscillating reactions applying different catalysts. Z Phys Chem (Leipzig) 264:43–48
21. Ginn BT, Steinbock B, Kahveci M, Steinbock O (2004) Microfluidic Systems for the Belousov–Zhabotinsky Reaction. J Phys Chem A 108:1325–1332
22. Gorecka J, Gorecki J (2003) T-shaped coincidence detector as a band filter of chemical signal frequency. Phys Rev E 67:067203
23. Gorecka J, Gorecki J (2005) On one dimensional chemical diode and frequency generator constructed with an excitable surface reaction. Phys Chem Chem Phys 7:2915–2920
24. Gorecka J, Gorecki J (2006) Multiargument logical operations performed with excitable chemical medium. J Chem Phys 124:084101 (1–5)
25. Gorecka J, Gorecki J, Igarashi Y (2007) One dimensional chemical signal diode constructed with two nonexcitable barriers. J Phys Chem A 111:885–889
26. Gorecki J, Kawczynski AL (1996) Molecular dynamics simulations of a thermochemical system in bistable and excitable regimes. J Phys Chem 100:19371–19379
27. Gorecki J, Yoshikawa K, Igarashi Y (2003) On chemical reactors that can count. J Phys Chem A 107:1664–1669
28. Gorecki J, Gorecka JN, Yoshikawa K, Igarashi Y, Nagahara H (2005) Sensing the distance to a source of periodic oscillations in a nonlinear chemical medium with the output information coded in frequency of excitation pulses. Phys Rev E 72:046201 (1–7)
29. Gorecki J, Gorecka JN (2006) Information processing with chemical excitations – from instant machines to an artificial chemical brain. Int J Unconv Comput 2:321–336
30. Haken H (2002) Brain Dynamics. In: Springer Series in Synergetics. Springer, Berlin
31. Häusser M, Spruston N, Stuart GJ (2000) Diversity and Dynamics of Dendritic Signaling. Science 290:739–744
32. Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Addison-Wesley, Redwood City
33. Ichino T, Igarashi Y, Motoike IN, Yoshikawa K (2003) Different operations on a single circuit: Field computation on an Excitable Chemical System. J Chem Phys 118:8185–8190
34. Igarashi Y, Gorecki J, Gorecka JN (2006) Chemical information processing devices constructed using a nonlinear medium with controlled excitability. Lect Note Comput Science 4135:130–138
35. Kapral R, Showalter K (1995) Chemical Waves and Patterns. Kluwer, Dordrecht
36. Kindzelskii AL, Petty HR (2003) Intracellular calcium waves accompany neutrophil polarization, formylmethionylleucylphenylalanine stimulation, and phagocytosis: a high speed microscopy study. J Immunol 170:64–72
37. Krischer K, Eiswirth M, Ertl GJ (1992) Oscillatory CO oxidation on Pt(110): modelling of temporal self-organization. J Chem Phys 96:9161–9172
38. Krug HJ, Pohlmann L, Kuhnert L (1990) Analysis of the modified complete Oregonator accounting for oxygen sensitivity and photosensitivity of Belousov–Zhabotinskii systems. J Phys Chem 94:4862–4866
39. Kuramoto Y (1984) Chemical oscillations, waves, and turbulence. Springer, Berlin
40. Kusumi T, Yamaguchi T, Aliev RR, Amemiya T, Ohmori T, Hashimoto H, Yoshikawa K (1997) Numerical study on time delay for chemical wave transmission via an inactive gap. Chem Phys Lett 271:355–360
41. Lázár A, Noszticzius Z, Försterling H-D, Nagy-Ungvárai Z (1995) Chemical pulses in modified membranes I. Developing the technique. Physica D 84:112–119; Volford A, Simon PL, Farkas H, Noszticzius Z (1999) Rotating chemical waves: theory and experiments. Physica A 274:30–49
42. Luengviriya C, Storb U, Hauser MJB, Müller SC (2006) An elegant method to study an isolated spiral wave in a thin layer

of a batch Belousov–Zhabotinsky reaction under oxygen-free conditions. Phys Chem Chem Phys 8:1425–1429

43. Manz N, Müller SC, Steinbock O (2000) Anomalous dispersion of chemical waves in a homogeneously catalyzed reaction system. J Phys Chem A 104:5895–5897; Steinbock O (2002) Excitable Front Geometry in Reaction-Diffusion Systems with Anomalous Dispersion. Phys Rev Lett 88:228302
44. Maselko J, Reckley JS, Showalter K (1989) Regular and irregular spatial patterns in an immobilized-catalyst Belousov–Zhabotinsky reaction. J Phys Chem 93:2774–2780
45. Mikhailov AS, Showalter K (2006) Control of waves, patterns and turbulence in chemical systems. Phys Rep 425:79–194
46. Morozov VG, Davydov NV, Davydov VA (1999) Propagation of Curved Activation Fronts in Anisotropic Excitable Media. J Biol Phys 25:87–100
47. Motoike IN, Adamatzky A (2005) Three-valued logic gates in reaction–diffusion excitable media. Chaos, Solitons & Fractals 24:107–114
48. Motoike I, Yoshikawa K (1999) Information Operations with an Excitable Field. Phys Rev E 59:5354–5360
49. Motoike IN, Yoshikawa K, Iguchi Y, Nakata S (2001) Real–Time Memory on an Excitable Field. Phys Rev E 63:036220 (1–4)
50. Motoike IN, Yoshikawa K (2003) Information operations with multiple pulses on an excitable field. Chaos, Solitons & Fractals 17:455–461
51. Murray JD (1989) Mathematical Biology. Springer, Berlin
52. Nagai Y, Gonzalez H, Shrier A, Glass L (2000) Paroxysmal Starting and Stopping of Circulatong Pulses in Excitable Media. Phys Rev Lett 84:4248–4251
53. Nagahara H, Ichino T, Yoshikawa K (2004) Direction detector on an excitable field: Field computation with coincidence detection. Phys Rev E 70:036221 (1–5)
54. Nagumo J, Arimoto S, Yoshizawa S (1962) An Active Pulse Transmission Line Simulating Nerve Axon. Proc IRE 50: 2061–2070
55. Noszticzuis Z, Horsthemke W, McCormick WD, Swinney HL, Tam WY (1987) Sustained chemical pulses in an annular gel reactor: a chemical pinwheel. Nature 329:619–620
56. Plaza F, Velarde MG, Arecchi FT, Boccaletti S, Ciofini M, Meucci R (1997) Excitability following an avalanche-collapse process. Europhys Lett 38:85–90
57. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical Recipes, 3rd edn. The Art of Scientific Computing. Available via http://www.nr.com.
58. Rambidi NG, Maximychev AV (1997) Towards a Biomolecular Computer. Information Processing Capabilities of Biomolecular Nonlinear Dynamic Media. BioSyst 41:195–211
59. Rambidi NG, Yakovenchuk D (1999) Finding paths in a labyrinth based on reaction-diffusion media. BioSyst 51:67–72; Rambidi NG, Yakovenchuk D (2001) Chemical reaction-diffusion implementation of finding the shortest paths in a labyrinth. Phys Rev E 63:026607
60. Rambidi NG (2005) Biologically Inspired Information Processing Technologies: Reaction-Diffusion Paradigm. Int J Unconv Comp 1:101–121
61. Rovinsky AB, Zhabotinsky AM (1984) Mechanism and mathematical model of the oscillating bromate-ferroin-bromomalonic acid reaction. J Phys Chem 88:6081–6084
62. Rovinsky AB (1986) Spiral waves in a model of the ferroin catalyzed Belousov–Zhabotinskii reaction. J Phys Chem 90:217–219
63. Sielewiesiuk J, Gorecki J (2002) Chemical Waves in an Excitable Medium: Their Features and Possible Applications in Information Processing. In: Klonowski W (ed) Attractors, Signals and Synergetics. 1st European Interdisciplinary School on Nonlinear Dynamics for System and Signal Analysis Euroattractor 2000, Warsaw, 6–15 June 2000. Pabst, Lengerich, pp 448–460
64. Sielewiesiuk J, Gorecki J (2002) On complex transformations of chemical signals passing through a passive barrier. Phys Rev E 66:016212; Sielewiesiuk J, Gorecki J (2002) Passive barrier as a transformer of chemical signal frequency. J Phys Chem A 106:4068–4076
65. Sielewiesiuk J, Gorecki J (2001) Chemical impulses in the perpendicular junction of two channels. Acta Phys Pol B 32:1589–1603
66. Sielewiesiuk J, Gorecki J (2001) Logical functions of a cross junction of excitable chemical media. J Phys Chem A 105:8189–8195
67. Steinbock O, Toth A, Showalter K (1995) Navigating complex labyrinths – optimal paths from chemical waves. Science 267:868–871
68. Steinbock O, Kettunen P, Showalter K (1995) Anisotropy and Spiral Organizing Centers in Patterned Excitable Media. Science 269:1857–1860
69. Steinbock O, Kettunen P (1996) Chemical clocks on the basis of rotating pulses. Measuring irrational numbers from period ratios. Chem Phys Lett 251:305–308
70. Steinbock O, Kettunen P, Showalter K (1996) Chemical Pulse Logic Gates. J Phys Chem 100:18970–18975
71. Suzuki K, Yoshinobu T, Iwasaki H (1999) Anisotropic waves propagating on two–dimensional arrays of Belousov–Zhabotinsky oscillators. Jpn J Appl Phys 38:L345–L348
72. Suzuki K, Yoshinobu T, Iwasaki H (2000) Unidirectional propagation of chemical waves through microgaps between zones with different excitability. J Phys Chem A 104:6602–6608
73. Suzuki R (1967) Mathematical analysis and application of iron-wire neuron model. IEEE Trans Biomed Eng 14:114–124
74. Taylor AF, Armstrong GR, Goodchild N, Scott SK (2003) Propagation of chemical waves across inexcitable gaps. Phys Chem Chem Phys 5:3928–3932
75. Toth A, Showalter K (1995) Logic gates in excitable media. J Chem Phys 103:2058–2066
76. Toth A, Horvath D, Yoshikawa K (2001) Unidirectional wave propagation in one spatial dimension. Chem Phys Lett 345: 471–474
77. Yoshikawa K, Nagahara H, Ichino T, Gorecki J, Gorecka JN, Igarashi Y (2009) On Chemical Methods of Direction and Distance Sensing. Int J Unconv Comput 5:53–65

Books and Reviews

Hjelmfelt A, Ross J (1994) Pattern recognition, chaos, and multiplicity in neural networks of excitable systems. Proc Natl Acad Sci USA 91:63–67

Nakata S (2003) Chemical Analysis Based on Nonlinearity. Nova, New York

Rambidi NG (1998) Neural network devices based on reaction–diffusion media: an Approach to artificial retina. Supramol Sci 5:765–767

Storb U, Müller SC (2004) Scroll waves. In: Scott A (ed) Encyclopedia of nonlinear sciences. Routledge, Taylor and Francis Group, New York, pp 825–827

Computing with Solitons

DARREN RAND[1], KEN STEIGLITZ[2]
[1] Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, USA
[2] Computer Science Department, Princeton University, Princeton, USA

Article Outline

Glossary

Integrable This term is generally used in more than one way and in different contexts. For the purposes of this article, a partial differential equation or system of partial differential equations is *integrable* if it can be solved explicitly to yield *solitons* (qv).

Manakov system A system of two cubic Schrödinger equations where the self- and cross-phase modulation terms have equal weight.

Nonlinear Schrödinger equation A partial differential equation that has the same form as the Schrödinger equation of quantum mechanics, with a term nonlinear in the dependent variable, and for the purposes of this article, interpreted classically.

Self- and cross-phase modulation Any terms in a nonlinear Schrödinger equation that involve nonlinear functions of the dependent variable of the equation, or nonlinear functions of a dependent variable of another (coupled) equation, respectively.

Solitary wave A solitary wave is a wave characterized by undistorted propagation. Solitary waves do not in general maintain their shape under perturbations or collisions.

Soliton A soliton is a solitary wave which is also robust under perturbations and collisions.

Turing equivalent Capable of simulating any Turing Machine, and hence by Turing's Thesis capable of performing any computation that can be carried out by a sequence of effective instructions on a finite amount of data. A machine that is Turing equivalent is therefore as powerful as any digital computer. Sometimes a device that is Turing equivalent is called "universal."

Definition of the Subject

Solitons are localized, shape-preserving waves characterized by robust collisions. First observed as water waves by John Scott Russell [29] in the Union Canal near Edinburgh and subsequently recreated in the laboratory, solitons arise in a variety of physical systems, as both temporal pulses which counteract dispersion and spatial beams which counteract diffraction.

Solitons with two components, vector solitons, are computationally universal due to their remarkable collision properties. In this article, we describe in detail the characteristics of Manakov solitons, a specific type of vector soliton, and their applications in computing.

Introduction

In this section, we review the basic principles of soliton theory and spotlight relevant experimental results. Interestingly, the phenomena of soliton propagation and collision occur in many physical systems despite the diversity of mechanisms that bring about their existence. For this reason, the discussion in this article will treat temporal and spatial solitons interchangeably, unless otherwise noted.

Scalar Solitons

A pulse in optical fiber undergoes dispersion, or temporal spreading, during propagation. This effect arises because the refractive index of the silica glass is not constant, but is rather a function of frequency. The pulse can be decomposed into a frequency range—the shorter the pulse, the broader its spectral width. The frequency dependence of the refractive index will cause the different frequencies of the pulse to propagate at different velocities, giving rise to dispersion. As a result, the pulse develops a chirp, meaning that the individual frequency components are not evenly distributed throughout the pulse. There are two types of dispersion: normal and anomalous. If the longer wavelengths travel faster, the medium is said to have *normal*dispersion. If the opposite is true, the medium has *anomalous* dispersion.

The response of a dielectric such as optical fiber is nonlinear. Most of the nonlinear effects in fiber originate from nonlinear refraction, where the refractive index n depends on the intensity of the propagating field according to the relation

$$n = n_0 + n_2|E|^2 , \tag{1}$$

where n_0 is the linear part of the refractive index, $|E^2|$ is the optical intensity, and n_2 is the coefficient of nonlinear contribution to the refractive index. Because the material responds almost instantaneously, on the order of femtoseconds, and because the phase shift $\Delta\phi$ is proportional to n, each component of an intense optical pulse sees a phase shift proportional to its intensity. Since the frequency shift $\delta\omega = -(\partial\Delta\phi)/(\partial t)$, the leading edge of the pulse is red-shifted ($\delta\omega < 0$), while the trailing edge is blue-shifted ($\delta\omega > 0$), an effect known as self-phase modulation (SPM). As a result, if the medium exhibits normal dispersion, the pulse is broadened; for anomalous dispersion, the pulse is compressed. Under the proper conditions, this pulse compression can exactly cancel the linear, dispersion-induced broadening, resulting in distortionless soliton propagation. For more details, see the book by Agrawal [3].

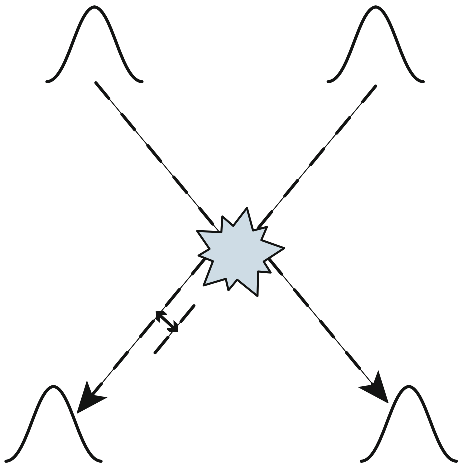
The idealized mathematical model for this pulse propagation is the nonlinear Schrödinger equation (NLSE):

$$i\frac{\partial u}{\partial z} \pm \frac{1}{2}\frac{\partial^2 u}{\partial x^2} + |u|^2 u = 0, \tag{2}$$

where $u(z,x)$ is the complex-valued field envelope, z is a normalized propagation distance and x is normalized time propagating with the group velocity of the pulse. The second and third terms describe dispersion and the intensity-dependent Kerr nonlinearity, respectively. The coefficient of the dispersion term is positive for anomalous dispersion and negative for normal dispersion. Equation (2), known as the *scalar* NLSE, is integrable—that is, it can be solved analytically, and collisions between solitons are 'elastic,' in that no change in amplitude or velocity occurs as a result of a collision. Zakharov and Shabat [38] first solved this equation analytically using the inverse scattering method. It describes, for example, the propagation of picosecond or longer pulses propagating in lossless optical fiber.

Two solitons at different wavelengths will collide in an optical fiber due to dispersion-induced velocity differences. A schematic of such a collision is depicted in Fig. 1. The scalar soliton collision is characterized by two phenomena— a position and phase shift—both of which can be understood in the same intuitive way. During collision, there will be a local increase in intensity, causing a local increase in the fiber's refractive index, according to Eq. (1). As a result, both the soliton velocity and phase will be affected during the collision.

From an all-optical signal processing perspective, the phase and position shifts in a soliton collision are not useful. This is because these effects are independent of any soliton properties that are changed by collision; that is, the result of one collision will not affect the result of subsequent collisions. Scalar solitons are therefore not useful for complex logic or computing, which depend on multiple, cascaded interactions.



Computing with Solitons, Figure 1
Schematic of a scalar soliton collision, in which amplitude and velocities are unchanged. The two soliton collision effects are a position shift (depicted through the translational shift in the soliton path) and phase shift (not pictured)

Despite this setback, it was discovered later that a system similar to the scalar NLSE, the Manakov system [19], possesses very rich collisional properties [26] and is integrable as well. Manakov solitons are a specific instance of two-component vector solitons, and it has been shown that collisions of Manakov solitons are capable of transferring information via changes in a complex-valued polarization state [16].

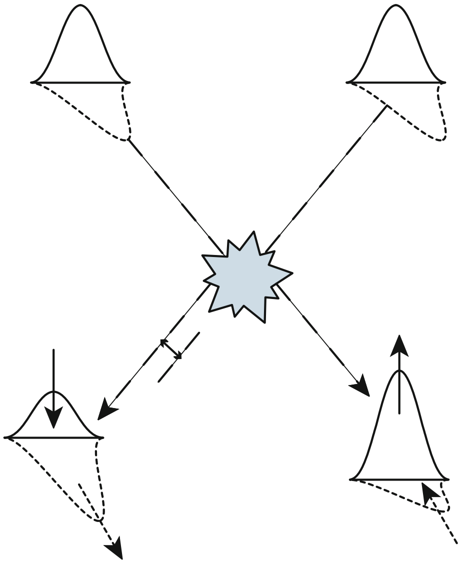
Vector Solitons

When several field components, distinguished by polarization and/or frequency, propagate in a nonlinear medium, the nonlinear interaction between them must be considered as well. This interaction between field components results in intensity-dependent nonlinear coupling terms analogous to the self-phase modulation term in the scalar case. Such a situation gives rise to a set of coupled nonlinear Schrödinger equations, and may allow for propagation of *vector* solitons. For the case of two components propagating in an ideal sl medium with no higher-order effects and only intensity-dependent nonlinear coupling, the equations become:

$$\begin{aligned} i\frac{\partial u_1}{\partial z} + \frac{\partial^2 u_1}{\partial x^2} + 2\mu(|u_1|^2 + \alpha|u_2|^2)u_1 = 0\,, \\ i\frac{\partial u_2}{\partial z} + \frac{\partial^2 u_2}{\partial x^2} + 2\mu(|u_2|^2 + \alpha|u_1|^2)u_2 = 0\,, \end{aligned} \tag{3}$$

where $u_1(z,x)$ and $u_2(z,x)$ are the complex-valued pulse envelopes for each component, μ is a nonlinearity parameter, and α describes the ratio between self- and cross-phase modulation contributions to the overall nonlinearity. Only for the special case of $\alpha = 1$ are Eqs. (3) integrable. First solved using the method of inverse scattering by Manakov [19], Eqs. (3) admit solutions known as Manakov solitons. For nonintegrable cases ($\alpha \neq 1$), some analytical solitary-wave solutions are known for specific cases, although in general a numerical approach is required [36]. The specific case of $\alpha = 2/3$, for example, corresponds to linearly birefringent polarization maintaining fiber, and will be considered in more detail in Sect. "Experiments".

Due to their multicomponent structure, vector solitons have far richer collision dynamics than their scalar, one-component counterparts. Recall that scalar collisions are characterized by phase and position shifts only. Vector soliton collisions also exhibit these effects, with the added feature of possible intensity redistributions between the component fields [19,26]. This process is shown schematically in Fig. 2. In the collision, two conservation relations are satisfied: (i) the energy in each soliton is conserved and (ii) the energy in each component is conserved. It can be seen that when the amplitude of one component in a soliton increases as a result of the collision, the other component decreases, with the opposite exchange in the second soliton. The experimental observation of this effect will be discussed in Sect. "Experiments". In addition to fundamental interest in such solitons, collisions of vector solitons make possible unique applications, including collision-based logic and universal computation [16,27,34,35], as discussed in Sect. "Manakov Soliton Computing".



Computing with Solitons, Figure 2
Schematic of a vector soliton collision, which exhibits a position shift and phase shift (not pictured), similar to the scalar soliton collision (cf. Fig. 1). Vector soliton collisions also display an energy redistribution among the component fields, shown here as two orthogonal polarizations. Arrows indicate direction of energy redistribution

Manakov Solitons

As mentioned in Sect. "Introduction", computation is possible using vector solitons because of an energy redistribution that occurs in a collision. In this section, we provide the mathematic background of Manakov soliton theory, in order to understand soliton computing and a remarkable way to achieve bistability using soliton collisions as described in Sects. "Manakov Soliton Computing" and "Multistable Soliton Collision Cycles", respectively.

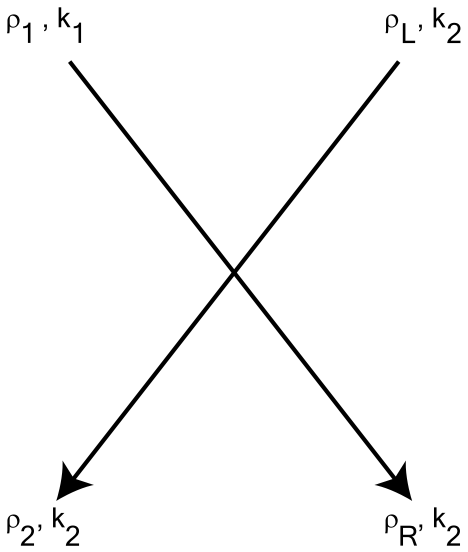
The Manakov system consists of two coupled NLSEs [19]:

$$\begin{aligned} i\frac{\partial q_1}{\partial z} + \frac{\partial^2 q_1}{\partial x^2} + 2\mu(|q_1|^2 + |q_2|^2)q_1 = 0\,, \\ i\frac{\partial q_2}{\partial z} + \frac{\partial^2 q_2}{\partial x^2} + 2\mu(|q_1|^2 + |q_2|^2)q_2 = 0\,, \end{aligned} \tag{4}$$

where $q_1(x,z)$ and $q_2(x,z)$ are two interacting optical components, μ is a positive parameter representing the strength of the nonlinearity, and x and z are normalized space and propagation distance, respectively. As mentioned in Sect. "Vector Solitons", the Manakov system is a special case of Eqs. (3) with $\alpha = 1$. The two components can be thought of as components in two polarizations, or, as in the case of a photorefractive crystal, two uncorrelated beams [11].

Manakov first solved Eqs. (4) by the method of inverse scattering [19]. The system admits single-soliton, two-component solutions that can be characterized by a complex number $k \equiv k_R + ik_I$, where k_R represents the energy of the soliton and k_I the velocity, all in normalized units. The additional soliton parameter is the complex-valued polarization state $\rho \equiv q_1/q_2$, defined as the (z- and x-independent) ratio between the q_1 and q_2 components.

Figure 3 shows the schematic for a general two-soliton collision, with initial parameters ρ_1, k_1 and ρ_L, k_2, corresponding to the right-moving and left-moving solitons, respectively. The values of k_1 and k_2 remain constant during collision, but in general the polarization state changes. Let ρ_1 and ρ_L denote the respective soliton states before impact, and suppose the collision transforms ρ_1 into ρ_R, and ρ_L into ρ_2. It turns out that the state change undergone



Computing with Solitons, Figure 3
Schematic of a general two-soliton collision. Each soliton is characterized by a complex-valued polarization state *ρ* and complex parameter *k*. Reprinted with permission from [34]. Copyright by the American Physical Society

by each colliding soliton takes on the very simple form of a linear fractional transformation (also called a bilinear or Möbius transformation). Explicitly, the state of the emerging left-moving soliton is given by [16]:

$$\rho_2 = \frac{[(1-g)/\rho_1^* + \rho_1]\rho_L + g\rho_1/\rho_1^*}{g\rho_L + (1-g)\rho_1 + 1/\rho_1^*} , \tag{5}$$

where

$$g \equiv \frac{k_1 + k_1^*}{k_2 + k_1^*} . \tag{6}$$

The state of the right-moving soliton is obtained similarly, and is

$$\rho_R = \frac{[(1-h^*)/\rho_L^* + \rho_L]\rho_1 + h^*\rho_L/\rho_L^*}{h^*\rho_1 + (1-h^*)\rho_L + 1/\rho_L^*} , \tag{7}$$

where

$$h \equiv \frac{k_2 + k_2^*}{k_1 + k_2^*} . \tag{8}$$

We assume here, without loss of generality, that k_{1R}, $k_{2R} > 0$.

Several properties of the linear fractional transformations in Eqs. (5) and (7) are derived in [16], including the characterization of inverse operators, fixed points, and implicit forms. In particular, when viewed as an operator every soliton has an *inverse*, which will undo the effect of the operator on the state. Note that this requires that the inverse operator have the same k parameter as the original, a condition that will hold in our application of computing in the next section.

These state transformations were first used by Jakubowski et al. [16] to describe logical operations such as NOT. Later, Steiglitz [34] established that arbitrary computation was possible through time gating of Manakov $(1+1)$-dimensional spatial solitons. We will describe this in Sect. "Manakov Soliton Computing".

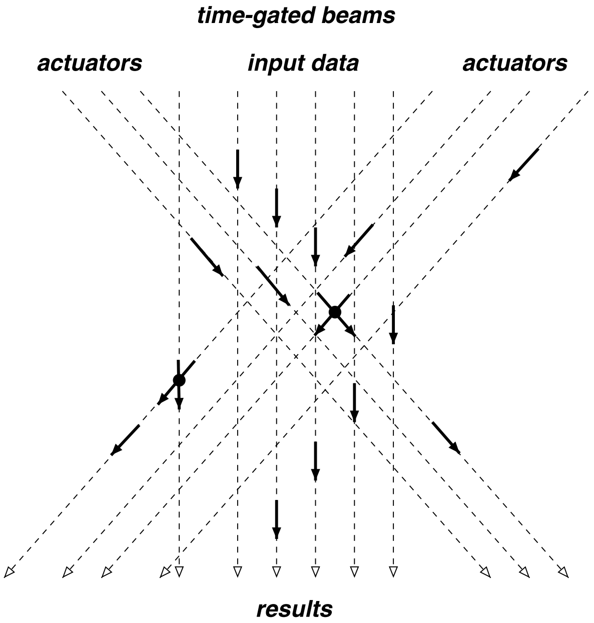
There exist several candidates for the physical realization of Manakov solitons, including photorefractive crystals [4,5,9,11,30], semiconductor waveguides [17], quadratic media [33], and optical fiber [23,28]. In Sect. "Experiments", we discuss in detail an experiment with vector solitons in linearly birefringent optical fiber.

Manakov Soliton Computing

We described in the previous section how collisions of Manakov solitons can be described by transformations of a complex-valued state which is the ratio between the two Manakov components. We show in this section that general computation is possible if we use (1+1)-dimensional spatial solitons that are governed by the Manakov equations and if we are allowed to time-gate the beams input to the medium. The result is a dynamic computer without spatially fixed gates or wires, which is unlike most present-day conceptions of a computer that involve integrated circuits, in which information travels between logical elements that are fixed spatially through fabrication on a silicon wafer. We can call such a scheme "nonlithographic," in the sense that there is no architecture imprinted on the medium.

The requirements for computation include cascadability, fanout, and Boolean completeness. The first, cascadability, requires that the output of one device can serve as input to another. Since any useful computation consists of many stages of logic, this condition is essential. The second, fanout, refers to the ability of a logic gate to drive at least two similar gates. Finally, Boolean completeness makes it possible to perform arbitrary computation.

We should emphasize that although the model we use is meant to reflect known physical phenomena, at least in the limit of ideal behavior, the result is a mathematical one. Practical considerations of size and speed are not considered here, nor are questions of error propagation. In this sense the program of this article is analogous to Fredkin and Toffoli [13] for ideal billiard balls, and Shor [31] for quantum mechanics. There are however several can-



Computing with Solitons, Figure 4

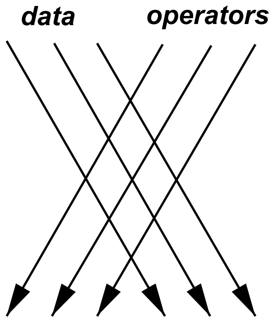
The general physical arrangement considered in this paper. Time-gated beams of spatial Manakov solitons enter at the top of the medium, and their collisions result in state changes that reflect computation. Each solid arrow represents a beam segment in a particular state. Reprinted with permission from [34]. Copyright by the American Physical Society

didates for physical instantiation of the basic ideas in this paper, as noted in the previous section.

Although we are describing computation embedded in a homogeneous medium, and not interconnected *gates* in the usual sense of the word, we will nevertheless use the term *gates* to describe prearranged sequences of soliton collisions that effect logical operations. We will in fact adopt other computer terms to our purpose, such as *wiring* to represent the means of moving information from one place to another, and *memory* to store it in certain ways for future use.

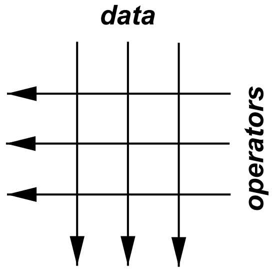
We will proceed in the construction of what amounts to a complete computer in the following stages: First we will describe a basic gate that can be used for FANOUT. Then we will show how the same basic configuration can be used for NOT, and finally, NAND. Then we will describe ways to use time gating of the input beams to interconnect signals. The NAND gate, FANOUT, and interconnect are sufficient to implement any computer, and we conclude with a layout scheme for a general-purpose, and hence Turing-equivalent computer. The general picture of the physical arrangement is shown in Fig. 4.

Figure 5 shows the usual picture of colliding solitons, which can work interchangeably for the case of temporal or spatial solitons. It is convenient for visualization purposes to turn the picture and adjust the scale so the axes are horizontal and vertical, as in Fig. 6. We will use binary logic, with two distinguished, distinct complex numbers representing TRUE and FALSE, called 1 and 0, respectively. In fact, it turns out to be possible to use complex 1 and 0 for these two state values, and we will do that throughout this paper, but this is a convenience and not at all a necessity. We will thus use complex polarization states 1 and 0 and logical 1 and 0 interchangeably.



Computing with Solitons, Figure 5

Colliding spatial solitons. Reprinted with permission from [34]. Copyright by the American Physical Society

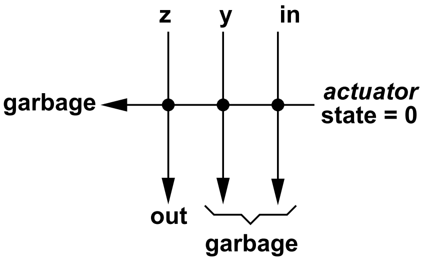


Computing with Solitons, Figure 6

Convenient representation of colliding spatial solitons. Reprinted with permission from [34]. Copyright by the American Physical Society

FANOUT

We construct the FANOUT gate by starting with a COPY gate, implemented with collisions between three down-moving, vertical solitons and one left-moving horizontal soliton. Figure 7 shows the arrangement. The soliton state labeled *in* will carry a logical value, and so be in one of the two states 0 or 1. The left-moving soliton labeled *actuator* will be in the fixed state 0, as will be the case throughout this paper. The plan is to adjust the (so far) arbitrary states z and y so that *out* = *in*, justifying the name COPY. It is reasonable to expect that this might be possible, be-



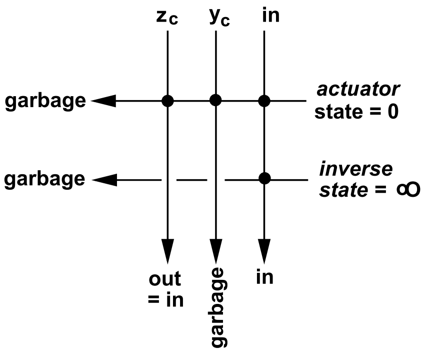
Computing with Solitons, Figure 7
COPY gate. Reprinted with permission from [34]. Copyright by the American Physical Society

cause there are four degrees of freedom in the two complex numbers z and y, and two complex equations to satisfy: that *out* be 1 and 0 when *in* is 1 and 0, respectively. Values that satisfy these four equations in four unknowns were obtained numerically. We will call them z_c and y_c. It is not always possible to solve these equations; Ablowitz et al. [1] showed that a unique solution is guaranteed to exist in certain parameter regimes. However, explicit solutions have been found for all the cases used in this section, and are given in Table 1.

To be more specific about the design problem, write Eq. (5) as the left-moving product $\rho_2 = L(\rho_1, \rho_L)$, and similarly write Eq. (7) as $\rho_R = R(\rho_1, \rho_L)$. The successive left-moving products in Fig. 7 are $L(in, 0)$ and $L(y, L(in, 0))$. The *out* state is then $R(z, L(y, L(in, 0)))$. The stipulation that 0 maps to 0 and 1 maps to 1 is expressed by the following two simultaneous complex equations in two complex unknowns

$$R(z, L(y, L(0, 0))) = 0 \,,$$
$$R(z, L(y, L(1, 0))) = 1 \,.$$

It is possible to solve for z as a function of y and then eliminate z from the equations, yielding one complex equation in the one complex unknown y. This is then solved numerically by grid search and successive refinement. There is no need for efficiency here, since we will require solutions in only a small number of cases.



Computing with Solitons, Figure 8
FANOUT gate. Reprinted with permission from [34]. Copyright by the American Physical Society

To make a FANOUT gate, we need to recover the input, which we can do using a collision with a soliton in the state which is the inverse of 0, namely ∞ [16]. Figure 8 shows the complete FANOUT gate. Notice that we indicate collisions with a dot at the intersection of paths, and require that the continuation of the inverse soliton not intersect the continuation of z that it meets. We indicate that by a broken line, and postpone the explanation of how this "wire crossing" is accomplished. It is immaterial whether the continuation of the inverse operator hits the continuation of y, because it is not used later. We call such solitons *garbage* solitons.

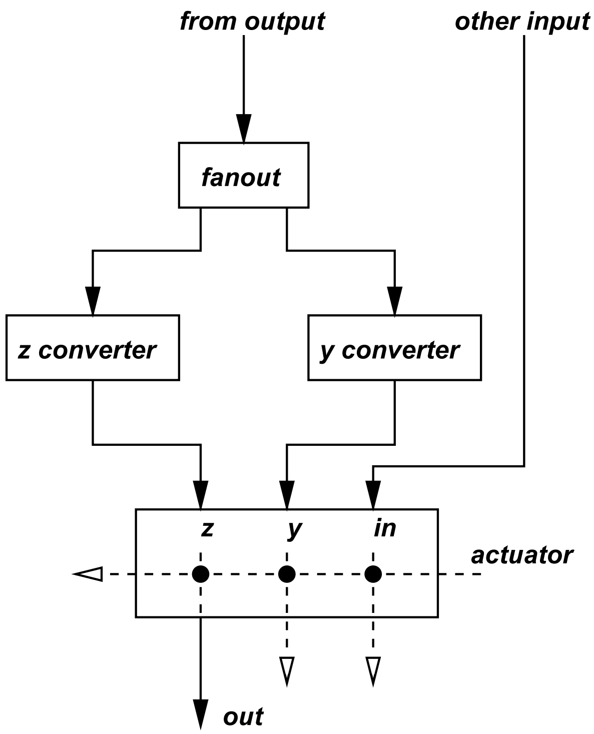
NOT and ONE Gates

In the same way we designed the complex pair of states (z_c, y_c) to produce a COPY and FANOUT gate, we can find a pair (z_n, y_n) to get a NOT gate, mapping 0 to 1 and 1 to 0; and a pair (z_1, y_1) to get a ONE gate, mapping both 0 and 1 to 1. These (z, y) values are given in Table 1.

We should point out that the ONE gate in itself, considered as a one-input, one-output gate, is not invertible, and could never be achieved by using the continuation of one particular soliton through one, or even many collisions. This is because such transformations are always nonsin-

Computing with Solitons, Table 1
Parameters for gates when soliton speeds are 1

gate	z	y
COPY	−0.24896731 − 0.62158212 · I	2.28774210 + 0.01318152 · I
NOT	−0.17620885 + 0.38170630 · I	0.07888703 − 1.26450654 · I
ONE	−0.45501471 − 1.37634227 · I	1.43987094 + 0.64061349 · I
Z-CONV	0.31838068 − 0.43078735 · I	−0.04232340 + 2.17536612 · I
Y-CONV	1.37286955 + 0.88495501 · I	−0.58835758 − 0.18026939 · I



Computing with Solitons, Figure 9
A NAND gate, using converter gates to couple copies of one of its inputs to its z and y parameters. Reprinted with permission from [34]. Copyright by the American Physical Society

gular linear fractional transformations, which are invertible [16]. The transformation of state from the input to the continuation of z is, however, much more complicated and provides the flexibility we need to get the ONE gate. It turns out that this ONE gate will give us a row in the truth table of a NAND, and is critical for realizing general logic.

Output/Input Converters, Two-Input Gates, and NAND

To perform logic of any generality we must of course be able to use the output of one operation as the input to another. To do this we need to convert logic (0/1) values to some predetermined z and y values, the choice depending on the type of gate we want. This results in a two-input, one-output gate.

As an important example, here's how a NAND gate can be constructed. We design a z-converter that converts 0/1 values to appropriate values of z, using the basic three-collision arrangement shown in Fig. 7. For a NAND gate, we map 0 to z_1, the z value for the ONE gate, and map 1 to z_n, the z value for the NOT gate. Similarly, we construct a y-converter that maps 0 to y_1 and 1 to y_n. These z- and y-converters are used on the fanout of one of the inputs, and the resulting two-input gate is shown in Fig. 9. Of course these z- and y-converters require z and y values themselves, which are again determined by numerical search (see Table 1).

The net effect is that when the left input is 0, the other input is mapped by a ONE gate, and when it is 1 the other input is mapped by a NOT gate. The only way the output can be 0 is if both inputs are 1, thus showing that this is a NAND gate. Another way of looking at this construction is that the 2×2 truth table of (left input)×(right input) has as its 0 row a ONE gate of the columns (1 1), and as its 1 row a NOT gate of the columns (1 0).

The importance of the NAND gate is that it is *universal* [20]. That is, it can be used with interconnects and fanouts to construct any other logical function. Thus we have shown that with the ability to "wire" we can implement any logic using the Manakov model.

We note that other choices of input converters result in direct realizations of other gates. Using input converters that convert 0 and 1 to (z_c, y_c) and (z_n, y_n), respectively, results in a truth table with first row (0 1) and second row (1 0), an XOR gate. Converting 0 and 1 to (z_c, y_c) and (z_1, y_1), respectively, results in an OR gate, and so on.

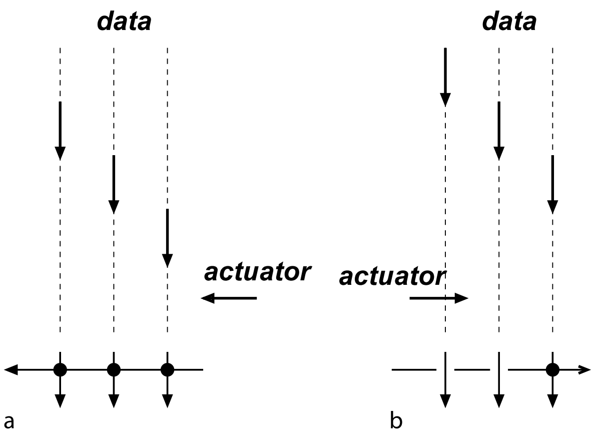
Time Gating

We next take up the question of interconnecting the gates described above, and begin by showing how the continuation of the input in the COPY gate can be restored without affecting the other signals. In other words, we show how a simple "wire crossing" can be accomplished in this case.

For spatial solitons, the key flexibility in the model is provided by assuming that input beams can be time-gated; that is, turned on and off. When a beam is thus gated, a finite segment of light is created that travels through the medium. We can think of these finite segments as finite light pulses, and we will call them simply *pulses* in the remainder of this article.

Figure 10a shows the basic three-collision gate implemented with pulses. Assuming that the actuator and data pulses are appropriately timed, the actuator pulse hits all three data pulses, as indicated in the projection below the space-space diagram. The problem is that if we want a later actuator pulse to hit the rightmost data pulse (to invert the state, for example, as in the FANOUT gate), it will also hit the remaining two data pulses because of the way they must be spaced for the earlier three collisions.

We can overcome this difficulty by sending the actuator pulse from the left instead of the right. Timing it appropriately early it can be made to miss the first two data



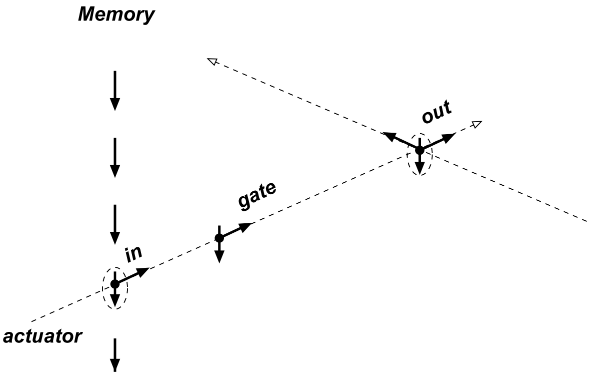
Computing with Solitons, Figure 10
a When entered from the right and properly timed, the actuator pulse hits all three data pulses, as indicated in the projection at the bottom; b When entered from the left and properly timed, the actuator pulse misses two data pulses and hits only the rightmost data pulse, as indicated in the projection at the bottom. Reprinted with permission from [34]. Copyright by the American Physical Society

pulses, and hit the third, as shown in Fig. 10b. It is easy to check that if the velocity of the right-moving actuator solitons is algebraically above that of the data solitons by the same amount that the velocity of the data solitons is algebraically above that of the left-moving actuator solitons, the same state transformations will result. For example, if we choose the velocities of the data and left-moving actuator solitons to be $+1$ and -1, we should choose the velocity of the right-moving actuator solitons to be $+3$. This is really a consequence of the fact that the g and h parameters of Eqs. (6) and (8) in the linear fractional transformation depend only on the difference in the velocities of the colliding solitons.

Wiring

Having shown that we can perform FANOUT and NAND, it remains only to show that we can "wire" gates so that any outputs can be fed to any inputs. The basic method for doing this is illustrated in Fig. 11. We think of data as stored in the down-moving pulses in a column, which we can think of as "memory". The observer moves with this frame, so the data appears stationary.

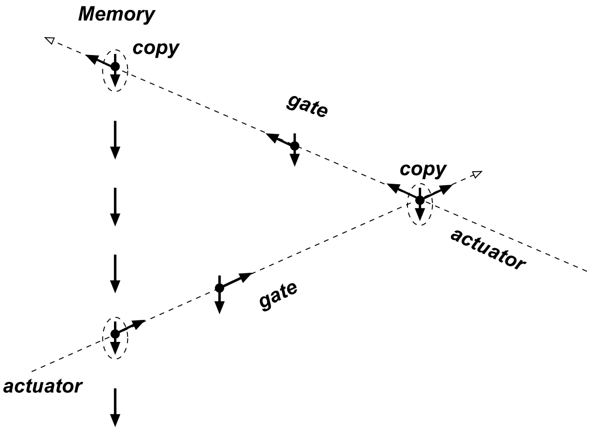
Pulses that are horizontal in the three-collision gates shown in previous figures will then appear to the observer to move upward at inclined angles. It is important to notice that these upward diagonally moving pulses are evanescent in our picture (and hence their paths are shown dashed in the figure). That is, once they are used, they do



Computing with Solitons, Figure 11
The frame of this figure is moving down with the data pulses on the left. A data pulse in memory is operated on with a three-collision gate actuated from the left, and the result deposited to the upper right. Reprinted with permission from [34]. Copyright by the American Physical Society

not remain in the picture with a moving frame and hence cannot interfere with later computations. However, all vertically moving pulses remain stationary in this picture.

Once a diagonal trajectory is used for a three-collision gate, reusing it will in general corrupt the states of all the stationary pulses along that diagonal. However, the original data pulse (gate input) can be restored with a pulse in the state inverse to the actuator, either along the same diagonal as the actuator, provided we allow enough time for the result (the gate output, a stationary z pulse) to be used, or along the other diagonal.



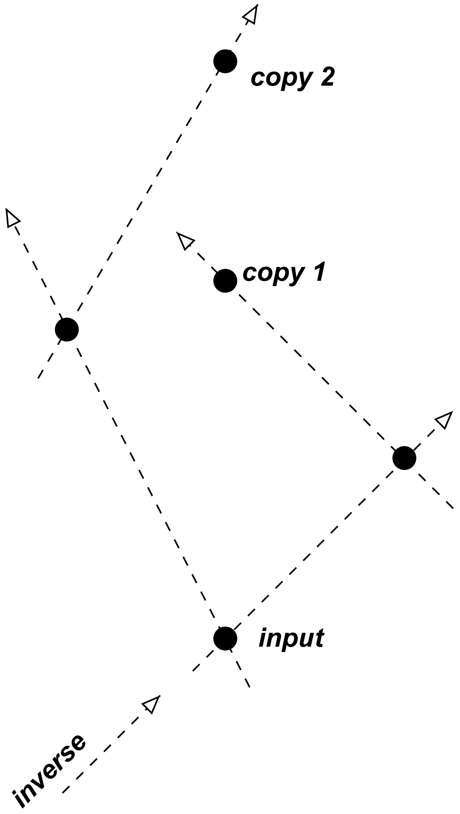
Computing with Solitons, Figure 12
A data pulse is copied to the upper right, this copy is copied to the upper left, and the result put at the top of memory. The original data pulse can then be restored with an inverse pulse and copied to the left in the same way. Reprinted with permission from [34]. Copyright by the American Physical Society

Suppose we want to start with a given data pulse in the memory column and create two copies above it in the memory column. Figure 12 shows a data pulse at the lower left being copied to the upper right with a three-collision COPY gate, initiated with an actuator pulse from the left. This copy is then copied again to the upper left, back to a waiting z pulse in the memory column. After the first copy is used, an inverse pulse can be used along the lower left to upper right diagonal to restore the original data pulse. The restored data pulse can then be copied to the left in the same way, to a height above the first copy, say, and thus two copies can be created and deposited in memory above the original.

A Second Speed and Final FANOUT and NAND

There is one problem still remaining with a true FANOUT: When an original data pulse in memory is used in a COPY operation for FANOUT, two diagonals are available, one from the lower left to the upper right, and the other from the lower right to the upper left. Thus, two copies can be made, as was just illustrated. However, when a data pulse is deposited in the memory column as a result of a logic operation, the logical operation itself uses at least one diagonal, which leaves at most one free. This makes a FANOUT of the *output* of a gate impossible with the current scheme.

A simple solution to this problem is to introduce another speed, using velocities ± 0.5, say, in addition to ± 1. This effectively provides four rather than two directions in which a pulse can be operated on, and allows true FANOUT and general interconnections. Figure 13 shows such a FANOUT; the data pulse at the lower left is copied to a position above it using one speed, and to another position, above that, using another.



Computing with Solitons, Figure 13
The introduction of a second speed makes true FANOUT possible. For simplicity, in this and the next figure, data and operator pulses are indicated by solid dots, and the y operator pulses are not shown. The paths of actuator pulses are indicated by dashed lines. Reprinted with permission from [34]. Copyright by the American Physical Society

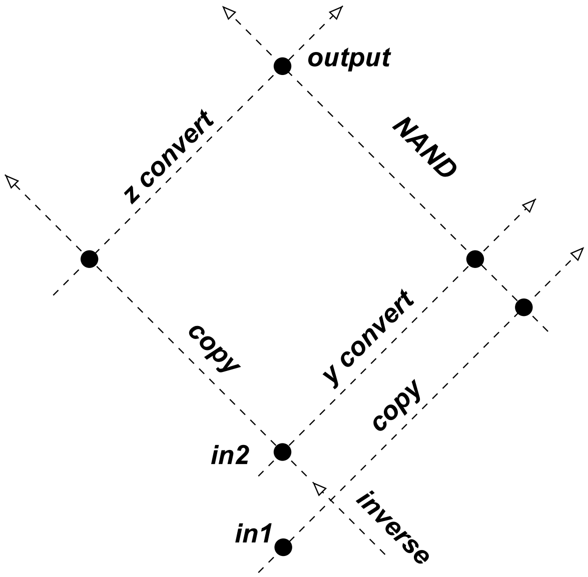
Finally, a complete NAND gate is shown in Fig. 14. The gate can be thought of as composed of the following steps:

- input 2 is copied to the upper left, and that copy transformed by a z-converter to the upper right, placing the z pulse for the NAND gate at the top of the figure;
- after the copy of input 2 is used, input 2 is restored with an inverse pulse to the upper left;
- input 2 is then transformed to the upper right by a y-converter;
- input 1 is copied to the upper right, to a position collinear with the z- and y-converted versions of the other input;
- a final actuator pulse converts the z pulse at the top to the output of the NAND gate.

Note that the output of the NAND has used two diagonals, which again shows why a second speed is needed if we are to use the NAND output as an input to subsequent logical operations. The y operator pulses, middle components in the three-collision COPY and converter gates, are not shown in the figure, but room can always be made for them to avoid accidental collisions by adding only a constant amount of space.

Universality

It should be clear now that any sequence of three-collision gates can be implemented in this way, copying data out of the memory column to the upper left or right, and performing NAND operations on any two at a time in the way shown in the previous section. The computation can proceed in a breadth-first manner, with the results of each successive stage being stored above the earlier results. Each additional gate can add only a constant amount of height



Computing with Solitons, Figure 14

Implementation of a NAND gate. A second speed will be necessary to use the output. Reprinted with permission from [34]. Copyright by the American Physical Society

and width to the medium, so the total area required is no more than proportional to the square of the number of gates.

The "program" consists of down-moving y and z operator pulses, entering at the top with the down-moving data, and actuator pulses that enter from the left or right at two different speeds. In the frame moving with the data, the data and operator pulses are stationary and new results are deposited at the top of the memory column. In the laboratory frame the data pulses leave the medium downward, and new results appear in the medium at positions above the old data, at the positions of newly entering z pulses.

Discussion

We have shown that in principle any computation can be performed by shining time-gated lasers into a completely homogeneous nonlinear optical medium. This result should be viewed as mathematical, and whether the physics of vector soliton collisions can lead to practical computational devices is a subject for future study. With regard to the economy of the model, the question of whether time gating is necessary, or even whether two speeds are necessary, is open.

We note that the result described here differs from the universality results for the ideal billiard ball model [13], the Game of Life [7], and Lattice Gasses [32], for example, in that no internal mirrors or structures of any kind are used inside the medium. To the author's knowledge, whether internal structure is necessary in these other cases is open.

Finally, we remark that the model used is reversible and dissipationless. The fact that some of the gate operations realized are not in themselves reversible is not a contradiction, since extra, "garbage" solitons [13] are produced that save enough state to run the computation backwards.

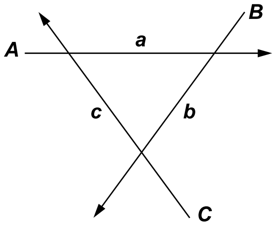
Multistable Soliton Collision Cycles

Bistable and multistable optical systems, besides being of some theoretical interest, are of practical importance in offering a natural "flip-flop" for noise immune storage and logic. We show in this section that simple cycles of collisions of solitons governed by the Manakov equations can have more than one distinct stable set of polarization states, and therefore these distinct equilibria can, in theory, be used to store and process information. The multistability occurs in the polarization states of the beams; the solitons themselves do not change shape and remain the usual sech-shaped solutions of the Manakov equations. This phenomenon is dependent only on simple soliton collisions in a completely homogeneous medium.

The basic configuration considered requires only that the beams form a closed cycle, and can thus be realized in any nonlinear optical medium that supports spatial Manakov solitons. The possibility of using multistable systems of beam collisions broadens the possibilities for practical application of the surprisingly strong interactions that Manakov solitons can exhibit, a phenomenon originally described in [26]. We show here by example that a cycle of three collisions can have two distinct foci surrounded by basins of attractions, and that a cycle of four collisions can have three.

The Basic Three-Cycle and Computational Experiments

Figure 15 shows the simplest example of the basic scheme, a cycle of three beams, entering in states A, B, and C, with intermediate beams a, b, and c. For convenience, we will refer to the beams themselves, as well as their states, as A, B, C, etc. Suppose we start with beam C initially turned off, so that $A = a$. Beam a then hits B, thereby transforming it to state b. If beam C is then turned on, it will hit A, closing the cycle. Beam a is then changed, changing b, etc., and the cycle of state changes propagates clockwise. The question we ask is whether this cycle converges, and if so, whether



Computing with Solitons, Figure 15
The basic cycle of three collisions. Reprinted with permission from [35]. Copyright by the American Physical Society

it will converge with any particular choice of complex parameters to exactly zero, one, two, or more foci. We answer the question with numerical simulations of this cycle.

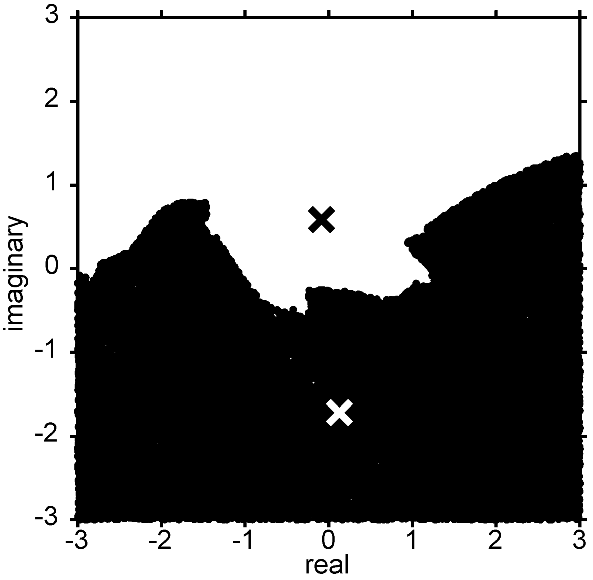
A typical computational experiment was designed by fixing the input beams *A*, *B*, *C*, and the parameters k_1 and k_2, and then choosing points *a* randomly and independently with real and imaginary coordinates uniformly distributed in squares of a given size in the complex plane. The cycle described above was then carried out until convergence in the complex numbers *a*, *b*, and *c* was obtained to within 10^{-12} in norm. Distinct foci of convergence were stored and the initial starting points *a* were categorized by which focus they converged to, thus generating the usual picture of basins of attraction for the parameter *a*. Typically this was done for 50,000 random initial values of *a*, effectively filling in the square, for a variety of parameter choices *A*, *B*, and *C*. The following results were observed:

- In cases with one or two clear foci, convergence was obtained in every iteration, almost always within one or two hundred iterations.
- Each experiment yielded exactly one or two foci.
- The bistable cases (two foci) are somewhat less common than the cases with a unique focus, and are characterized by values of k_R between about 3 and 5 when the velocity difference Δ was fixed at 2.

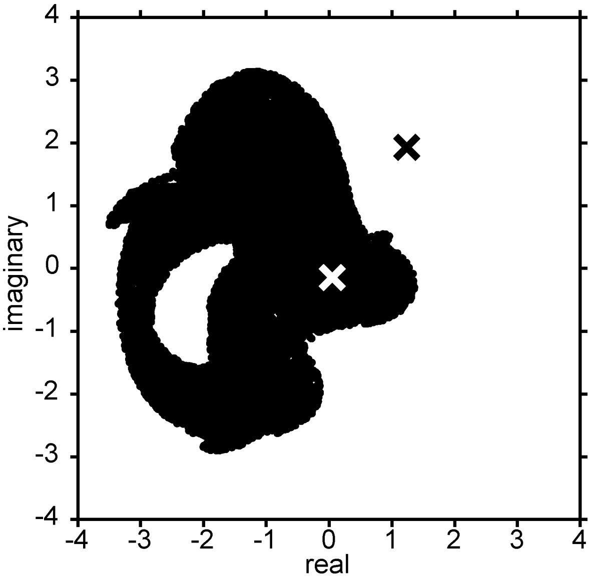
Figure 16 shows a bistable example, with the two foci and their corresponding basins of attraction. The parameter *k* is fixed in this and all subsequent examples at 4 ± *i* for the right- and left-moving beams of any given collision, respectively. The second example, shown in Fig. 17, shows that the basins are not always simply connected; a sizable island that maps to the upper focus appears within the basin of the lower focus.

A Tristable Example Using a Four-Cycle

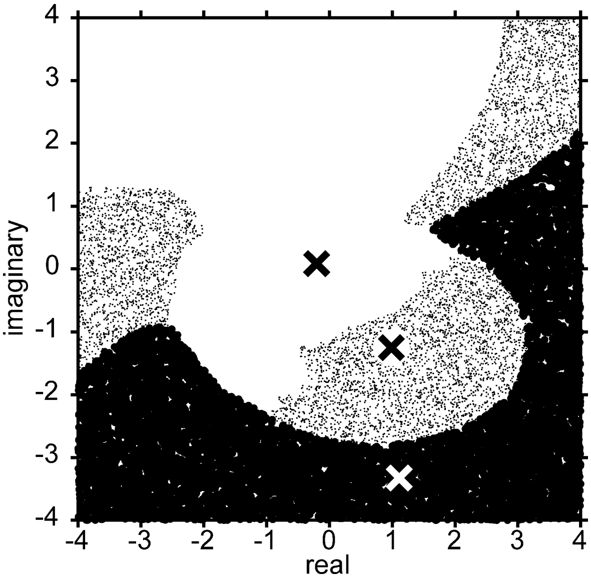
Collision cycles of length four seem to exhibit more complex behavior than those of length three, although it is dif-



Computing with Solitons, Figure 16
The two foci and their corresponding basins of attraction in the first example, which uses a cycle of three collisions. The states of the input beams are $A = -0.8 - i \cdot 0.13$, $B = 0.4 - i \cdot 0.13$, $C = 0.5 + i \cdot 1.6$; and $k = 4 \pm i$. Reprinted with permission from [35]. Copyright by the American Physical Society



Computing with Solitons, Figure 17
A second example using a cycle of three collisions, showing that the basins need not be simply connected. The states of the input beams are $A = 0.7 - i \cdot 0.3$, $B = -1.1 - i \cdot 0.5$, $C = 0.4 + i \cdot 0.81$; and $k = 4 \pm i$. Reprinted with permission from [35]. Copyright by the American Physical Society



Computing with Solitons, Figure 18
A case with three stable foci, for a collision cycle of length four. The states of the input beams are $A = -0.39 - i \cdot 0.45$, $B = 0.22 - i \cdot 0.25$, $C = 0.0 + i \cdot 0.25$, $D = -0.51 + i \cdot 0.48$; and $k = 4 \pm i$. Reprinted with permission from [35]. Copyright by the American Physical Society

ficult to draw any definite conclusions because the parameter spaces are too large to be explored exhaustively, and there is at present no theory to predict such highly nonlinear behavior. If one real degree of freedom is varied as a control parameter, we can move from bistable to tristable solutions, with a regime between in which one basin of attraction disintegrates into many small separated fragments. Clearly, this model is complex enough to exhibit many of the well-known features of nonlinear systems.

Fortunately, it is not difficult to find choices of parameters that result in very well behaved multistable solutions. For example, Fig. 18 shows such a tristable case. The smallest distance from a focus to a neighboring basin is on the order of 25% of the interfocus distance, indicating that these equilibria will be stable under reasonable noise perturbations.

Discussion

The general phenomenon discussed in this section raises many questions, both of a theoretical and practical nature. The fact that there are simple polarization-multistable cycles of collisions in a Manakov system suggests that similar situations occur in other vector systems, such as photorefractive crystals or birefringent fiber. Any vector system with the possibility of a closed cycle of soliton collisions becomes a candidate for multistability, and there is at this point really no compelling reason to restrict attention to the Manakov case, except for the fact that the explicit state-change relations make numerical study much easier.

The simplified picture we used of information traveling clockwise after we begin with a given beam *a* gives us stable polarization states when it converges, plus an idea of the size of their basins of attractions. It is remarkable that in all cases in our computational experience, except for borderline transitional cases in going from two to three foci in a four-cycle, this circular process converges consistently and quickly. But understanding the actual dynamics and convergence characteristics in a real material requires careful physical modeling. This modeling will depend on the nature of the medium used to approximate the Manakov system, and is left for future work. The implementation of a practical way to switch from one stable state to another is likewise critically dependent on the dynamics of soliton formation and perturbation in the particular material at hand, and must be studied with reference to a particular physical realization.

We remark also that no iron-clad conclusions can be drawn from computational experiments about the numbers of foci in any particular case, or the number possible for a given size cycle—despite the fact that we regularly used 50,000 random starting points. On the other hand, the clear cases that have been found, such as those used as examples, are very characteristic of universal behavior in other nonlinear iterated maps, and are sufficient to establish that bi- and tristability, and perhaps higher-mode multistability, is a genuine mathematical characteristic, and possibly also physically realizable. It strongly suggests experimental exploration.

We restricted discussion in this section to the simplest possible structure of a single closed cycle, with three or four collisions. The stable solutions of more complicated configurations are the subject of continuing study. A general theory that predicts this behavior is lacking, and it seems at this point unlikely to be forthcoming. This forces us to rely on numerical studies, from which, as we point out above, only certain kinds of conclusions can be drawn. We are fortunate, however, in being able to find cases that look familiar and which are potentially useful, like the bistable three-cycles with well separated foci and simply connected basins of attraction.

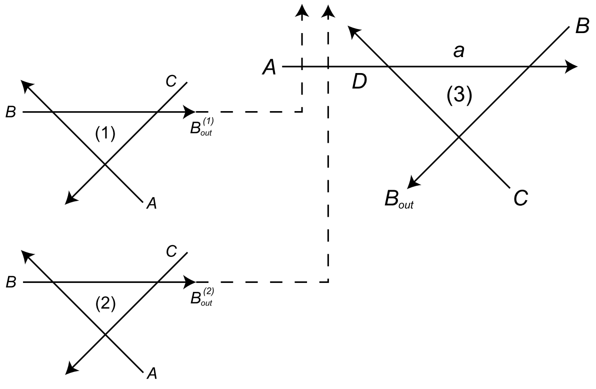
It is not clear however, just what algorithms might be used to find equilibria in collision topologies with more than one cycle. It is also intriguing to speculate about how collision configurations with particular characteristics can be designed, how they can be made to interact, and how they might be controlled by pulsed beams. There is promise that when the ramifications of complexes of vector soliton collisions are more fully understood they might be useful for real computation in certain situations.

Application to Noise-Immune Soliton Computing

Any physical instantiation of a computing technology must be designed to be immune from the effects of noise buildup from logic stage to logic stage. In the familiar computers of today, built with solid-state transistors, the noise-immunity is provided by physical state restoration, so that voltage levels representing logical "0" and "1" are restored by bistable circuit mechanisms at successive logic stages. This is state restoration at the physical level.

For another example, proposed schemes for quantum computing would be impractical without some means of protecting information stored in qubits from inevitable corruption by the rest of the world. The most common method proposed for accomplishing this is error correction at the software level, state restoration at the logical level.

In the collision-based scheme for computing with Manakov solitons described in Sect. "Manakov Soliton Computing", there is no protection against buildup of error from stage to stage, and some sort of logical state-restoration would be necessary in a practical realization. The bistable collision cycles of Manakov solitons described in this section, however, offer a natural computational building block for soliton computation with physical state restoration. This idea is explored in [27]. Figure 19 illustrates the approach with a schematic diagram of a NAND gate, implemented with bistable cycles to represent bits. The input bits are stored in the collision cycles (1) and (2), which have output beams that can be made to collide with input beam *A* of cycle (3), which represents the output bit of the gate. These inputs to the gate, shown as dashed lines, change the state of beam *A* of the ordinarily bistable cycle (3) so that it becomes *monostable*. The state of cycle (3) is then steered to a known state. When the input beams are turned off, cycle (3) returns to its normal bistable condition, but with a known input state. Its state then evolves to one of two bits, and the whole system of three collision cycles can be engineered so that the final state of cycle (3) is the NAND of the two bits represented by input cycles (1) and (2). (See [27] for details.)



Computing with Solitons, Figure 19
Schematic of NAND gate using bistable collision cycles. Reprinted with permission from [27]. Copyright by Old City Publishing

A computer based on such bistable collision cycles is closer in spirit to present-day ordinary transistor-based computers, with a natural noise-immunity and state-restoration based on physical bistability. As mentioned in the previous subsection, however, the basic bistable cycle phenomenon awaits laboratory verification, and much remains to be learned about the dynamics, and eventual speed and reliability of such systems.

Experiments

The computation schemes described in the previous sections obviously rely on the correct mathematical modeling of the physics proposed for realization. We next describe experiments that verify some of the required soliton phenomenology in optical fibers. Specifically, we highlight the experimental observation of temporal vector soliton propagation and collision in a birefringent optical fiber [28]. This is both the first demonstration of temporal vector solitons with two mutually-incoherent component fields, and of vector soliton collisions in a Kerr nonlinear medium.

Temporal soliton pulses in optical fiber were first predicted by Hasegawa and Tappert [14], followed by the first experimental observation by Mollenauer et al. [24]. In subsequent work, Menyuk accounted for the birefringence in polarization maintaining fiber (PMF) and predicted that vector solitons, in which two orthogonally polarized components trap each other, are stable under the proper operating conditions [21,22]. For birefringent fibers, self-trapping of two orthogonally polarized pulses can occur when XPM-induced nonlinearity compensates the birefringence-induced group velocity difference, causing the pulse in the fiber's fast axis to slow down and the pulse in the slow axis to speed up. The first demonstration of temporal soliton trapping was performed in the sub picosecond regime [15], in which additional ultrashort pulse effects such as Raman scattering are present. In particular, this effect results in a red-shift that is linearly proportional to the propagation distance, as observed in a later temporal soliton trapping experiment [25]. Recently, soliton trapping in the picosecond regime was observed with equal amplitude pulses [18]. However, vector soliton propagation could not be shown, because the pulses propagated for less than 1.5 dispersion lengths. In other work, phase-locked vector solitons in a weakly birefringent fiber laser cavity with nonlinear coherent coupling between components were observed [12].

The theoretical model for linearly birefringent fiber is the following coupled nonlinear Schrödinger equation (CNLSE):

$$i\left(\frac{\partial A_x}{\partial z}+\beta_{1x}\frac{\partial A_x}{\partial t}\right)-\frac{\beta_2}{2}\frac{\partial^2 A_x}{\partial t^2}+\gamma\left(|A_x|^2+\frac{2}{3}|A_y|^2\right)A_x = 0\,,$$

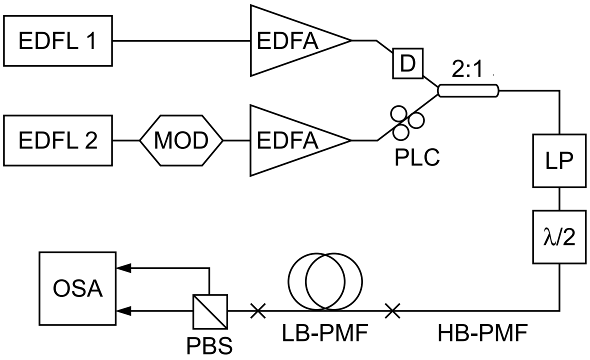
$$i\left(\frac{\partial A_y}{\partial z}+\beta_{1y}\frac{\partial A_y}{\partial t}\right)-\frac{\beta_2}{2}\frac{\partial^2 A_y}{\partial t^2}+\gamma\left(|A_y|^2+\frac{2}{3}|A_x|^2\right)A_y = 0\,, \quad (9)$$

where t is the local time of the pulse, z is propagation distance along the fiber, and $A_{x,y}$ is the slowly varying pulse envelope for each polarization component. The parameter $\beta_{1x,y}$ is the group velocity associated with each fiber axis, and β_2 represents the group velocity dispersion, assumed equal for both polarizations. In addition, we neglect higher order dispersion and assume a loss less medium with an instantaneous electronic response, valid for picosecond pulses propagating in optical fiber.

The last two terms of Eqs. (9) account for the nonlinearity due to SPM and XPM, respectively. In linearly birefringent optical fiber, a ratio of 2/3 exists between these two terms. When this ratio equals unity, the CNLSE becomes the integrable Manakov system of Eqs. (4). On the other hand, solutions of Eqs. (9) are, strictly speaking, solitary waves, not solitons. However, it was found in [36] that the family of symmetric, single-humped (fundamental or first-order) solutions, to which the current investigation in this section belongs, are all stable. Higher-order solitons, characterized by multiple humps, are unstable. Furthermore, it was shown in [37] that collisions of solitary waves in Eqs. (9) can be described by application of perturbation theory to the integrable Manakov equations, indicating the similarities between the characteristics of these two systems.

Experimental Setup and Design

The experimental setup is shown in Fig. 20. We synchronized two actively mode-locked erbium-doped fiber lasers (EDFLs)—EDFL1 at 1.25 GHz repetition rate, and EDFL2 at 5 GHz. EDFL2 was modulated to match with the lower repetition rate of EDFL1. Each pulse train, consisting of 2 ps pulses, was amplified in an erbium-doped fiber amplifier (EDFA) and combined in a fiber coupler. To align polarizations, a polarization loop controller (PLC) was used in one arm, and a tunable delay line (TDL) was employed to temporally align the pulses for collision. Once combined, both pulse trains passed through a linear polarizer (LP) and a half-wave plate to control the input polarization to the PMF. Approximately 2 m of high birefringence (HB) PMF preceded the specially designed 500 m of low birefringence (LB) PMF used to propagate vector solitons. Although this short length of HB-PMF will introduce some pulse splitting (on the order of 2–3 ps), the birefringent axes of the HB- and LB-PMF were swapped in order to counteract this effect. At the output, each component of the vector soliton was then split at a polarization beam splitter, followed by an optical spectrum analyzer (OSA) for measurement.



Computing with Solitons, Figure 20
Experimental setup. EDFL: Erbium-doped fiber laser; EDFA: Erbium-doped fiber amplifier; MOD: modulator; D: tunable delay line; PLC: polarization loop controller; 2:1: fiber coupler; LP: linear polarizer; $\lambda/2$: half-wave plate; HB-PMF and LB-PMF: high and low birefringence polarization maintaining fiber; PBS: polarization beam splitter; OSA: optical spectrum analyzer. Reprinted with permission from [28]. Copyright by the American Physical Society

The design of the LB-PMF required careful control over three characteristic length scales: the (polarization) beat length, dispersion length L_{d}, and nonlinear length L_{nl}. A beat length $L_{\mathrm{b}} = \lambda/\Delta n = 50$ cm was chosen at a wavelength of 1550 nm, where Δn is the fiber birefringence. According to the approximate stability criterion of [8], this choice allows stable propagation of picosecond vector solitons. By avoiding the sub picosecond regime, ultrashort pulse effects such as intrapulse Raman scattering will not be present. The dispersion $D = 2\pi c\beta_2/\lambda^2 =$ 16 ps/km nm and $L_{\mathrm{d}} = 2T_0^2/|\beta_2| \approx 70$ m, where $T_0 = T_{\mathrm{FWHM}}/1.763$ is a characteristic pulse width related to the full width at half maximum (FWHM) pulse width. Since $L_d \gg L_b$, degenerate four-wave mixing due to coherent coupling between the two polarization components can be neglected [23]. Furthermore, the total propagation distance is greater than 7 dispersion lengths.

Polarization instability, in which the fast axis component is unstable, occurs when $L_{nl} = (\gamma P)^{-1}$ is of the same order of magnitude or smaller than L_b, as observed in [6]. The nonlinearity parameter $\gamma = 2\pi n_2/\lambda A_{eff} = 1.3$ (km W)$^{-1}$, with Kerr nonlinearity coefficient $n_2 = 2.6 \times 10^{-20}$ m^2/W and measured effective mode area $A_{eff} = 83\,\mu m^2$. In the LB-PMF, the fundamental vector soliton power $P \approx 14$ W, thus $L_{nl} = 55$ m $\gg L_b$, mitigating the effect of polarization instability.

Vector Soliton Propagation

We first studied propagation of vector solitons using both lasers independently. The wavelength shift for each component is shown in Fig. 21a as a function of the input polarization angle ϕ, controlled through the half-wave plate. Due to the anomalous dispersion of the fiber at this wavelength, the component in the slow (fast) axis will shift to shorter (longer) wavelengths to compensate the birefringence. The total amount of wavelength shift between components $\Delta\lambda_{xy} = \Delta\beta_1/D = 0.64$ nm, where $\Delta\beta_1 = |\beta_{1x} - \beta_{1y}| = 10.3$ ps/km is the birefringence-induced group velocity difference and dispersion $D = 2\pi c\beta_2/\lambda^2 = 16$ ps/km nm.

As ϕ approaches 0° (90°), the vector soliton approaches the scalar soliton limit, and the fast (slow) axis does not shift in wavelength, as expected. At $\phi = 45°$, a symmetric shift results. For unequal amplitude solitons, the smaller component shifts more in wavelength than the larger component, because the former experiences more XPM. Numerical simulations of Eqs. (9), given by the dashed lines of Fig. 21a, agree very well with the experimental results. Also shown in Fig. 21 are two cases, $\phi = 45°$ and 37°, as well as the numerical prediction. The experimental spectra show some oscillatory features at 5 GHz, which are a modulation of the EDFL2 repetition rate on the optical spectrum. A sample input pulse spectrum from EDFL1 is shown in the inset of Fig. 21, which shows no modulation due to the limited resolution of the OSA. Vector solitons from both lasers produced similar results. In this and all subsequent plots in this section, the slow and fast axis components are depicted by solid and dashed lines, respectively.

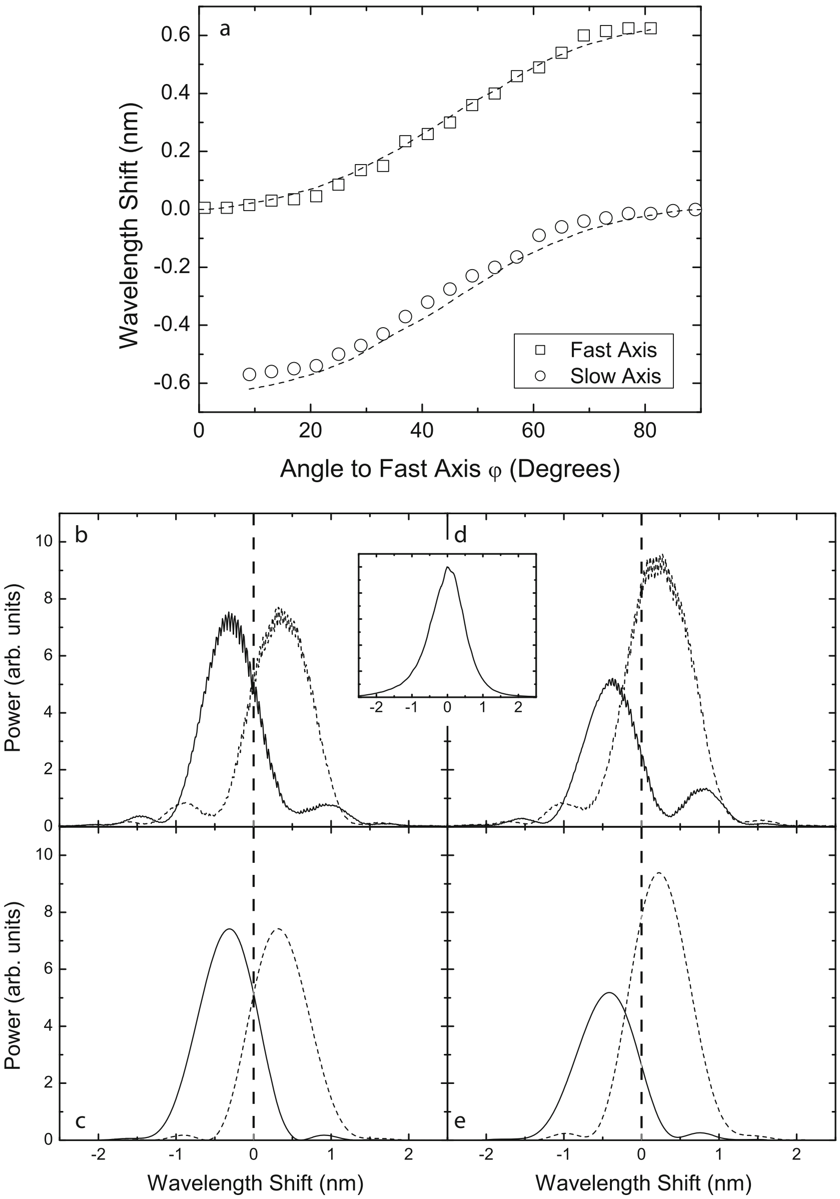
As the two component amplitudes become more unequal, satellite peaks become more pronounced in the smaller component. These features are also present in the simulations, but are not as dominant (cf. Fig. 21d and e). We attribute this to the input pulse, which is calibrated for the $\phi = 45°$ case, because the power threshold for vector soliton formation in this case is largest due to the 2/3 factor between SPM and XPM nonlinear terms in the CNLSE. As the input is rotated towards unequal components, there will be extra power in the input pulse, which will radiate in the form of dispersive waves as the vector soliton forms. Due to the nature of this system, these dispersive waves can be nonlinearly trapped, giving rise to the satellite features in the optical spectra. This effect is not as prevalent in the simulations because the threshold was numerically determined at each input angle ϕ.

Vector Soliton Collision

To prepare the experiment for a collision, we operated both lasers simultaneously, detuned in wavelength to allow for dispersion-induced walkoff, and adjusted the delay line in such a way that the collision occurred halfway down the fiber. We define a collision length $L_{coll} = 2T_{FWHM}/D\Delta\lambda$, where $\Delta\lambda$ is the wavelength separation between the two vector solitons. For our setup, $\Delta\lambda = 3$ nm, and $L_{coll} = 83.3$ m. An asymptotic theory of soliton collisions, in which a full collision takes place, requires at least 5 collision lengths. The total fiber length in this experiment is equal to 6 collision lengths, long enough to ensure sufficient separation of solitons before and after collision. In this way, results of our experiments can be compared to the asymptotic theory, even though full numerical simulations will be shown for comparison. To quantify our results, we introduce a quantity $R \equiv \tan^2\phi$, defined as the amplitude ratio between the slow and fast components.

Recall that in Sect. "Manakov Solitons", we introduced the Manakov equations (Eqs. (4)), and described collision-induced transformations of the polarization state of the soliton, which come about due to the asymptotic analysis of the soliton collision. The polarization state is the ratio between the two components $\rho \equiv A_x/A_y = \cot\phi \exp(i\Delta\theta)$, and is therefore a function of the polarization angle ϕ and the relative phase $\Delta\theta$ between the two components. In the context of the experiments described in this section, these state transformations (Eqs. (5) and (7)) predict that the resulting energy exchange will be a function of amplitude ratios $R_{1,2}$, wavelength separation $\Delta\lambda$, and the relative phase $\Delta\theta_{1,2}$ between the two components of each soliton, where soliton 1 (2) is the shorter (longer) wavelength soliton.

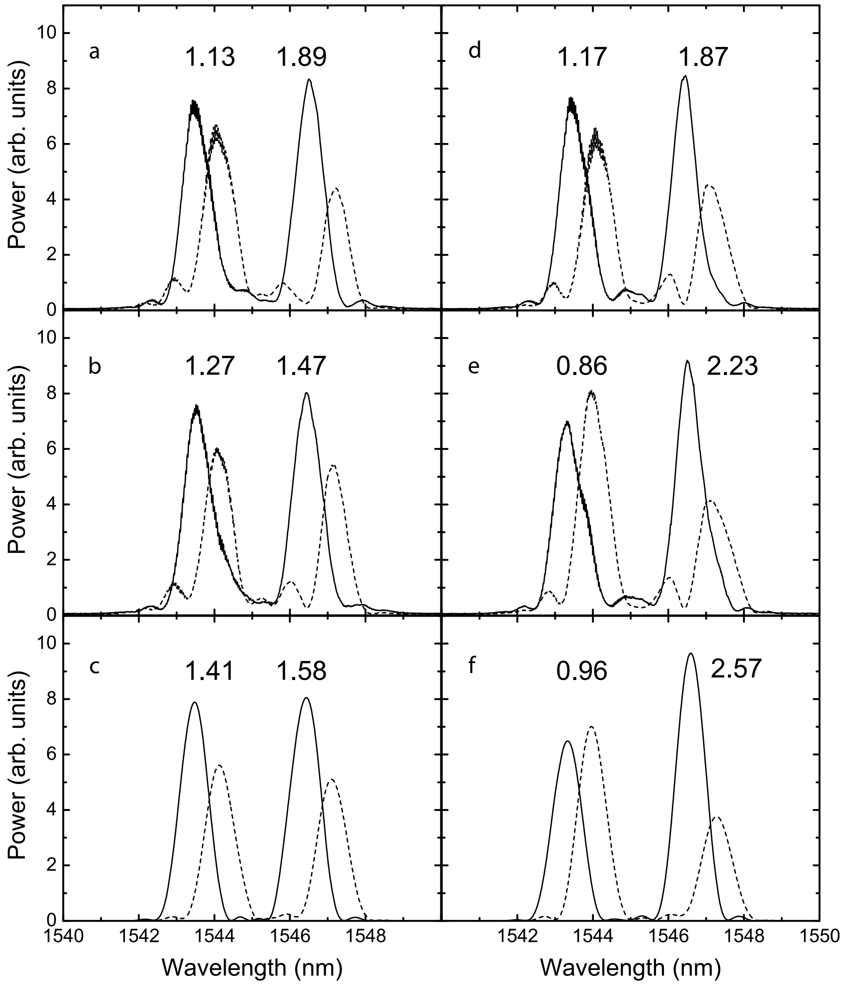
A word of caution is in order at this point. An interesting consequence of the 2/3 ratio between SPM and XPM, which sets the birefringent fiber model apart from



Computing with Solitons, Figure 21

Arbitrary-amplitude vector soliton propagation. a Wavelength shift vs. angle to fast axis ϕ, numerical curves given by dashed lines; b and d experimental results for $\phi = 45°$ and 37° with EDFL2, respectively. *Inset:* input spectrum for EDFL1; c and e corresponding numerical simulations of $\phi = 45°$ and 37°, respectively. The slow and fast axis components are depicted by solid and dashed lines, respectively. Reprinted with permission from [28]. Copyright by the American Physical Society

the Manakov model, is the relative phase between the two components. For the Manakov soliton, each component 'feels' the same amount of total nonlinearity, because the strengths of both SPM and XPM are equal. Therefore, regardless of the polarization angle, the amount of total nonlinear phase shift for each component is the same (even though the contributions of SPM and XPM phase shifts are in general not equal). As a result, the relative phase between the two components stays constant during propagation, as does the polarization state. This is *not* the case for vector solitons in birefringent fiber. For the case of equal amplitudes, each component does experience the same amount of nonlinear phase shift, and therefore the polarization state is constant as a function of propagation



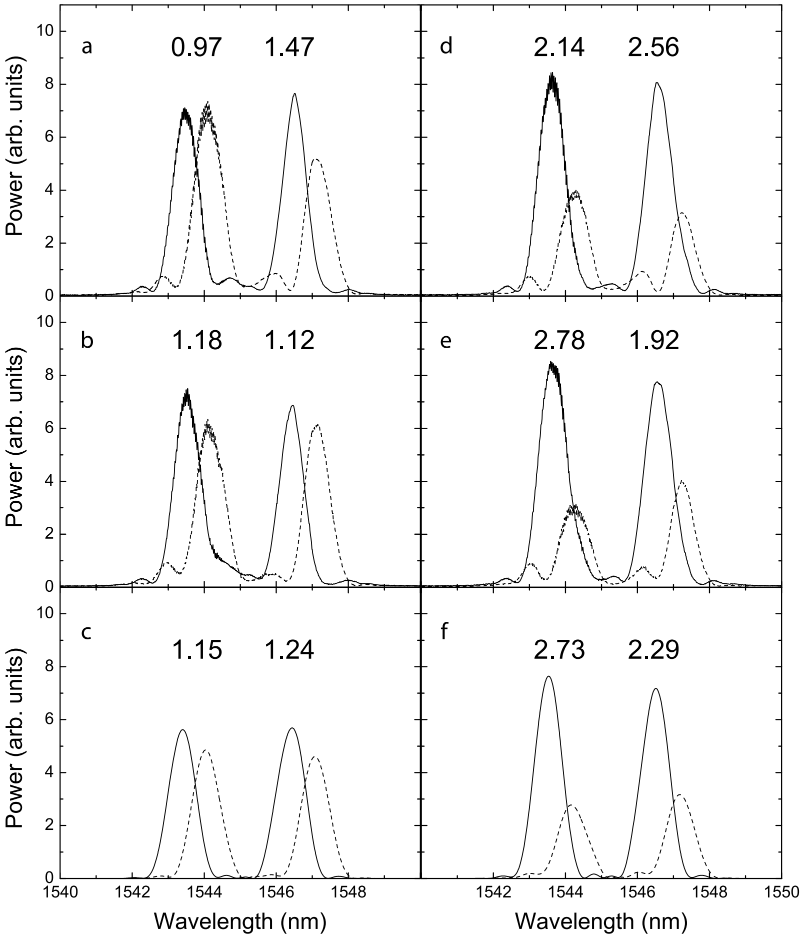
Computing with Solitons, Figure 22

Demonstration of phase-dependent energy-exchanging collisions. a–c Short HB-PMF; d–f long HB-PMF; a,d experiment, without collision; b, e experiment, with collision; c,f simulated collision result with c $\Delta\theta_2 = 90°$ and f $\Delta\theta_2 = 50°$. Values of slow-fast amplitude ratio *R* are given above each soliton. The slow and fast axis components are depicted by solid and dashed lines, respectively. Reprinted with permission from [28]. Copyright by the American Physical Society

distance. However, for arbitrary (unequal) amplitudes, the total phase shift for each component will be different. Consequently, the relative phase will change *linearly* as a function of propagation distance, and the polarization state will not be constant. As a result, the collision-induced change in polarization state, while being a function of the amplitude ratios $R_{1,2}$ and wavelength separation $\Delta\lambda$, will also depend upon the collision position due to the propagation dependence of the relative phase $\Delta\theta_{1,2}(z)$. To bypass this complication, we ensure that all collisions occur at the same spatial point in the fiber.

Because only one half-wave plate is used in our experiment (see Fig. 20), it was not possible to prepare each vector soliton individually with an arbitrary *R*. In addition, due to the wavelength dependence of the half-wave plate, it was not possible to adjust $\Delta\lambda$ without affecting *R*.

First, we investigated the phase dependence of the collision. This was done by changing the length of the HB-PMF entering the LB-PMF, while keeping *R* and $\Delta\lambda$ constant. As a result, we could change $\Delta\theta_{1,2}$ due to the birefringence of the HB-PMF. Approximately 0.5 m of HB-PMF was added to ensure that the total amount of temporal pulse splitting did not affect the vector soliton formation. The results are shown in Fig. 22, where Fig. 22a–c and d–f correspond to the short and long HB-PMFs, respectively. Figure 22a and d show the two vector solitons, which propagate independently when no collision occurs; as expected, the two results are similar because the OSA



Computing with Solitons, Figure 23

Additional energy-exchanging collisions. a,d Experiment, without collision; b,e experiment, with collision; c,f simulated collision result, using $\Delta\theta_2 = 90°$ inferred from the experiment of Fig. 22. Values of slow-fast amplitude ratio *R* are given above each soliton. The slow and fast axis components are depicted by solid and dashed lines, respectively. Reprinted with permission from [28]. Copyright by the American Physical Society

measurement does not depend on $\Delta\theta_{1,2}$. The result of the collision is depicted in Fig. 22b and e, along with the corresponding simulation results in Fig. 22c and f.

In both of these collisions, an energy exchange between components occurs, and two important relations are satisfied: the total energy in each soliton and in each component is conserved. It can be seen that when one component in a soliton increases as a result of the collision, the other component decreases, with the opposite exchange in the second soliton. The difference between these two collisions is dramatic, in that the energy redistributes in opposite directions. For the simulations, idealized sech pulses for each component were used as initial conditions, and propagation was modeled without accounting for losses. The experimental amplitude ratio was used, and (without loss of generality [16,19,26]) $\Delta\theta_2$ was varied while $\Delta\theta_1 = 0$. Best fits gave $\Delta\theta_2 = 90°$ (Fig. 22c) and 50° (Fig. 22f). Despite the model approximations, experimental and numerical results all agree to within 15%.

In the second set of results (Fig. 23), we changed R while keeping all other parameters constant. More specifically, we used the short HB-PMF, with initial phase difference $\Delta\theta_2 = 90°$, and changed the amplitude ratio. In agreement with theoretical predictions, the same direction of energy exchange is observed as in Fig. 22a–c.

Spatial Soliton Collisions

We mention here analogous experiments with spatial solitons in photorefractive media by Anastassiou et al. In [4], it is shown that energy is transferred in a collision of vector spatial solitons in a way consistent with the predictions for the Manakov system (although the medium is a saturable one, and only approximates the Kerr nonlinearity). The experiment in [5] goes one step farther, showing that one soliton can be used as an intermediary to transfer energy from a second soliton to a third. We thus are now at a point where the ability of both temporal and spatial vector solitons to process information for computation has been demonstrated.

Future Directions

This article discussed computing with solitons, and attempted to address the subject from basic physical principles to applications. Although the nonlinearity of fibers is very weak, the ultralow loss and tight modal confinement make them technologically attractive. By no means, however, are they the only potential material for soliton-based information processing. Others include photorefractive crystals, semiconductor waveguides, quadratic media, and Bose–Einstein condensates, while future materials research may provide new candidate systems.

From a computing perspective, scalar soliton collisions are insufficient. Although measurable phase and position shifts do occur, these phenomena cannot be cascaded to affect future soliton collisions and therefore cannot transfer information from one collision to the next. Meaningful computation using soliton collisions requires a new degree of freedom; that is, a new component. Collisions of vector solitons display interesting energy-exchanging effects between components, which can be exploited for arbitrary computation and bistability.

The vector soliton experiments of Sect. "Experiments" were proof-of-principle ones. The first follow-up experiments with temporal vector solitons in birefringent fiber can be directed towards a full characterization of the collision process. This can be done fairly simply using the experimental setup of Fig. 20 updated in such a way as to allow independent control of two vector soliton inputs. This would involve separate polarizers and half-waveplates, followed by a polarization preserving fiber coupler.

Cascaded collisions of temporal solitons also await experimental study. As demonstrated in photorefractive crystals with a saturable nonlinearity [5], one can show that information can be passed from one collision to the next. Beyond a first demonstration of two collisions is the prospect of setting up a multi-collision feedback cycle. Discussed in Sect. "Multistable Soliton Collision Cycles", these collision cycles can be bistable and lead to interesting applications in computation.

Furthermore, the recent work of Ablowitz et al. [2] shows theoretically that the useful energy-redistribution properties of vector soliton collisions extend perfectly to the semi-discrete case: that is, to the case where space is discretized, but time remains continuous. This models, for example, propagation in an array of coupled nonlinear waveguides [10]. The work suggests alternative physical implementations for soliton switching or computing, and also hints that the phenomenon of soliton information processing is a very general one.

Bibliography

1. Ablowitz MJ, Prinari B, Trubatch AD (2004) Soliton interactions in the vector nls equation. Inverse Problems 20(4):1217–1237
2. Ablowitz MJ, Prinari B, Trubatch AD (2006) Discrete vector solitons: Composite solitons, Yang–Baxter maps and computation. Studies in Appl Math 116:97–133
3. Agrawal GP (2001) Nonlinear Fiber Optics, 3rd edn. Academic Press, San Diego
4. Anastassiou C, Segev M, Steiglitz K, Giordmaine JA, Mitchell M, Shih MF, Lan S, Martin J (1999) Energy-exchange interactions between colliding vector solitons. Phys Rev Lett 83(12):2332–2335
5. Anastassiou C, Fleischer JW, Carmon T, Segev M, Steiglitz K (2001) Information transfer via cascaded collisions of vector solitons. Opt Lett 26(19):1498–1500
6. Barad Y, Silberberg Y (1997) Phys Rev Lett 78:3290
7. Berlekamp ER, Conway JH, Guy RK (1982) Winning ways for your mathematical plays, Vol. 2. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London
8. Cao XD, McKinstrie CJ (1993) J Opt Soc Am B 10:1202
9. Chen ZG, Segev M, Coskun TH, Christodoulides DN (1996) Observation of incoherently coupled photorefractive spatial soliton pairs. Opt Lett 21(18):1436–1438
10. Christodoulides DN, Joseph RI (1988) Discrete self-focusing in nonlinear arrays of coupled waveguides. Opt Lett 13:794–796
11. Christodoulides DN, Singh SR, Carvalho MI, Segev M (1996) Incoherently coupled soliton pairs in biased photorefractive crystals. Appl Phys Lett 68(13):1763–1765
12. Cundiff ST, Collings BC, Akhmediev NN, Soto-Crespo JM, Bergman K, Knox WH (1999) Phys Rev Lett 82:3988
13. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theor Phys 21(3/4):219–253
14. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers 1: Anomalous dispersion. Appl Phys Lett 23(3):142–144
15. Islam MN, Poole CD, Gordon JP (1989) Opt Lett 14:1011
16. Jakubowski MH, Steiglitz K, Squier R (1998) State transformations of colliding optical solitons and possible application to computation in bulk media. Phys Rev E 58(5):6752–6758
17. Kang JU, Stegeman GI, Aitchison JS, Akhmediev N (1996) Observation of Manakov spatial solitons in AlGaAs planar waveguides. Phys Rev Lett 76(20):3699–3702
18. Korolev AE, Nazarov VN, Nolan DA, Truesdale CM (2005) Opt Lett 14:132
19. Manakov SV (1973) On the theory of two-dimensional stationary self-focusing of electromagnetic waves. Zh Eksp Teor Fiz 65(2):505–516, [Sov. Phys. JETP 38, 248 (1974)]
20. Mano MM (1972) Computer Logic Design. Prentice-Hall, Englewood Cliffs

21. Menyuk CR (1987) Opt Lett 12:614
22. Menyuk CR (1988) J Opt Soc Am B 5:392
23. Menyuk CR (1989) Pulse propagation in an elliptically birefringent Kerr medium. IEEE J Quant Elect 25(12):2674–2682
24. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45(13):1095–1098
25. Nishizawa N, Goto T (2002) Opt Express 10:1151-1160
26. Radhakrishnan R, Lakshmanan M, Hietarinta J (1997) Inelastic collision and switching of coupled bright solitons in optical fibers. Phys Rev E 56(2):2213–2216
27. Rand D, Steiglitz K, Prucnal PR (2005) Signal standardization in collision-based soliton computing. Int J of Unconv Comp 1:31–45
28. Rand D, Glesk I, Brès CS, Nolan DA, Chen X, Koh J, Fleischer JW, Steiglitz K, Prucnal PR (2007) Observation of temporal vector soliton propagation and collision in birefringent fiber. Phys Rev Lett 98(5):053902
29. Russell JS (1844) Report on waves. In: Report of the 14th meeting of the British Association for the Advancement of Science, Taylor and Francis, London, pp 331–390
30. Shih MF, Segev M (1996) Incoherent collisions between two-dimensional bright steady-state photorefractive spatial screening solitons. Opt Lett 21(19):1538–1540
31. Shor PW (1994) Algorithms for quantum computation: Discrete logarithms and factoring. In: 35th Annual Symposium on Foundations of Computer Science, IEEE Press, Piscataway, pp 124–134
32. Squier RK, Steiglitz K (1993) 2-d FHP lattice gasses are computation universal. Complex Systems 7:297–307
33. Steblina VV, Buryak AV, Sammut RA, Zhou DY, Segev M, Prucnal P (2000) Stable self-guided propagation of two optical harmonics coupled by a microwave or a terahertz wave. J Opt Soc Am B 17(12):2026–2031
34. Steiglitz K (2000) Time-gated Manakov spatial solitons are computationally universal. Phys Rev E 63(1):016608
35. Steiglitz K (2001) Multistable collision cycles of Manakov spatial solitons. Phys Rev E 63(4):046607
36. Yang J (1997) Physica D 108:92–112
37. Yang J (1999) Multisoliton perturbation theory for the Manakov equations and its applications to nonlinear optics. Phys Rev E 59(2):2393–2405
38. Zakharov VE, Shabat AB (1971) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. Zh Eksp Teor Fiz 61(1):118–134, [Sov. Phys. JETP 34, 62 (1972)]

Conduction and Diffusion in Percolating Systems

Barry D. Hughes
Department of Mathematics and Statistics,
University of Melbourne, Melbourne, Australia

Article Outline

Glossary

Backbone In a lattice percolation problem above the percolation threshold, that fraction of the infinite cluster with two disjoint connections to infinity.

Correlation length (ξ) Length scale in a randomly structured system over which the system cannot be regarded as homogeneous.

Critical exponent Exponent characterizing the dominant behavior of an observable quantity near a percolation threshold; e. g. $\xi \sim \text{constant} \times (p_c - p)^{-\nu}$ as $p \uparrow p_c$.

Effective medium approximation Approximate description of an inhomogeneous system obtained by matching averaged local fluctuations in properties in a self-consistent manner.

Flux Vector-valued function (in a continuum) or signed scalar (in a discrete system) quantifying transport or conduction rates.

Lattice Discrete structure (network/graph) of sites (nodes/vertices) connected by bonds (links/edges), including periodic lattices, random lattices, and tree-like or self-similar pseudolattices.

Percolation theory Idealized model of a random medium. In the classical discrete case, the bonds of a lattice are independently open with probability p (Bernoulli bond percolation) or the sites of a lattice are independently occupied with probability p (Bernoulli site percolation). There are various continuum analogues.

Percolation threshold (p_c) Dividing point in parameter space separating cases where long-range connectivity is precluded (infinite connected sets exist with probability 0) from those where long-range connectivity occurs (infinite connected sets exist with positive probability).

Percolative system Random two-phase continuous or discrete system in which one phase is deemed void or non-conducting; usually a percolation threshold exists in such systems.

Potential (V) Function distributed over space or over the sites of a network from which steady-state transport or conduction may be determined.

Pseudolattice A non-periodic discrete structure of sites (nodes/vertices) and bonds (links/edges), most commonly either topologically tree-like or geometrically self-similar.

Random walk Model of random motion, especially on lattices, consisting of a sequence of steps separated by constant or random time intervals.

Recurrent random walk Random walk process on a lattice for which the walker returns to the starting site with probability 1.

Renormalization Converting a system, either exactly or approximately, to a related system with a different characteristic length scale.

Transient random walk Random walk process on a lattice for which the walker has probability less than 1 of returning to the starting site.

Definition of the Subject

The problem of determining the macroscopic structural and transport properties of microscopically non-uniform materials has a long history and is of such central importance that it finds applications in an astonishingly wide range of areas of science and technology, from developmental biology to xerography [109,140,141]. Consider systems that consist of two phases, each of which is homogeneous in its properties, mixed in some way to create an inhomogeneous material. Let one phase occupy a fraction ϕ of the volume. If $\phi \ll 1$, so that the inhomogeneous structure is in some sense dilute, the possibility arises of determining the effective properties of the material as an expansion in powers of ϕ. Simple examples of this, when the dilute phase consists of identical spheres distributed in some reasonable manner, are the prediction variously associated with the names of Maxwell, Clausius, Mossotti, Lorenz, Lorentz and others that spheres of dielectric permittivity ϵ_1 embedded in a matrix of permittivity ϵ_0 produce an effective permittivity ϵ^* given by

$$\frac{\epsilon^* - \epsilon_0}{\epsilon^* + 2\epsilon_0} \approx \frac{\epsilon_1 - \epsilon_0}{\epsilon_1 + 2\epsilon_0}\phi \; ; \qquad (1)$$

and the prediction of Einstein [43,44] that the effective viscosity η^* of a random suspension of rigid, spheres occupying a volume fraction ϕ in a liquid of viscosity η is given by $\eta^* = \eta\left[1 + \frac{5}{2}\phi + o(\phi)\right]$ as $\phi \to 0$. Analogous calculations can be performed for many other "dilute" systems, leading to predictions at low volume fraction for effective electrical conductivity and other attributes. Simple examples demonstrate that as the volume fraction ϕ increases, quantitative information on exactly how the phases are distributed becomes essential for accurate prediction of, or even construction of decent bounds for, effective properties—knowing the volume fractions of the phases is not enough.

We address inhomogeneous systems and especially two-phase systems—both continua and lattices—that are randomly structured, with particular emphasis on systems in which only one phase (with relative abundance ϕ) sustains transport. The simplest such models arise by defining steady-state transport processes or random motions on the occupied phase in the lattice percolation model (▶ Percolation Phase Transition) of Broadbent and Hammersley [25,62,63,64] or its continuum analogues. (▶ Continuum Percolation). We call all such models *percolative models*. They exhibit threshold behavior at $\phi = \phi_c$, separating globally nonconducting states when the transport-sustaining phase is sufficiently rare $(0 \le \phi < \phi_c)$ from conducting states when the transport-sustaining phase is sufficiently abundant $(\phi_c < \phi \le 1)$. When the transport-sustaining phase only sparsely spans a large region, that is, just above the percolation threshold ϕ_c, asymptotic power-law dependence of the transport properties on $\phi - \phi_c$ is observed, but many results now deemed well-known still evade rigorous proof.

Introduction

We address problems of transport and conduction, both in continua and on lattices, where there are randomly distributed local transport properties, especially the cases where there are two phases present (for continua) or there are bonds or sites of two types present (for lattices). Although we later pay most attention to the lattice case, for the present and in Sect. "Continuum Models: Steady-State Phenomena" we discuss primarily continua in which two material phases with different properties are present. Frequently, one of these phases is empty space (void space). Within regions comprising only one phase, all properties of the system are uniform. The properties of regions containing both phases depend not only on the relative proportions of each phase present, but also on the way the phases are distributed. The fundamental question to be answered is this: if we pretend that the system is homogeneous, as it might indeed appear if we viewed a very large piece of it from a distance, what are the effective transport properties of the system?

A complete solution of the problem is not to be expected if the microstructure is elaborate or especially subtle, but sufficient progress has been made since the 1950s that for a number of conceptually simple models, some exact results and many approximate results of decent quality are available. The two cases amenable to treatment are periodic microstructure and random microstructure. The former case, which is not our present concern, presents few conceptual challenges and is increasingly amenable to numerical calculation, since the determination of the global properties of the system can usually be reduced to a study of a single finite region (the fundamental repeat unit) and this is usually computationally tractable.

Interest in modeling systems in which small (usually identical) pieces of one phase are randomly distributed in some way and the remaining phase occupies the rest of space goes back at least to James Clerk Maxwell (1831–1879), Ludvig Valentin Lorenz (1829–1891) and Hendrik Antoon Lorentz (1853–1928). The work of these authors in the latter half of the nineteenth century and some antecedents, culminating in results such as Eq. (1), is reviewed by Landauer [94], Markov [102] and Milton [109]. In the first two decades of the twentieth century the problem was addressed by Einstein [43,44] and others, but four major conceptual developments were needed to bring the subject beyond its infancy.

(a) In 1935 Bruggeman [27], building on ideas of earlier authors but arguing with much greater clarity, developed an *effective medium approximation*, in which an unknown overall property of the composite system, such as its effective conductivity σ_{eff}, is computed in an approximate but self-consistent manner that takes account of fluctuations in the corresponding local property. In the extreme case in which one phase is nonconducting, Bruggeman's approach predicts that for sufficiently small volume fraction ϕ of the conducting phase, there is no conductivity, but there is a critical volume fraction ϕ_c of the conducting phase above which there is conduction, with effective conductivity $\sigma_{\text{eff}} \sim \text{constant} \times (\phi - \phi_c)$ as $\phi \downarrow \phi_c$. Bruggeman's original treatment pays no attention to the manner in which the phases are randomly or periodically distributed.

(b) In 1965, Beran [11,12] established the foundations of a statistical study of effective properties of random media, in which the random placement of phases is statistically quantified. In particular, Beran brought out the importance of considering individual realizations of a random system equipped with a proper probability structure, and drew attention to the important issue of ergodicity, which in this context concerns the relation between volume averages (averages of properties over large volumes in a single realization) and ensemble averages (averages over many realizations).

(c) In 1957, Broadbent and Hammersley [25,62,63,64] considered a lattice model of random media, motivated by a problem of gas mask design for British coal miners. Although quantified transport was not part of their original model—they only addressed connectivity—this work is at the conceptual heart of modern discussions. In their model, now known as, the *percolation theory* lattice is randomly stripped so that either individual bonds remain with probability p (bond percolation) or individual sites remain with probability p (site percolation); the parameter p is the analogue of volume fraction ϕ in a two-phase random continuum, with what remains of the lattice viewed as a conducting phase in the original system. There is a well-defined relative abundance p of the conducting phase, called the *percolation threshold* p_c, above which there is long-distance connectivity and transport is possible. Although the percolation threshold plays for the lattice model a role analogous to the critical volume fraction predicted by Bruggeman's effective medium approximation (and indeed effective medium approximations can be set up for lattice systems), the existence of a precise percolation threshold p_c was rigorously established in 1957, and its value is known exactly for some lattices and approximately but to very high precision for other lattices. The probability that a given site belongs to an infinite connected cluster of the active phase when that phase has relative abundance p is the percolation probability $P_\infty(p)$, and it is accepted that $P_\infty(p) \sim \text{constant} \times (p - p_c)^\beta$ as $p \downarrow p_c$, with $\beta = 5/36$ in simple two-dimensional lattice systems and $\beta \leq 1$ more generally. A number of the techniques and concepts that have developed over 50 years for lattice-based percolation theory have now found their way into natural continuous-medium analogues [104]. Both the lattice and continuum percolation models have many applications [109,138,140,166].

(d) That percolation ideas would be significant in physics was realized almost immediately [5] following the 1957 paper of Broadbent and Hammersely. The final conceptual step needed was first articulated in 1968 by Ziman [174]: the stochastic geometry underlying percolation theory, though essential to the description of transport in random media, does not completely characterize the transport properties: there is also the specific contribution of the transport mechanism. The natural attempt to connect a transport coefficient such as effective conductivity $\sigma_{\text{eff}}(p)$ to the percolation probability $P_\infty(p)$ by writing $\sigma_{\text{eff}}(p) \propto P_\infty(p)$ [42] was refuted by experiments in 1971 of Last and Thouless [95] on the conductivity of a sheet of colloidal graphite with holes randomly punched in it in a manner appropriate to simulate site percolation on the square lattice. The experiments suggested that $\sigma_{\text{eff}}(p) \sim \text{constant} \times (p - p_c)^t$ as $p \downarrow p_c$, with $t > 1 \gg \beta$. The subtlety of the problem arises from the fact that just above the percolation threshold, the sample-spanning structure is tenuous and tortuous, with significant implications on observable quantities that characterize the transport process.

The developments stemming from the approximate work of Bruggeman and the rigorous work of Beran [(a) and (b) above] are well covered in the expansive texts of Milton [109] and Torquato [166]. The bulk of the present article is devoted to random medium problems—both continuous and discrete, though favoring the latter—with a percolative aspect [(c) and (d) above]. Steady-state continuum problems are discussed first in Sect. "Continuum Models: Steady-State Phenomena", with the physical contexts addressed in Sect. "Contexts". These contexts are also claimed by advocates of the analogous lattice models discussed in Sect. "Lattice Models: Steady-State Phenomena", to which we pay more attention. The model on which we primarily focus is the *random resistor network* (introduced in Sect. "The Random Resistor Problem"), which reveals most of the key features of random media, especially the existence of percolation thresholds and the subtlety of the active paths in the system close to the percolation threshold. Although many important and elegant results are now available for geometrical and topological aspects of the underlying percolation process, as discussed in the books by Bollobas and Riordan [21], Grimmett [59] and Hughes [75], and ▶ Percolation Phase Transition, considerable work is still needed to resolve major questions concerning the random resistor network and related systems.

We do not consider the mechanical properties (such as elastic modulii or fracture resistance) of random media, although many of the ideas and results presented here have been extended to mechanical properties in the literature. Less restricted surveys of the properties of random media can be found in the major texts of Milton [109], Sahimi [140,141] and Torquato [166]. Time-evolving problems in randomly structured media are of great interest, but our discussion in Sect. "Random Motion in a Random Environment" strongly emphasizes the model of random walk on a random lattice derived from a percolation process, poetically described as the *ant in the labyrinth* [24,33,75].

Continuum Models: Steady-State Phenomena

Despite the enormous efforts expended on discrete inhomogeneous and percolative systems, most scientific applications that motivate these studies arise in materials perhaps most naturally treated as continua. We briefly summarize several contexts in which percolative systems arise and review selected findings for continua that expose major concepts later discussed more extensively for the more tractable discrete analogues. Examples considered

(i) are scalar, so that we have scalar transport coefficients rather than tensors;

(ii) involve real potentials, so we consider only steady direct current electrical conduction, rather than frequency-dependent alternating current conduction;

(iii) have no explicit time dependence, and

(iv) have linear constitutive response.

The reader may refer to the expansive texts of Milton [109] and Sahimi [140,141] for discussions without the restrictions (i)–(iv).

Contexts

Problems Equivalent to Electrical Conduction In a variety of physical contexts the following mathematical problem arises. Let Ω be a domain (a connected spatial region), let $\sigma(\mathbf{r})$ (which we shall call a *transport coefficient*) be a prescribed function of position $\mathbf{r}$ in Ω, and let ∇ denote the usual gradient operator. We are to find a *potential* $V(\mathbf{r})$ that satisfies the equation

$$\nabla \cdot (\sigma \nabla V) = 0, \quad \mathbf{r} \in \Omega\,, \tag{2}$$

subject to prescribed conditions on the boundary $\partial\Omega$ of Ω, usually either the Dirichlet boundary condition [$V(\mathbf{r})$ prescribed], the Neumann [$\mathbf{n} \cdot \nabla V$ prescribed, where $\mathbf{n}$ is a unit vector normal to $\partial\Omega$ and, for consistency, the surface integral of $\mathbf{n} \cdot \nabla V$ over $\partial\Omega$ is zero], or Dirichlet and Neumann conditions applied to disjoint components of $\partial\Omega$. The potential V, which is uniquely defined for the Dirichlet boundary-value problem and is unique up to an additive constant for the pure Neumann boundary-value problem, is associated with a vector field

$$\mathbf{E} = -\nabla V\,, \tag{3}$$

and the field $\mathbf{E}$ and the transport coefficient σ determine a flux vector

$$\mathbf{J} = \sigma \mathbf{E}\,. \tag{4}$$

Equation (2) can be interpreted as a statement of a conservation law under steady-state conditions for a substance carried by the flux vector $\mathbf{J}$. Where the transport coefficient has surfaces of discontinuity (as is the case in two-phase media), one interprets the partial differential equations as holding in the distributional sense and the boundary conditions of continuity of V and of $\mathbf{n} \cdot \sigma \nabla V$ (where $\mathbf{n}$ is normal to the phase interface) follow.

The canonical example of the problem embodied in Eqs. (2)–(4) is steady state (direct current) electrical conduction, with V the electrostatic potential, $\mathbf{E}$ the electric field, $\mathbf{J}$ the electric current, and σ the electrical conductivity. Six different interpretations of the same mathematical

Conduction and Diffusion in Percolating Systems, Table 1
Mathematically equivalent potential theory problems

Transport coefficient	Flux law	Potential–field relation
Electrical conductivity σ	$\mathbf{J} = \sigma\mathbf{E}$	$\mathbf{E} = -\nabla V$ from Maxwell's $\nabla \times \mathbf{E} = \mathbf{0}$
Dielectric permittivity ϵ	$\mathbf{D} = \epsilon\mathbf{E}$	$\mathbf{E} = -\nabla V$ from Maxwell's $\nabla \times \mathbf{E} = \mathbf{0}$
Magnetic permeability μ	$\mathbf{B} = \mu\mathbf{H}$	$\mathbf{H} = -\nabla V$ from Maxwell's $\nabla \times \mathbf{H} = \mathbf{0}$

Transport coefficient	Flux–potential relation
Thermal conductivity κ	Fourier's Law (temperature T) $\mathbf{Q} = -\kappa\nabla T$
Diffusivity D	Fick's Law (pressure P, viscosity η) $\mathbf{q} = -(k/\eta)\nabla P$
Permeability k	Darcy's Law (concentration c) $\mathbf{J} = -D\nabla c$

problem are given in Table 1. For the first three interpretations, the potential V is the consequence of a fundamental physical equation of the form $\nabla \times \mathbf{E} = \mathbf{0}$. In the remaining examples in Table 1), the flux law $\mathbf{J} = \sigma\mathbf{E}$ and the potential–field relation $\mathbf{E} = -\nabla V$ do not have individual fundamental physical interpretations.

Caveats to the Electrical Interpretation In our discussion we shall use the terminology appropriate to electrical conductivity, however in interpreting the results in other contexts several caveats are needed. For time-dependent diffusive processes, we really wish to solve the equation

$$\frac{\partial c}{\partial t} = \nabla \cdot (D\nabla c)\,; \tag{5}$$

the electrical conduction analogue applies only to the steady state. Similar issues arise for heat conduction. For diffusion, problems in which the tracer is injected at an isolated point (a point source) are of great interest. In the extreme case in which one phase sustains diffusion and the other does not, the effect of placing the source at an arbitrary point in the phase that sustains diffusion depends critically on whether the source falls in a conducting region of finite extent or of infinite extent and these issues are not covered by a discussion confined to the electrical conduction analogue.

Although we have included porous medium permeability in Table 1, this exact equivalence to the electrical conduction problem only applies when we work at length scales large compared to pore sizes, so that the notion of permeability—whether constant or variable—is meaningful. A problem of great interest arises when one considers fluid flow through the void space of a medium consisting of an impermeable solid phase and a void phase [139]. In this case, there is no local partial differential equation at all for the solid phase, while in the void space the fluid is subject to the low Reynolds number limit of the Navier–Stokes equations.

The Idea of Homogenization

If the local conductivity σ is bounded, but fluctuates in some manner, one may seek a "homogenized" description of the medium, replacing the variable conductivity problem $\nabla \cdot (\sigma\nabla V) = 0$ by a uniform conductivity problem $\nabla \cdot (\sigma_{\text{eff}}\nabla V) = 0$ that matches it in some sense that needs to be made precise, with the constant *effective conductivity* σ_{eff} to be determined. A porous medium, viewed as a two-phase void/solid system as mentioned above, homogenizes in a different manner, since the equations for the desired uniform system, embodying Darcy's Law, are structurally different from the equations that govern the flow in the pore space [139]. We do not discuss this case here.

There are two basic rigorous approaches. One, on which we focus in the present article, is introduced in Sect. "Effective Conductivities". The other, which is generally described as *homogenization theory*, may be briefly summarized as follows. There are three length scales in the problem: a microscale, on which the structure fluctuates (perhaps strongly), a mesoscale over which unstructured continuum behavior emerges, and a macroscale over which parameters defined at the mesoscale may vary slowly. One may define volume averages $\langle\mathbf{E}\rangle$ and $\langle\mathbf{J}\rangle$ of the electric field over some representative domain large compared to the length scale on which microstructure fluctuates (that is, a mesoscale domain) and introduce a homogenized conductivity $\widehat{\sigma}$ defined by

$$\langle\mathbf{J}\rangle = \widehat{\sigma}\langle\mathbf{E}\rangle\,. \tag{6}$$

Homogenization is successful if, in an appropriate limit as the microscale dimension is sent to zero, a well-defined $\widehat{\sigma}$ emerges. The theory of homogenization has been well worked out for periodic microstructures [4,10], which are not our concern here, but is more challenging for the random microstructure case, although considerable progress has also been made there [56,79]; surveys of recent work will be found in the texts of Milton [109] and

Torquato [166]. The first strong results were derived for uniformly elliptic partial differential operators, which in the present context means that for some constants α and β we have $0 < \alpha \leq \sigma \leq \beta < \infty$, precluding the analysis of percolative systems, but extensions to percolative systems are discussed in the book of Jikov, Kozlov and Oleinik [79]. Percolative systems are subtle because the length scale on which the inhomogeneity is important—the correlation length—diverges at the percolation threshold; and the detailed geometrical arrangement of the conducting and non-conducting phases is very important except perhaps for very low or very high volume fractions of the conducting phase. It should be emphasized that in homogenizing random media, the determination of exact values of $\widehat{\sigma}$, either by formula or in terms of an algorithm for easy, highly precise computation, largely remains a distant goal. Establishing the existence of a well-defined $\widehat{\sigma}$ and determining any of its nontrivial qualitative properties is already a major achievement.

Effective Conductivities

As an alternative to the homogenization approach outlined above, one may simply define effective properties of an inhomogeneous system by a black box approach. In the electrical analogy, and considering three dimensions for definiteness, take a finite rectangular prism $-M < x < M$, $-M < y < M$, $0 \leq z \leq L$, occupied by the inhomogeneous conductor, with all boundaries save for $z = 0$ and $z = L$ being insulated ($\mathbf{n} \cdot \sigma \nabla V = 0$). If one specifies the potential to be $V = V_0$ at $z = 0$ and $V = 0$ at $z = L$, and the z-component of the current through the surface $x = 0$ is $J(x, y)$ per unit area, then a reasonable definition of the effective conductivity σ_{eff} is

$$\sigma_{\text{eff}} = \lim_{L\to\infty} \lim_{M\to\infty} \frac{1}{V_0/L} \int_{-M}^{M} \int_{-M}^{M} \frac{J(x, y)\mathrm{d}x\mathrm{d}y}{(2M)^2} , \qquad (7)$$

where L/M is held constant as L, $M \to \infty$. This is the most natural definition for physical experiments or computer simulation and the one we adopt below in our more detailed discussion of discrete models. Actually, for a randomly-structured two-phase system, unless one establishes that for the particular statistical model of the microstructure, the limit is well-defined and takes the same value for almost all realizations of the microstructure (that is, the limit exists and takes a unique value with probability 1), the integral has to be averaged over all realizations of the microstructure before the limit $L, M \to \infty$ is taken.

An alternative definition replaces the insulated boundary conditions on all faces other than $z = 0$ or $z = L$ by $V = V_0(L - z)/L$. One could alternatively set up a definition of effective conductivity by prescribing a uniform injection of current per unit area across the face $z = 0$ and studying the induced potential difference.

Golden and Papanicolaou [56] have examined the relation between the black box definition of effective conductivity for a finite region and the definition from homogenization, and shown their equivalence under the conditions of uniform ellipticity. Stronger results applicable to percolative systems will be found in Jikov et al. [79].

While results on the existence of a well-defined effective conductivity σ_{eff} are of course of significant interest, one really wants to determine its value.

Exact Results Few nontrivial exact results on the effective conductivity are available. Consider a one-dimensional conductor of length L, comprising N independent random conducting elements of length Δ, with the kth element having conductivity σ_k and sustaining a potential drop V_k. Then the current flowing is $J = \sigma_k V_k/\Delta$, the same in each element, and the field is

$$\frac{1}{L}\sum_{k=1}^{N} V_k = \frac{1}{N\Delta}\sum_{k=1}^{N} \frac{J\Delta}{\sigma_k} \to J\langle\sigma^{-1}\rangle , \qquad (8)$$

where we have taken the limit $N \to \infty$ and used the Strong Law of Large Numbers [47]. Hence we have the exact result that

$$\sigma_{\text{eff}} = \frac{1}{\langle\sigma^{-1}\rangle} . \qquad (9)$$

For two-dimensional systems there are a number of "phase interchange" or "duality" results, due to Keller [82] and others, discussed carefully by Milton [109]. In particular, an infinite chessboard, with the white squares having conductivity σ_{w} and the black squares having conductivity σ_{b} produces an exact effective conductivity

$$\sigma_{\text{eff}} = \sqrt{\sigma_{\text{w}}\sigma_{\text{b}}} . \qquad (10)$$

Bounds The electrical energy dissipation rate associated with a potential distribution V and corresponding electric field $\mathbf{E}$ and current $\mathbf{J}$ in the finite $2M \times 2M \times L$ rectangular prism of Sect. "Effective Conductivities" is

$$\begin{aligned}\mathcal{E}\{V\} &= \int_{-M}^{M}\int_{-M}^{M}\int_{0}^{L} \mathbf{J}\cdot\mathbf{E}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &= \int_{-M}^{M}\int_{-M}^{M}\int_{0}^{L} \sigma|\nabla V|^2\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z .\end{aligned} \qquad (11)$$

Consider a potential V that satisfies the partial differential equation $\nabla \cdot (\sigma \nabla V) = 0$ in the interior and the boundary conditions $v = V_0$ on $z = 0$, $V = 0$ on $z = L$ and $\mathbf{n} \cdot \sigma \nabla V = 0$ on other surfaces: we call this an *admissible potential*. Let W, a *trial potential*, be any other function that satisfies the same boundary conditions but may not satisfy the partial differential equation. If we write $\eta = W - V$, expand $\mathcal{E}\{W\} = \mathcal{E}\{V + \eta\}$ and use the Divergence Theorem and the boundary conditions, we easily show that $\mathcal{E}\{W\} \geq \mathcal{E}\{V\}$, so that admissible potentials minimize energy dissipation. That is, we have a variational principle. It can be shown under reasonable conditions that

$$\sigma_{\text{eff}} \leq \frac{1}{(V_0/L)^2 (2M)^2 L} \int_{-M}^{M} \int_{-M}^{M} \int_{0}^{L} \sigma |\nabla W|^2 \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\,, \tag{12}$$

so that trial potentials can be used to construct rigorous upper bounds on the effective conductivity. A complementary variational principle that leads to lower bounds on the effective conductivity can be derived by analyzing trial current distributions. The variational approach has its origin in work in 1955 of Brown [26] and has been developed in diverse contexts by many authors (Brown, Beran, Prager, Hashin and Shtrikman, Davies, ...) with varying degrees of rigor [109,166]. Bergman [15] introduced an alternative approach to variational methods for deriving bounds on effective conductivities, based on complex variable techniques. Bergman's work and many important consequences are reviewed by Milton [109].

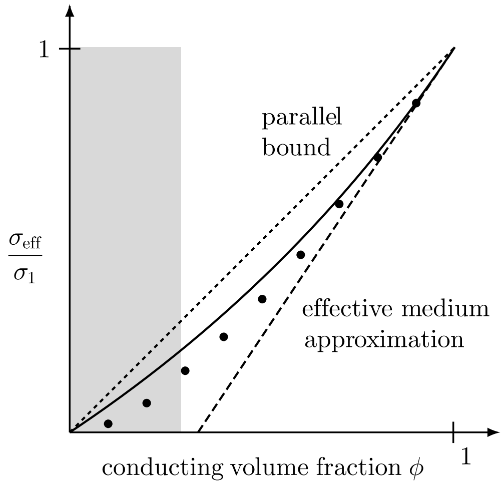
The most easily derived bounds hold without restriction on the distribution of phases in a multiphase system, and also apply for continuously varying local conductivities:

$$\frac{1}{\langle \sigma \rangle} \leq \sigma_{\text{eff}} \leq \langle \sigma \rangle\,, \tag{13}$$

where angle brackets denote the volume average. These bounds, first identified in 1912 by Wiener [168], are the best possible general bounds, since they become equalities in the case of slabs of homogeneous material parallel to the face $z = 0$ (conductors in series—$\sigma_{\text{eff}} = 1/\langle \sigma^{-1} \rangle$) or no variation in conductivity with z (conductors in parallel—$\sigma_{\text{eff}} = \langle \sigma \rangle$). If no additional assumptions are made except for statistical isotropy, the 1962 bounds of Hashin and Shtrikman [67] become the best possible. For high and low conductivity phases with volume fractions ϕ_{high} and ϕ_{low} and conductivities $\sigma_{\text{high}} \geq \sigma_{\text{low}}$, the Hashin–Shtrikman bounds in d dimensions ($d = 2$ or 3) are

$$\begin{aligned} \sigma_{\text{low}} + \frac{d\phi_{\text{high}}\sigma_{\text{low}}(\sigma_{\text{high}} - \sigma_{\text{low}})}{d\sigma_{\text{low}} + \phi_{\text{low}}(\sigma_{\text{high}} - \sigma_{\text{low}})} &\leq \sigma_{\text{eff}} \\ &\leq \sigma_{\text{high}} - \frac{d\phi_{\text{low}}\sigma_{\text{high}}(\sigma_{\text{high}} - \sigma_{\text{low}})}{d\sigma_{\text{high}} - \phi_{\text{high}}(\sigma_{\text{high}} - \sigma_{\text{low}})}\,. \end{aligned} \tag{14}$$

If we consider the case of a percolative system, where we have $\sigma_{\text{low}} = 0$, $\sigma_{\text{high}} = \sigma_1$ and for brevity we write $\phi_{\text{high}} = \phi$ (that is, ϕ is the volume fraction for the conducting phase) the lower bounds in (13) and (14) become trivial ($\sigma_{\text{eff}} \geq 0$) and neither upper bound is very strong, as shown by Fig. 1. For the case of spheres with independently located centers (technically, Poisson points) and either constant radii, or random radii with a reasonable distribution, it can be rigorously established [104] that for sufficiently small ϕ, the probability that there is an infinite connected region of the conducting phase is zero and consequently the effective conductivity will be zero below a percolation threshold ϕ_c. Numerical simulations with Poisson-centered conducting spheres of constant radius show that $\phi_c \approx 0.2895 \pm 0.0005$ [135]: the gray rectan-



Conduction and Diffusion in Percolating Systems, Figure 1

Bounds on the effective conductivity σ_{eff} for a conducting phase of conductivity σ_1 randomly and homogeneously distributed at volume fraction ϕ, with remaining space occupied by a nonconducting phase. The *gray* area lies below the percolation threshold estimate $\phi_c \approx 0.2895 \pm 0.0005$ of Rintoul and Torquato [135] for overlapping conducting spheres in a nonconducting matrix. The solid discs are numerical estimates of Kim and Torquato [87] for overlapping non-conducting spheres in a conducting matrix, where $\phi_c \approx 0.03$. *Upper broken line*: elementary ("parallel") bound $\sigma_{\text{eff}}/\sigma_1 \leq \phi$. Solid curve: Hashin–Shtrikman bound $\sigma_{\text{eff}}/\sigma_1 \leq 2\phi/(3 - \phi)$ [67]. Lower broken line: effective medium approximation $\sigma_{\text{eff}}/\sigma_1 \approx (3/2)(\phi - 1/3)$ for $\phi \geq 1/3$ from work of Bruggeman [27]

gle in Fig. 1 spans the interval $0 \leq \phi \leq \phi_c$ for this system, and parts of the upper bound curves (13) and (14) that intersect the Gray rectangle miss the threshold behavior entirely and are misleadingly uninformative for this specific system.

In Fig. 1 we also show as discs numerical estimates [87] of the conductivity in the case where the randomly placed spheres are non-conducting, the rest of space is conducting and the percolation threshold is much smaller ($\phi_c \approx 0.03$ [87]); physicists call this the *Swiss cheese model*. The Hashin–Shtrikman upper bound (14) works better because the percolation threshold is so low, but the Hashin–Shtrikman lower bound remains useless. Much better bounds than those shown are available for non-percolative problems, especially if additional information about correlations between phase locations is available [109,166].

Approximations Clarifying the work of earlier authors, in 1935 Bruggeman [27] derived an *effective medium approximation* for the effective conductivity. This approach represents an uncontrolled approximation—it does not give rigorous bounds and a priori estimation of its accuracy is not possible. One considers a spherical inclusion of unspecified constant conductivity within a uniform medium whose conductivity is taken to be the unknown effective conductivity σ_{eff}, which will be estimated "self-consistently". For a prescribed constant field **E** at infinity, the field within the inclusion is calculated. The requirement that, when averaging over material properties within the inclusion, the average field in the inclusion does not differ from the field at infinity yields an approximate equation for σ_{eff},

$$\left\langle \frac{3\sigma_{\text{eff}}}{\sigma + 2\sigma_{\text{eff}}} \right\rangle = 1\,. \tag{15}$$

In the case in which $\sigma = \sigma_1$ with probability ϕ, corresponding to a volume fraction ϕ of the conducting phase, this approximation predicts that for a percolative system

$$\sigma_{\text{eff}} = \frac{3}{2}\left(\phi - \frac{1}{3}\right)\sigma_1\,. \tag{16}$$

The physical requirement that $\sigma_{\text{eff}} \geq 0$ leads one to interpret this as predicting that $\sigma_{\text{eff}} = 0$ for $0 \leq \phi \leq 1/3$, so that there is a threshold for the conductivity at $\phi = 1/3$.

As shown in Fig. 1, this approximation closely conforms to the Hashin–Shtrikman upper bound near $\phi = 1$. Although it has the desirable feature of predicting a conductivity threshold, the actual predicted threshold (1/3) is not particularly close to numerical estimate for randomly placed spheres with independent centers (0.2895) and is very far from the estimated threshold for the Swiss cheese model (0.03).

Differences Between Continuum and Discrete Models

The considerable emphasis in the literature on discrete (lattice) models is partly motivated by the folklore of *universality*, under which, apart from some reasonable conditions that exclude direct, long-range connections and other oddities, qualitative features and associated critical exponents of analogous problems depend only on dimensionality. While this appears to be broadly true, there are some significant exceptions, two of which we mention here.

(a) A parameter value where properties of the infinite system which are elsewhere analytic lose analyticity is called a *critical point*. In the percolation model of Broadbent and Hammersley [25,62,63,64] on periodic lattices, there is only one critical point associated with global connectedness [106], namely the percolation threshold, and this coincides with the only critical point for the effective conductivity. Following on work of Kozlov [93], Jikov et al. [79] have discussed carefully a random chessboard, where the plane is divided into equal squares, colored independently. Squares are black (conductivity σ_b) with probability ϕ and white (conductivity σ_w) with probability $1 - \phi$. For $\phi = 1/2$, Eq. (10) holds, but more importantly, in the random chessboard model there are effectively two thresholds for a given color, one for long-distance connectivity using corner connections and one for long-distance connectivity through adjoining edges of the squares. If we let $p_c \approx 0.59$ denote the site percolation threshold of the square lattice (▶ Percolation Thresholds, Exact), then the following results hold as $\sigma_b \to 0$ with σ_w held fixed. For $0 \leq \phi \leq 1 - p_c$, the white phase has large components connected via edges of squares and

$$\lim_{\sigma_b \to 0} \sigma_{\text{eff}} = \sigma_w f(\phi) > 0\,.$$

For $1 - p_c < \phi < p_c$ neither color has large components connected via edges of squares, but both colors have large components connected via corners, and

$$c_1(\phi)\sqrt{\sigma_w \sigma_b} \leq \sigma_{\text{eff}} \leq c_2(\phi)\sqrt{\sigma_w \sigma_b}\,.$$

For $p_c < \phi < 1$, we have $\sigma_b \leq \sigma_{\text{eff}} \leq c_3(\phi)\sigma_b$. Berlyand and Golden [16] have established the stronger result that for $1 - p_c < \phi < p_c$, $\sigma_{\text{eff}} = \sqrt{\sigma_w \sigma_b} + O(\sigma_b)$ as $\sigma_b \to 0$ with σ_w held fixed. Torquato et al. [167] have reported numerical studies.

(b) The universality concept presupposes that all analogous systems in the same dimension have the same critical exponents for a given system property. Thus, for example, in a percolative continuous system where ϕ is the volume fraction of the conducting phase, we expect to see

$$\sigma_{\text{eff}} \sim \text{constant} \times (\phi - \phi_c)^t \quad \text{as } \phi \downarrow \phi_c\,. \tag{17}$$

where the conductivity exponent t (the general existence of which remains to be established, even for specific two and three dimensional systems) is independent of microstructure, and the same for lattices and continua in a given dimension. This expectation turns out to be wrong, with hopes being dashed by computational studies of Feng et al. [48] on the Swiss cheese model introduced in Sect. "Bounds": in two dimensions the lattice and Swiss cheese continuum model exponents are close together and possibly coincident; in three dimensions the exponent t for the Swiss cheese continuum is about 0.5 larger than the exponent t for the simple cubic lattice.

Lattice Models: Steady-State Phenomena

Notwithstanding the caveats exposed in Sect. "Differences Between Continuum and Discrete Models", we turn to a detailed discussion of discrete systems. We consider *lattices* (also known as networks or graphs) of *sites* (also known as nodes or vertices) connected by *bonds* (also called links or edges). The number of bonds attached to a site is its *coordination number* (also called degree or valence), and two sites that are attached to the same bond are called *nearest-neighbour sites*. The most commonly studied lattices are periodic in structure, the simplest example being the *hypercubic lattice* $\mathbb{Z}^d$ of sites with integer coordinates, connected by bonds of unit length, so that each site has coordination number $2d$. Other important infinite networks are sometimes called *pseudolattices* to distinguish them from periodic structures. The most important examples of pseudolattices are self-similar or fractal structures [100] (► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation) and tree-like structures, especially the *Cayley tree* or *Bethe lattice* [165] of coordination number z, in which there are no closed loops.

Key Results from Percolation Theory

In percolation theory for lattices sites, are declared *occupied* or *vacant* in some random way and bonds are declared *open* or *closed* in some random way. A site is deemed to belong to a cluster of size $n \geq 1$ (where n may be finite or infinite) if it is present and connected via open bonds to $n-1$ other sites; a vacant site is a cluster of size 0. In the (Bernoulli) *bond problem*, all sites are occupied, and bonds are independently declared open with probability p and closed with probability $1-p$. In the (Bernoulli) *site problem*, sites are occupied with probability p and vacant with probability $1-p$, and a bond is declared open if and only if the two sites it joins are both occupied. Every bond percolation problem is equivalent to a site percolation problem on a related lattice, so that it suffices to state general results for site percolation, although in many examples modeling transport and conduction, the language of bond percolation is more appropriate.

Viewed at sufficiently low resolution, a realization of bond or site percolation other than at the percolation threshold appears homogeneous. A measure of the length scale over which inhomogeneous structure is important is furnished by the *correlation length* $\xi(p)$. One way of defining this precisely on $\mathbb{Z}^d$ is to let τ_m denote the probability that the sites at $\mathbf{0}$ and $(m, 0, 0, \ldots, 0)$ belong to a common finite cluster. Then $\xi(p) = -\lim_{m\to\infty} m^{-1} \ln \tau_m$.

Rigorous General Results The following results are rigorously proven [21,59,75] for Bernoulli site percolation, provided that the lattice or pseudolattice is homogeneous in the sense that all sites are equivalent (as is the case for $\mathbb{Z}^d$ and the Bethe lattice), and the number of sites at most n bonds distant from a given site grows no faster than $c \exp(an^\kappa)$, where $a > 0$, $c > 0$ and $\kappa < 1$.

(i) There exists a percolation threshold $p_c > 0$ such that the probability $P_\infty(p)$ of a chosen site belonging to an infinite cluster is zero for $p < p_c$ and strictly greater than zero for $p > p_c$.

(ii) The probability $P_n(p)$ that a chosen site belongs to a cluster of size n decays exponentially rapidly with increasing n when $p < p_c$.

The exponential decay of $P_n(p)$ ensures the finiteness of the mean cluster size $\chi(p) = \sum_{n=0}^{\infty} nP_n(p)$ for $p < p_c$. The sum remains meaningful for $p > p_c$ if we exclude the infinite cluster $n = \infty$ and $\chi(p)/(1 - P_\infty(p))$ becomes the mean cluster size, conditioned on the cluster being finite. A result related to (ii) establishes that on $\mathbb{Z}^d$, the correlation length is well-defined and finite for $p < p_c$.

(iii) The mean cluster size $\chi(p)$ is finite for $p < p_c$ but $\chi(p) \geq pp_c(p_c - p)^{-1}$ when $p < p_c$, so that $\chi(p) \to \infty$ as $p \uparrow p_c$.

(iv) For a given value of p, there is a number k_0 (which may take no values other than 0, 1 or ∞) such that the number of distinct infinite clusters of occupied sites on the lattice takes the value k_0 with probability 1. When $P_\infty(p) = 0$, $k_0 = 0$.

For some pseudolattices, such as the Bethe lattice, if $P_\infty(p) > 0$ then $k_0 = \infty$; however, for a class of lattices which includes normal periodic lattices, if $0 < P_\infty(p) < 1$

then $k_0 = 1$. That is, when an infinite cluster exists it is unique.

(v) For $p > p_c$ a site is part of the backbone if it lies on the infinite cluster, and within that cluster there are two independent paths to infinity. If $B(p)$ is the probability that a given site belongs to the backbone then $B(p) \le p^{-1} P_\infty(p)^2$.

For a discussion of the determination of and actual values of site and percolation thresholds for particular lattices, see ► Percolation Thresholds, Exact.

Heuristics A heuristic scaling theory for percolation theory (► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation introduced in 1979 by Stauffer [153] and analogies between percolation and statistical mechanics first exhibited in 1972 by Fortuin and Kasteleyn [51,52] lead to the following folklore [75]. There are *critical exponents* that describe the non-analytic behavior of important attributes of the percolation model at the percolation threshold. If we use the proportionality symbol $\propto$ as a short notation, so that $f(p) \propto g(p)$ as $p \to p_c$ means that $f(p)/g(p) \to$ nonzero constant, then the critical exponents are defined as follows:

cluster size distribution at p_c

$$P_n(p_c) \propto n^{-1-1/\delta} \qquad \text{as } n \to \infty\,; \tag{18}$$

mean cluster size below p_c

$$\chi(p) \propto (p_c - p)^{-\gamma} \qquad \text{as } p \uparrow p_c\,; \tag{19}$$

mean cluster size above p_c

$$\chi(p) \propto (p_c - p)^{-\gamma'} \qquad \text{as } p \downarrow p_c\,; \tag{20}$$

correlation length below p_c

$$\xi(p) \propto (p_c - p)^{-\nu} \qquad \text{as } p \uparrow p_c\,; \tag{21}$$

correlation length above p_c

$$\xi(p) \propto (p_c - p)^{-\nu'} \qquad \text{as } p \downarrow p_c\,; \tag{22}$$

percolation probability

$$P_\infty(p) \propto (p - p_c)^{\beta} \qquad \text{as } p \downarrow p_c\,; \tag{23}$$

backbone probability

$$B(p) \propto (p - p_c)^{\beta_{\text{bb}}} \qquad \text{as } p \downarrow p_c\,. \tag{24}$$

For the Bethe lattice, it is established that $\delta = 2, \gamma = 1$, $\beta = 1$ and $\beta_{\text{bb}} = 2$ [75], but for standard lattices even the existence of critical exponents in the sense explained above is not rigorously established, although as noted below, a few weaker results consistent with this picture, such as $\lim_{p \downarrow p_c} \ln[P_\infty(p)]/\ln(p - p_c) = \frac{5}{36}$ for site percolation on the triangular lattice, are now proven.

Folklore asserts that exponents are the same for the all site and bond problems in a given dimension (universality), vary significantly with the dimension d in low dimensions, and are independent of dimension for $d \ge 6$, where they take the so-called *mean-field values* $\beta = \gamma = 1$ and $\nu = \frac{1}{2}$. (These values are rigorously established by work of Hara and Slade [66] for standard percolation for $d \ge 19$, and for a spread-out model of percolation for $d \ge 6$.) At $d = 6$, the prefactors in the dominant asymptotic forms that define the critical exponents may be modified by the inclusion of a power of $\ln|p - p_c|$. It is generally believed that $\gamma = \gamma'$ and $\nu = \nu'$. On modest hypotheses on lattice structure it is rigorously established [75] from work of various authors that, provided the exponents exist, $\beta \le 1$, $\gamma \ge 1, \gamma/d \le \nu \le \gamma$, and $\nu \ge 2/d$, but two much more informative scaling relations are accepted:

$$\gamma + 2\beta = \beta(\delta + 1) \tag{25}$$

in all cases, and for $d \le 6$,

$$d\nu = 2\beta + \gamma\,. \tag{26}$$

All credible numerical evidence supports the existence of critical exponents consistent with these scaling laws, they are rigorously established for site percolation on the triangular lattice [148], and natural probabilistic conditions equivalent to (26) for bond percolation on $\mathbb{Z}^d$ are known [22].

Rigorous Results for Two Dimensions For several two-dimensional lattices a number of exact results are known: in particular $p_c = 1/2$ for site percolation on the triangular lattice (coordination number 6) and bond percolation on the square lattice (coordination number 4). In these particular problems it has also been rigorously established that the percolation probability $P_\infty(p)$ is continuous for $0 \le p \le 1$, so that there is no infinite cluster present right at the percolation threshold. This result is believed to be true, but remains unproven, in three dimensions.

Major innovations in probability since 2000, primarily associated with the work of Lawler, Schramm, Smirnov and Werner, have led to the following results being rigorously established for the case of site percolation on the two-dimensional triangular lattice [148]:

$$P_\infty(p) = (p - \tfrac{1}{2})^{5/36 + o(1)} \qquad \text{as } p \downarrow \tfrac{1}{2}\,; \tag{27}$$

$$\chi(p) = |p - \tfrac{1}{2}|^{-43/18 + o(1)} \qquad \text{as } p \to \tfrac{1}{2}\,; \tag{28}$$

$$\xi(p) = |p - \tfrac{1}{2}|^{-4/3 + o(1)} \qquad \text{as } p \to \tfrac{1}{2}\,. \tag{29}$$

This is almost (but not quite) a proof of existence of critical exponents in the sense of physicists, and moreover shows that the scaling relations (25) and (26) hold in this one particular case and the accepted folklore for two dimensions that

$$\beta = \tfrac{5}{36}, \qquad \gamma = \gamma' = \tfrac{43}{18}, \qquad \nu = \nu' = \tfrac{4}{3}, \tag{30}$$

arrived at from scaling arguments, statistical mechanical arguments, heuristic arguments and numerical evidence, is indeed correct in this one particular case. The exact backbone exponent β_{bb} remains unknown, but it has been shown to be determined by the leading eigenvalue of a differential operator [96].

The Random Resistor Problem

The canonical discrete model of transport and conduction is the *random resistor network*, in which the bonds of a lattice or some less regular network are resistors of random resistance. A more extended discussion of this model, including proofs of many results stated here, will be found in Hughes [75]; for the Dirichlet, Thomson and Rayleigh Principles discussed below see also Doyle and Snell [40]. The first study of random resistor networks may fairly be ascribed to 1971 work of Kirkpatrick [88,89], although an equivalent system had already been studied as early as 1956 by Fatt [45] as a model for porous media, with hydraulic conductivities of bonds playing the role of resistance and pressure playing the role of voltage.

Problem Definition If we use Greek subscripts to index bonds of the lattice (each bond being assigned a nominal direction of orientation), then the (direct) current I and the potential difference V across a particular bond α are related by Ohm's Law $I = G_\alpha V$, where G_α is the conductance of bond α (the reciprocal of the resistance). Current is a signed quantity. If the potential at site $\mathbf{s}'$ is less than that at site $\mathbf{s}$ then the current flowing from $\mathbf{s}$ to $\mathbf{s}'$ is positive.

It is customary to assume that the conductances of different bonds are independent, identically distributed random variables. Note that one might consider more generally the case in which $I = \mathcal{G}_\alpha(V)$ where $\mathcal{G}_\alpha$ is a nonlinear function with several random parameters, but this problem is both subtle [29] and difficult [19,20,61,83,105,161] and will not be discussed here; see Sahimi [141].

If the bonds of a random resistor network have independent, identically distributed conductances governed by the probability density function

$$f(g) = (1-p)\delta_+(g) + p h(g), \quad 0 < p < 1, \tag{31}$$

where $h(g)$ is the conditional probability of the conductance, given that the conductance is nonzero, we call the network the *general percolation conduction problem*. We use the notation $\delta_+(g)$ to distinguish a delta function that acts on the right from the usual symmetric delta function $\delta(x)$. The specific case in which $h(g) = \delta(g - g_0)$, that is,

$$f(g) = (1-p)\delta_+(g) + p\delta(g - g_0)), \quad 0 < p < 1, \tag{32}$$

will be called the *standard* percolation conduction problem. The definitions (31) and (32) are directly associated with bond percolation. A standard percolation problem for site percolation arises on assigning each bond that links two occupied sites the conductance g_0, and making all other bonds non-conducting.

In general, we have a lattice with a potential (voltage) $V_{\mathbf{s}}$ at each site $\mathbf{s}$ and a current $i_{\mathbf{rs}}$ flowing from site $\mathbf{r}$ to site $\mathbf{s}$ along a bond of conductance $G_{\mathbf{rs}} = G_{\mathbf{sr}}$ (or, equivalently, resistance $R_{\mathbf{rs}} = 1/G_{\mathbf{rs}}$). From Ohm's Law

$$i_{\mathbf{rs}} = G_{\mathbf{rs}}(V_{\mathbf{r}} - V_{\mathbf{s}}) \tag{33}$$

and Kirchhoff's Law of conservation of current $\sum_{\mathbf{s}} i_{\mathbf{rs}} = I_{\mathbf{r}}$, we have

$$\sum_{\mathbf{s}} G_{\mathbf{rs}}(V_{\mathbf{r}} - V_{\mathbf{s}}) = I_{\mathbf{r}}, \tag{34}$$

where $I_{\mathbf{r}}$ is the current supply into site $\mathbf{r}$, positive if current is injected and negative if it is withdrawn. At most sites $\mathbf{r}$, we have $I_{\mathbf{r}} = 0$, but we impose the additional physical requirement that $\sum_{\mathbf{r}} I_{\mathbf{r}} = 0$, the sum being taken over all sites where current is injected or withdrawn. When appropriate boundary conditions are imposed on finite pieces of periodic lattice, or on finite networks more generally, the problem is completely specified and numerical determination of the potential distribution over the sites and the associated currents is possible; from this estimates of an effective conductivity σ_{eff} that we define in Sect. "Lattices of Dimension $d \geq 2$" below emerge.

Bounds for General Resistor Networks Before addressing the specific case of periodic lattices, we consider arbitrary finite networks. The *effective conductance* $C_{\mathbf{ab}}^{\text{eff}}$ *between two sites* $\mathbf{a}$ and $\mathbf{b}$ and its reciprocal, the *effective resistance* $R_{\mathbf{ab}}^{\text{eff}}$ *between the sites*, are defined by

$$C_{\mathbf{ab}}^{\text{eff}} = \frac{1}{R_{\mathbf{ab}}^{\text{eff}}} = \frac{I}{V_{\mathbf{a}} - V_{\mathbf{b}}}, \tag{35}$$

if a current I injected at site $\mathbf{a}$ and withdrawn at site $\mathbf{b}$ is associated with a potential difference $V_{\mathbf{a}} - V_{\mathbf{b}}$ between the sites. (Effective conductivities arise by scaling effective conductances with appropriate measures of network size.)

Variational principles enable bounds on the effective conductance to be deduced.

Dirichlet's Principle. Consider a finite resistor network with bond conductances $G_{\mathbf{rs}}$ ($0 \leq G_{\mathbf{rs}} < \infty$) and with a specified potential difference V between sites **a** and **b**. Let $\{W_{\mathbf{s}}\}$ be a *trial potential distribution* , that is, any set of real numbers subject to the requirement that $W_{\mathbf{a}} - W_{\mathbf{b}} = V$. Let $\{V_{\mathbf{s}}\}$ be an admissible potential distribution, that is, a set of real numbers satisfying the Kirchhoff–Ohm Law (34) for $\mathbf{s} \neq \mathbf{a}$ or **b**, as well as the constraint $V_{\mathbf{a}} - V_{\mathbf{b}} = V$. Then the effective conductance between sites **a** and **b** is given by

$$C^{\mathrm{eff}}_{\mathbf{ab}} = \frac{1}{2V^2} \sum_{\mathbf{r},\mathbf{s}} G_{\mathbf{rs}}(V_{\mathbf{r}} - V_{\mathbf{s}})^2 \leq \frac{1}{2V^2} \sum_{\mathbf{r},\mathbf{s}} G_{\mathbf{rs}}(W_{\mathbf{r}} - W_{\mathbf{s}})^2 . \tag{36}$$

Thomson's Principle. Consider a finite resistor network with bond resistances $R_{\mathbf{rs}}$ ($0 \leq R_{\mathbf{rs}} < \infty$) and a specified current $I_{\mathbf{a}}$ injected at site **a** and withdrawn at site **b**. Let $\{j_{\mathbf{sr}}\}$ be a *trial current distribution*, that is: $j_{\mathbf{rs}} = -j_{\mathbf{sr}}$; $\sum_{\mathbf{s}} j_{\mathbf{rs}} = 0$ if $\mathbf{r} \neq \mathbf{a}$ and $\mathbf{r} \neq \mathbf{b}$; and $\sum_{\mathbf{s}} j_{\mathbf{as}} = I_{\mathbf{a}} = -\sum_{\mathbf{s}} j_{\mathbf{bs}}$. Let $\{i_{\mathbf{sr}}\}$ be an *admissible current distribution*, that is, a trial current distribution for which Ohm's Law $V_{\mathbf{r}} - V_{\mathbf{s}} = i_{\mathbf{rs}} R_{\mathbf{rs}}$ holds. Then the effective resistance between sites **a** and **b** is given by

$$R^{\mathrm{eff}}_{\mathbf{ab}} = \frac{1}{2I^2_{\mathbf{a}}} \sum_{\mathbf{r},\mathbf{s}} i^2_{\mathbf{rs}} R_{\mathbf{rs}} \leq \frac{1}{2I^2_{\mathbf{a}}} \sum_{\mathbf{r},\mathbf{s}} j^2_{\mathbf{rs}} R_{\mathbf{rs}} . \tag{37}$$

The sums in the inequalities in Dirichlet's and Thomson's Principles may be interpreted as the rate of energy dissipation corresponding to the potential distribution or the current distribution respectively. These two principles assert that admissible potential and current distributions minimize the rate of energy dissipation. From these two principles a number of important results follow, including the following.

Rayleigh's Monotonicity Principle. For finite resistor networks, the resistance between two sites does not decrease if the resistances of some or all of the bonds are increased.

Exactly Solvable Models In one dimension, N random resistors in series yield an effective conductivity (here the overall conductance per unit length) $\sigma_{\mathrm{eff}} = \langle G^{-1} \rangle^{-1}$; the argument is the same as that used above for a random one-dimensional continuum. A slightly less trivial problem for which exact results are available is the standard percolation conduction problem on a Bethe lattice of coordination number z [70,75,156,157]. Pick an arbitrary site $\mathbf{s}_0$ of the lattice, set that site to have potential V_0 and prescribe that the potential is zero at the periphery of the tree (that is, the potential decays asymptotically to zero as we move outwards from $\mathbf{s}_0$ along any infinite path of open bonds that may be present). If a current I_0 flows from site $\mathbf{s}_0$ into the tree, then the tree conductance may be defined as $T = I_0/V_0$. Where the average is taken over all realizations of the system, it can be shown rigorously that $\langle T \rangle$ is zero for $p < p_{\mathrm{c}} = (z-1)^{-1}$ and

$$\langle T \rangle \sim \frac{2zg_0 c}{z-2}[p(z-1)-1]^2 \quad \text{as } p \downarrow p_{\mathrm{c}} , \tag{38}$$

where $c \approx 0.761$ is a constant. Since $p_{\mathrm{c}} = 1/(z-1)$ for the Bethe lattice, this result is consistent with the hypothesis that long-distance conduction is possible in the standard percolation conduction problem (32) if and only if $p > p_{\mathrm{c}}$. It is sometimes argued that the Bethe lattice correctly represents the critical behavior of periodic lattices of sufficiently high dimensionality and that would imply that effective conductivity σ_{eff} defined in Sect. "Lattices of Dimension $d \geq 2$" satisfies $\sigma_{\mathrm{eff}} \propto (p - p_{\mathrm{c}})^2$ as $p \downarrow p_{\mathrm{c}}$. Straley [159] has identified the flaw in this argument—in a Bethe lattice of finite size, a significant fraction of all sites are at the boundary—and has given a definition of the effective conductivity σ_{eff} of a Bethe lattice that gives bonds in all regions of the lattice comparable significance and leads to $\sigma_{\mathrm{eff}} \propto (p - p_{\mathrm{c}})^3$ as $p \downarrow p_{\mathrm{c}}$.

For some remarks concerning self-similar pseudolattices for which exact results are available, see Sect. "Conduction on Fractals".

Lattices of Dimension $d \geq 2$

For random resistor problems more appropriately related to transport and conduction problems in d-dimensional space, we consider periodic lattices. There have been numerical investigations of the standard percolation conduction on topologically random lattices created by the Voronoï algorithm [77,78], but we shall not pursue these extensions of the theory.

Defining the Conductivity Consider the specific case of a finite hypercube Λ_L cut from from the simple hypercubic lattice $\mathbb{Z}^d$, with $(L+1)^d$ sites at locations $\mathbf{s} = (s_1, s_2, \ldots, s_d)$ with integer coordinates subject to the inequalities $0 \leq s_j \leq L$. The faces $s_1 = 0$ and $s_1 = L$ are subjected to a potential difference V, so that there is a potential gradient V/L. In three dimensions, this corresponds to a cubic array of conducting elements sandwiched between two perfectly conducting plates. Boundary conditions on the other faces should be unimportant when L is large, but in numerical work periodic boundary conditions are often used (the faces $s_j = 0$ and $s_j = L$ are joined for $j = 2, \dots d$). If a total current I flows in response to the imposed potential difference, then the current per unit area (that is, per site) of the face $s_1 = 0$ is $I/(L+1)^{d-1}$. We therefore define the (specific) conductivity of the lattice as

$$\lim_{L\to\infty} \frac{I/(L+1)^{d-1}}{V/L} = \lim_{L\to\infty} \frac{L^{2-d} I}{V} . \tag{39}$$

For the random resistor network, one needs to discuss the existence of this limit for each realization ω of the lattice and one is especially interested in the mean value of the limit, which we take as defining the *effective conductivity* of the system:

$$\sigma_{\text{eff}} = \lim_{L\to\infty} \langle L^{2-d} I/V \rangle = \lim_{L\to\infty} \langle L^{2-d} C_L \rangle , \tag{40}$$

where $C_L = I/V$ is the conductance between the faces of the hypercube Λ_L.

For a considerable period the existence of a well-defined effective conductivity remained unproven, with the case of percolative problems providing an especially severe challenge, but this hurdle was eventually overcome [79]. In particular, for percolative problems it is now established there is a well-defined $\sigma_{\text{eff}}(p)$. The averages over realizations in Eq. (40) can be dropped, with the limit existing with probability 1, and being 0 for $p < p_c$ and strictly positive for $p > p_c$

Bounding the Conductivity The Dirichlet and Thompson Principles and Rayleigh's Monotonicity Law are stated for the case in which current is injected at one site and withdrawn at another, but by introducing appropriate links of zero resistance, the laws apply equally well to the case in which a potential difference V is maintained across two parallel faces of the hypercube Λ_L used in the definition of the effective conductivity, and a current I flows between the parallel faces.

THE SERIES AND PARALLEL BOUNDS. For a random resistor network on the simple hypercubic lattice $\mathbb{Z}^d$ with independent, identically distributed bond conductances, the conductance C_L between two parallel faces of the hypercube Λ_L satisfies the inequalities

$$L^{-1}(L+1)^{d-1}\langle G^{-1}\rangle^{-1} \le \langle C_L\rangle \le L^{-1}(L+1)^{d-1}\langle G\rangle, \tag{41}$$

where the random variable G corresponds to the conductance of any one bond of the lattice.

Hence $\langle L^{2-d} C_L\rangle$ is bounded above as $L \to \infty$ if the individual bond conductances G have finite mean, and is also bounded away from zero if the individual bond resistances $(1/G)$ have finite mean. Under these conditions, if σ_{eff} exists, then from Eq. (40), we have

$$\langle G^{-1}\rangle^{-1} \le \sigma_{\text{eff}} \le \langle G\rangle . \tag{42}$$

Hammersley [65] has used Dirichlet's Principle to derive some stronger results than the series and parallel bounds (42) for independent conductances. Chayes and Chayes [30] used variational principles to derive a not directly useful lower bound on the conductivity for the standard percolation conduction problem, and the following more encouraging upper bound.

UPPER BOUND OF CHAYES AND CHAYES. For the standard bond percolation conduction problem on the d-dimensional hypercubic lattice, with open bonds having conductance g_0, the effective conductivity $\sigma_{\text{eff}}(p)$ satisfies the inequality

$$\sigma_{\text{eff}}(p) \le g_0 \Pr\{\text{a given bond is part of the backbone}\} . \tag{43}$$

The *conductivity exponent* for the percolation conduction problem is defined by writing

$$\sigma_{\text{eff}}(p) \propto (p - p_c)^t \quad \text{as } p \downarrow p_c . \tag{44}$$

There is no rigorous proof, even in two dimensions, of the existence of the critical exponent, even in the weak sense that $t = \lim_{p\downarrow p_c} \ln[\sigma_{\text{eff}}(p)]/\ln(p - p_c)$. However, from the result (v) in Sect. "Rigorous General Results" and the upper bound of Chayes and Chayes we know that the conductivity is identically zero below the percolation threshold, as one's intuition demands, and if the accepted critical exponents do exist then we have the inequality

$$t \ge \beta_{\text{bb}} \ge 2\beta , \tag{45}$$

where β_{bb} is the backbone exponent and β is the exponent for the percolation probability $P_\infty(p)$. From numerical estimates $\beta_{\text{bb}} \approx 0.52$ in two dimensions, and we know that $2\beta = \frac{5}{18} \approx 0.28$, while $t > 1$ (see Sect. "Numerical Estimates of the Conductivity Critical Exponent t"), so neither equality is very sharp.

Scaling Theory for the Conductivity Straley [158] developed a heuristic scaling argument to describe the conductivity of a random resistor network for which the probability density function for the individual bond conductances is

$$f(g) = (1 - p)\delta(g - a) + p\delta(g - b) , \tag{46}$$

where $a \ll b$. The effective conductivity σ_{eff} changes in the obvious manner when the units of conductivity or length change. A prefactor μ in the conductivity can be regarded as accounting for this. Straley proposed that σ_{eff} is a homogeneous function near the percolation threshold:

$$\sigma_{\text{eff}}(p) \approx \mu S([p - p_c]\lambda^{-1}, a\mu^{-1}\lambda^{-A}, \mu b^{-1}\lambda^{-B}) , \quad (47)$$

with the approximation assumed valid when each argument of S is less than unity. The function S is singular when any of its arguments vanishes. The parameter λ embodies the interrelations via scaling of the parameters a, b and $p - p_c$. Setting $a = 0$, $\lambda = |p - p_c|$ and $\mu = \lambda^B b$ gives

$$\sigma_{\text{eff}}(p) \approx |p - p_c|^B bS(\operatorname{sgn}[p - p_c], 0, 1) . \quad (48)$$

This case describes the conductivity of the classical percolation conduction problem, so we are led to the conclusion that $S(-1, 0, 1) = 0$ and $S(1, 0, 1) > 0$, and we identify B with the conductivity exponent t. If instead we set $1/b = 0$, $\lambda = |p - p_c|$ and $\mu = \lambda^{-A} a$, we obtain

$$\sigma_{\text{eff}}(p) \approx |p - p_c|^{-A} aS(\operatorname{sgn}[p - p_c], 1, 0) . \quad (49)$$

This case describes the superconductivity problem in which a fraction $1 - p$ of the bonds are normal conductors, while the remaining bonds have infinite conductance, so we conclude that $S(-1, 1, 0) < \infty$ and $S(1, 1, 0) = \infty$ and we identify A with the Superconductivity exponent s. Thus

$$\sigma_{\text{eff}}(p) \approx \mu S([p - p_c]\lambda^{-1}, a\mu^{-1}\lambda^{-s}, \mu b^{-1}\lambda^{-t}) . \quad (50)$$

If we now take $p = p_c$ and $0 < a < b < \infty$, then setting $\lambda^{t+s} = a/b$ and $\mu = b\lambda^t$, we obtain

$$\sigma_{\text{eff}}(p_c) \approx a^u b^{1-u} S(0, 1, 1) , \quad (51)$$

where the exponent u is given by

$$u = t/(s + t) . \quad (52)$$

Special Results for two Dimensions Each two-dimensional lattice L with well-defined faces (polygons with the bonds as sides, not crossed by any bonds) is associated with a *dual lattice* L^D, obtained by placing a site of L^D in every face of L and joining sites of L^D by bonds if the original faces of L share a common bond in L. The square lattice is its own dual lattice; the triangular lattice and the hexagonal lattice form a dual pair. Duality is central to the exact determination of percolation thresholds in two dimensions ► Percolation Thresholds, Exact, but also has useful implications for the random resistor problem [17,101,160], which parallel the phase interchange relations for two-dimensional continua (Sect. "Effective Conductivities"). For the square lattice, if the probability density function for bond conductances is

$$f(g) = (1 - p)\delta(g - a) + p\delta(g - b) , \quad (53)$$

then the effective conductivity $\sigma_{\text{eff}}(p)$ satisfies the equation

$$\sigma_{\text{eff}}(p)\sigma_{\text{eff}}(1 - p) = ab , \quad (54)$$

and in particular $\sigma_{\text{eff}}(\frac{1}{2}) = \sqrt{ab}$. In the notation of Straley's scaling theory (Sect. "Scaling Theory for the Conductivity"), this implies that $u = \frac{1}{2}$ for the square lattice, and so $s = t$ for the square lattice. Although this does not prove rigorously that the conductivity exponent t and superconductivity exponent s coincide for the square lattice, it is taken as evidence for the claim that $s = t$ for all reasonable two-dimensional lattices. Numerical evidence is strongly against this result remaining true in higher dimensions.

It can also be shown from duality that for the square lattice, if

$$f(g) = \frac{1}{gD\sqrt{2\pi}} \exp\left[-\frac{(\log g - \log g_0)^2}{2D^2}\right] , \quad (55)$$

then $\sigma_{\text{eff}} = g_0$.

Effective Medium Approximations The effective medium ideas of Sect. "Approximations" were adapted to the estimation of the conductivity of random resistor networks on periodic lattices by Kirkpatrick [88,89]. A much fuller modern account than given here of the implementation of effective medium ideas for lattice systems and the application of the ideas to more subtle problems and in different contexts, will be found in Sahimi [140].

The random network is replaced by a network that is uniform with unknown bond conductance g_*, except for one special bond, which has conductance G. Let the lattice have coordination number z. It can be shown that the average fluctuation in current or potential difference across the special bond due to its differing in conductance from g_* vanishes provided that

$$\left\langle \frac{G - g_*}{g_* + (2/z)(G - g_*)} \right\rangle = 0 , \quad (56)$$

and this gives a self-consistent approximate determination of g_*. For the square and simple cubic lattices, g_* is an approximation for σ_{eff}. For other lattices, σ_{eff} is a constant multiple of g_*. In the case of a percolative conductance distribution, the approximation produces an estimate of the percolation threshold.

For the standard percolation conduction problem (32), where a fraction p of the bonds has nonzero conductance g_0, the effective medium approximation produces $g_*/g_0 \approx (p - 2/z)/(1 - 2/z)$ and as bond conductances must be non-negative this prediction is interpreted as meaning that

$$\frac{g_*}{g_0} \approx \begin{cases} 0, & p \le 2/z, \\ \dfrac{p - 2/z}{1 - 2/z}, & p > 2/z . \end{cases} \tag{57}$$

That is, the effective medium approximation predicts that the percolation threshold is $2/z$. This is fortuitously exact for the square lattice bond problem and reasonable for the triangular bond and honeycomb bond problems (▶ Percolation Thresholds, Exact), but works less well in three dimensions. The construction of effective medium approximations for site percolation problems is more delicate [75].

The major qualitative deficiency of the effective medium approximation is the conductivity exponent prediction: it asserts that $t = 1$ for all dimensions. This is already a poor approximation in two dimensions ($t \approx 1.3$—see Table 2) and is grossly misleading in three dimensions ($t \approx 2$—see Table 3). The effective medium approximation also predicts that the critical exponent is unchanged if the standard percolation conduction problem is replaced by the more general model (31), provided that the mean resistance of conducting bonds is finite, that is,

$$\int_0^\infty g^{-1} h(g) \mathrm{d}g < \infty . \tag{58}$$

When the conditional probability density function for the conductance of bonds of nonzero conductance has the asymptotic form $h(g) \propto g^{-\alpha}$ as $g \downarrow 0$ with $0 < \alpha < 1$, so that the condition (58) is violated, the effective medium approximation predicts that $t = 1/(1 - \alpha)$ [75,92].

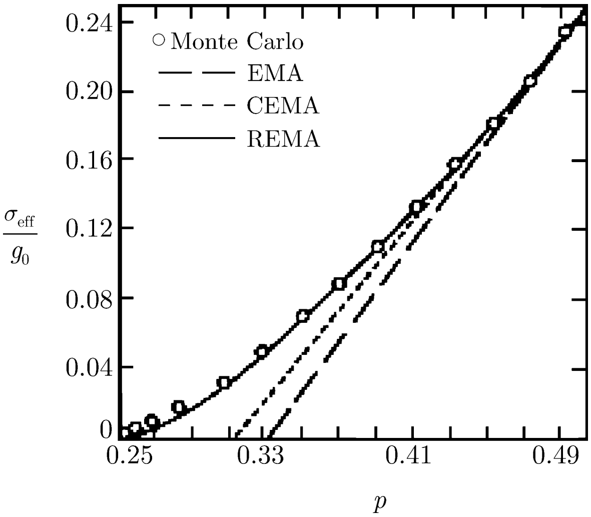
The effective medium approximation works well in non-percolative systems, and in percolative systems has some use when p is close to 1.

Renormalization By analogy with statistical mechanics applications, any process by which a lattice system is replaced by a similar lattice with bonds of different lengths is called (*real-space* or *position-space*) *renormalization*. Such a process can be performed exactly on some self-similar pseudolattices, but not on standard periodic lattices.

Numerical estimates of the effective conductivity based on simulation of finite lattice fragments often use an algorithm in which a finite subnetwork is replaced by a single effective bond for which the conductance may be calculated exactly [32,57,97]. This is especially effective near the percolation threshold, and is sometimes loosely described as an exact renormalization procedure [32]. The name is a little misleading, since the new random resistor network cannot generally be embedded in a natural way within the original underlying lattice structure.

There was considerable interest in real-space renormalization in the 1970s and 1980s, as it gave approximate predictions for critical exponents that differ from the so-called mean-field (high dimension) values of the geometrical exponents of percolation theory, and from the effective medium predictions of the conductivity exponent. Like the effective medium approximation, real-space renormalization is an uncontrolled approximation for which a priori estimates of the quality of the approximation are not available. For detailed accounts of real-space renormalization ideas applied to percolation and conduction, see Sahimi [140] and ▶ Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation.

The effect of renormalization is to reduce the correlation length in the system, moving the system away from the percolation threshold. Sahimi et al. [144] have given an improved, though still uncontrolled approximation, in which approximate renormalization is used to map the system to parameter values where the effective medium approximation is more accurate. This renormalized effective medium approximation effectively increases the interval in which the effective medium approximation works, but like all effective medium treatments cannot accurately



Conduction and Diffusion in Percolating Systems, Figure 2
Effective conductivity of the simple cubic lattice bond problem for the standard percolation conduction model (32). Monte Carlo simulation data (*circles* [89] give clear evidence that the conductivity exponent *t* exceeds 1. The curves represent uncontrolled approximations: EMA, simple effective medium approximation [89]; CEMA, cluster effective medium approximation [1]; REMA, hybrid theory (renormalized effective medium approximation) [144]. Figure reproduced from [144]

portray the ultimate asymptotic behavior as $p \downarrow p_c$. An example of the resulting approximation is given in Fig. 2, together with the simple effective medium approximation [89], an alternative improved effective medium approximation [1] and simulation data [89]. The reader should note from the simulation data, which the approximate curve of Sahimi et al. matches well except very close to the threshold, the clear nonlinear behavior of the conductivity, in striking contrast to the global linear behavior predicted by the simple effective medium theory and the ultimate linear behavior also predicted by the improved approximations near their predicted approximate percolation thresholds.

Numerical Estimates of the Conductivity Exponent t
Although it is possible to estimate the conductivity exponent t from an approximate determination of $\sigma_{\text{eff}}(p)$ on an interval including p_c, greater computational efficiency comes from the extension of the finite scaling ideas of Fisher [49] to the percolation conduction problem [32,57,97,112,143].

We explain the argument for the hypercubic lattice $\mathbb{Z}^d$, but the argument should apply for all periodic lattices. As the percolation threshold is approached, the correlation length $\xi(p)$ in the infinite lattice diverges:

$$\xi(p) \propto |p - p_c|^{-\nu} \,. \tag{59}$$

Although in principle the critical exponents for the correlation length for $p > p_c$ (ν') and for $p < p_c$ (ν) could differ, they have been proved rigorously to coincide for site percolation on the triangular lattice and there is no evidence to support their being different for other lattices. The argument is therefore written out in terms of ν.

Consider the finite hypercube $\Lambda_L \subset \mathbb{Z}^d$ introduced in Sect. "Defining the Conductivity". The finite lattice Λ_L will represent the infinite lattice to a good approximation provided that $\xi(p) \ll L$. We assume that the conductivity $\sigma_L(p)$ of the finite lattice depends on $|p - p_c|$ only through the ratio $\xi(p)/L \propto (L|p - p_c|^{\nu})^{-1}$, and we write

$$\sigma_L(p) \approx L^{-\kappa} f(L|p - p_c|^{\nu}) \,. \tag{60}$$

Since the conductivity is nonzero at $p = p_c$ for some realizations of the finite lattice, $f(0) \neq 0$. We assume that $f(u) \propto u^{\lambda}$ as $u \to \infty$ so that $\sigma_L(p) \propto L^{-\kappa}[L|p - p_c|^{\nu}]^{\lambda}$. To recover the appropriate asymptotic law (44) for the conductivity of the infinite lattice, that is, $\sigma_{\text{eff}} \propto (p - p_c)^t$ as $p \to p_c^+$, we require $\lambda = \kappa$ (to cancel out the L-dependence), and to obtain the correct dependence on $p - p_c$, we need $t = \nu\lambda$. Hence

$$\sigma_L(p) \approx L^{-t/\nu} f(L|p - p_c|^{\nu}) \tag{61}$$

and in particular

$$\sigma_L(p_c) \approx L^{-t/\nu} f(0) \,. \tag{62}$$

The best available estimates of t come from the use of this asymptotic relation to determine t/ν and then using the exact value $\nu = \frac{4}{3}$ for $d = 2$ or independent numerical estimates of ν for $d \geq 3$.

In comparison with estimating the purely geometric or topological parameters of a percolation model, conductivity estimates are comparatively expensive in terms of computer time. There is a trade-off between having high precision in the values of the average conductivity of the finite lattice and having large enough L values to use the asymptotic relation (62). The most precise estimates of t/ν presently available come from Monte Carlo simulations, in which a random subset of the possible realizations of the lattice is generated for a sequence of values of L. Although the difference equations for the potential can be solved by standard techniques of numerical linear algebra, it is more efficient to perform a sequence of transformations that replace locally complicated small sections of the conducting backbone of the system by equivalent single bonds [32,57,97]; this is especially effective in two dimensions. An alternative technique for two-dimensional systems can be devised using transfer matrices, with the conductivity of one realization of an infinite strip of width L being estimated for a sequence of values of L [38,39], and a generalization of this method that works in higher dimensions is also available [28,72]. Other efficient methods of solving the relevant linear equations are also known [91].

Finite-size scaling arguments also apply to the superconductivity problem in which a fraction p of the bonds have infinite conductance, and one finds that

$$\sigma_L(p_c) \propto L^{s/\nu} \quad \text{as } L \to \infty \,. \tag{63}$$

A selection of the better estimates for t/ν and s/ν and corresponding predictions for t and s are given in Table 2 ($d = 2$) and Table 3 ($d = 3$), together with a few estimates not derived via finite-size scaling. A number of plausible conjectures on relations between various exponents, based on heuristic arguments or phenomenological relations observed from early numerical estimates, are now comprehensively refuted [57,75]. Grassberger [57] has carefully assessed the relative merits of different schemes for estimating t/ν in two dimensions, and notes major issues with the quality of random number generators used for some simulations, and problems arising from inappropriate assumptions concerning correction to scaling terms for the

Conduction and Diffusion in Percolating Systems, Table 2

Estimates of the conductivity (t) and superconductivity (s) exponents in two dimensions. Where values of t/ν or s/ν are given, these values were obtained first by finite-size scaling or related ideas, and the values of t or s were subsequently deduced. In two dimensions it is known that $\nu = \frac{4}{3}$ exactly (asterisked entries use this value to compute t or s from t/ν or s/ν), and also that $s = t$ exactly, but no such results are available for $d = 3$

t/ν	t	Source
0.95 ± 0.05		Real metal-insulator mixtures [123]
0.95 ± 0.01	1.28 ± 0.03	Transfer matrix [38]
0.968 ± 0.005	$\approx 1.291^*$	transfer matrix (bond) [170]
0.970 ± 0.009	$\approx 1.293^*$	Enumerate random walks on backbone [73]
≈ 0.972	$\approx 1.296^*$	Simulation of random walks [131]
$0.973^{+0.005}_{-0.003}$	$1.297^{+0.007}_{-0.004}$	Finite-size scaling [98]
0.975 ± 0.005	$\approx 1.300^*$	transfer matrix (site)[170]
0.979 ± 0.006	$\approx 1.305^*$	finite-size scaling [132]
	1.31 ± 0.04	Monte Carlo [50]
0.9825 ± 0.0008	$1.3100 \pm 0.0011^*$	Finite-size scaling (bond and site) [57]
s/ν	s	Source
0.9745 ± 0.0015	1.299 ± 0.002	Special purpose computer [118]
0.977 ± 0.010	$\approx 1.303^*$	Transfer matrix [72]

Conduction and Diffusion in Percolating Systems, Table 3

Estimates of the conductivity (t) and superconductivity (s) exponents in three dimensions. Where values of t/ν or s/ν are given, these values were obtained first by finite-size scaling or related ideas, and the values of t or s were subsequently deduced. Neither ν nor p_c are known exactly and estimates are sensitive to choices made by the cited authors. All estimates are for the simple cubic lattice, except for the random walk estimate [136] for the simple cubic lattice site problem

t/ν	ν used	t	Source
2.095 ± 0.016	0.89 ± 0.01	1.867 ± 0.035	Finite-size scaling [143]
2.26 ± 0.04			Special purpose computer [120]
2.276 ± 0.012	0.88 ± 0.02	2.003 ± 0.047	Finite-size scaling [55]
2.282 ± 005			Current distribution moments [8]
2.305 ± 0.015	≈ 0.88	≈ 2.0	Finite-size scaling [32]
		2.02 ± 0.02	Monte Carlo as $p \rightarrow p_c$ [32]
≈ 2.315			Generalized transfer matrix [28]
2.48 ± 0.07			Random walks [136]
s/ν	ν used	s	Source
0.782 ± 0.019			Finite-size scaling [137]
0.85 ± 0.04	≈ 0.88	≈ 0.75	Tranfer matrix [72]
0.835 ± 0.005			Special purpose computer [119]

square lattice bond problem. The square lattice site problem, and the bond problems on the triangular and honeycomb lattices, are better behaved [57,133]. Early estimates of t and s by finite-size scaling for lattices with unknown percolation thresholds are partly compromised by inaccuracy in numerical estimates of thresholds, but in some cases numerical estimates of thresholds are now available to extravagant precision (e. g., square lattice site problem, $p_c = 0.592\,746\,5 \pm 0.000\,000\,4$; simple cubic lattice site problem, $p_c = 0.311\,607\,7 \pm 0.000\,000\,4$ [37]).

A weak inequality relating the critical exponents t and ν can be deduced on the basis of finite-size scaling. For finite fragments of the square lattice, a duality argument shows that there is probability $\frac{1}{2}$ that the fragment is spanned by a conducting path at $p = \frac{1}{2}$. Each such realization of the system gives a conductance between the connected sides of the lattice which is no smaller than aL^{-2}, this worst case arising when there is a single path passing through all lattice sites. It follows that for the square lattice, $\sigma_L(p_c) \geq (a/2)ML^{-2}$ and so $t \leq 2\nu$ in two dimensions.

Conduction on Fractals Some progress has been made for percolation conduction problems on self-similar fractal structures [100], which have been proposed as possible models for the backbone of large clusters at the percolation threshold [53]. The plane Sierpinski lattice (triangular gasket) has percolation threshold $p_c = 1$, but unlike the linear chain which also has $p_c = 1$, the Sierpinski lattice has nontrivial finite-size scaling behavior, and the analogue of t/ν is $\log(5/3)/\log(2) \approx 0.73$ [53]. The analogue of s/ν has been evaluated numerically for this system as 0.27 ± 0.03 [162]. For a further discussion of fractal structures in the context of percolation see ▶ Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation.

Structural Speculations

Despite recent spectacular advances in the understanding of two-dimensional percolation mentioned in Sect. "Rigorous Results for two Dimensions", an adequate theory for the conductivity exponent t and superconductivity exponent s remains elusive, even in two dimensions. A number of attempts have been made to model the structure of the backbone just above the percolation, and to model what physicists have called the "incipient infinite cluster" at the percolation threshold p_c, in the hope of predicting the values of t, s and related "dynamical" exponents of percolating systems.

The Backbone The backbone in a percolating system includes all sites with two independent connections to infinity. Only the backbone can carry current in the random resistor problem, although because of local symmetries, not all bonds of the backbone will carry nonzero current. On periodic lattices, the existence of a well-defined nonzero backbone probability $B(p)$ for $p > p_c$ ensures (as a consequence of ergodic theory [18]) a well-defined density of backbone sites, and so the number $B_L(p)$ of sites with two disjoint connections to the boundary in a hypercube of side L is asymptotically a multiple of $L^d B(p)$ as $L \to \infty$. Natural finite-size scaling analysis [155] (cf. Sect. "Numerical Estimates of the Conductivity Critical Exponent t" above) based on the functional form $B_L(p) = L^\kappa b(L|p-p_c|^\nu)$ with $b(z) \propto z^\lambda$ as $z \to \infty$ and the requirement that $B(p) \propto (p-p_c)^{\beta_{bb}}$ as $p \downarrow p_c$ leads to the conclusion that $d = \kappa + \lambda$ and $\beta_{bb} = \lambda\nu$. Thus, in particular, $B_L(p_c) \propto L^{d_{bb}}$, where $d_{bb} = d - \beta_{bb}/\nu$ is a formula believed to hold by physicists for $d \leq 6$, interpreted by them as the fractal dimension [100] of the backbone at the percolation threshold, and estimated by simulation. Grassberger [57] estimates that $d_{bb} = 1.6432 \pm 0.0008$ from a careful study of the square lattice site and bond percolation problems. The occurrence of this noninteger exponent suggests delicate structure of the backbone just above the percolation threshold.

The first attempts to model the backbone in detail and thereby make predictions about the conductivity exponent t date from 1975 and 1976 and are due to de Gennes [34] and Skal and Shklovskii [147]. In their view, the backbone consisted of a "superlattice" of nodes connected by macrobonds or links. If the node spacing is estimated as $\xi(p)$, the correlation length, and the conductivity of a macrobond is given by $g(p) \propto (p-p_c)^\zeta$ for some exponent ζ, then the effective conductivity is predicted to scale as

$$\sigma_{\text{eff}}(p) \approx \xi(p)^{2-d} g(p) \propto (p-p_c)^{(d-2)\nu+\zeta} , \qquad (64)$$

the factor of $\xi(p)^{2-d}$ being motivated by Eq. (39). This prediction implies that the conductivity exponent is $t = \nu(d-2) + \zeta$, and it was suggested [34] that $\zeta = 1$ in all dimensions. However, Chayes and Chayes [31] proved rigorously that $\zeta > \nu$ in all dimensions, implying that $t > \nu(d-1)$. For $d = 2$, this gives $t > \frac{4}{3}$, a result which is no longer credible given the precision recently attained in numerical estimates and the simple superlattice model is unviable.

Problems with the superlattice model were already noted by Kirkpatrick [90] in 1978. He proposed instead a self-similar fractal model, and this idea was further pursued by Gefen et al. [53]. An implication of the self-similar fractal model that $t \leq (d-1)\nu$ is consistent with numerical data in low dimensions, but fails for $d = 6$, where it is accepted that $\nu = \frac{1}{2}$ and $t = 3$.

The most plausible picture of the backbone structure just above the percolation threshold is inelegantly described as the "nodes–links–blobs" model [151,152], with approximately self-similar structures over smallest scales linked together in some way by tortuous, quasi-one-dimensional links but the structure is far from completely characterized.

The Incipient Infinite Cluster Since the percolation probability $P_\infty(p)$ is proven rigorously to be zero at the percolation threshold for several two-dimensional percolation problems, there is with probability 1 no infinite cluster present at the percolation threshold. The *incipient infinite cluster* at the percolation threshold in the sense originally used by physicists is therefore a dubious construct. Kesten has given a rigorous discussion of two mathematically respectable candidates for the incipient infinite cluster [59,86], using appropriately defined conditional probabilities.

(i) Take $p > p_c$ and work with conditional probabilities, conditioning on the event that the origin is part of an infinite cluster. Then take the limit $p \downarrow p_c$.

(ii) Take $p = p_c$ and work with conditional probabilities, conditioning on the event that the origin belongs to a connected cluster that reaches the boundary of a box of side length $2n$ centered on the origin. Then take the limit as $n \to \infty$.

For bond percolation on the square lattice, these two apparently different definitions have been proved consistent [86]. Moreover, for this specific problem, it can be shown that the expected number of sites of the incipient infinite cluster in a box of side $2n$ is (to within a slowly-varying prefactor) $n^{2-\eta}$, where the exponent η governs the decay of the probability that the origin is connected to a box of side $2n$. It has been proved that for the triangular lattice bond problem, $\eta = \frac{5}{24}$ [148], and this should be generally true for two dimensions. In physicists' terminology, this shows that the fractal dimension of the incipient infinite cluster in two dimensions is $d_{\text{iic}} = \frac{43}{24} \approx 1.792$, which is outside the 1983 estimates $d_{\text{iic}} = 1.90 \pm 0.1$ [130] and $d_{\text{iic}} = 1.900 \pm 0.009$ [81].

The theory of the incipient infinite cluster and its relation to finite-size scaling considerations are completely resolved in two dimensions [23,76]. Some illuminating if less complete conclusions are available in higher dimensions [23]. Physicists' scaling theories suggest that $d_{\text{iic}} = d - \beta/\nu$ for $2 \le d \le 6$ [155]. However, it is not the incipient infinite cluster but rather the backbone that is core to the determination of the conductivity exponent, so that these results do not go far enough for the present purposes.

Random Motion in a Random Environment

One may replace the steady-state transport and conduction processes discussed in Sect. "Continuum Models: Steady-State Phenomena" and "Lattice Models: Steady-State Phenomena" for continua and lattices respectively by time-dependent analogues. Our discussion of the continuum version is very brief.

Continuum Problems

As noted in Sect. "Caveats to the Electrical Interpretation", problems of diffusion in random media are easily defined. An account of the homogenization of the diffusion Eq. (5), including references to earlier work, will be found in Jikov et al. [79]. In particular for random media subject a uniform ellipticity condition (equivalent to bounding the local diffusivity away from zero) and with a spatially stationary distribution of the local diffusivity, the system can be homogenized for almost all realizations of the random diffusivity distribution, and the corresponding homogenized diffusivity is independent of the realization. Percolative problems, where the diffusivity is zero at a given arbitrary point with nonzero probability are more difficult to analyze, but Jikov et al. [79] give a proof that for the random chessboard (squares with diffusivities 0 or 1 with probabilities $1 - p$ and p respectively), the scaled displacement $\mathbf{X}(t)$ of a diffusion process commencing in the conducting phase converges in distribution to a zero mean, isotropic bivariate normal random variable. If D_0 denotes the homogenized diffusivity for the steady-state problem treated like a steady conduction problem, then the diffusivity governing the time-dependent problem is $D_0/P_\infty(p)$, where $P_\infty(p)$ is the density of the infinite conducting cluster (and indeed equal to the percolation probability for site percolation on the square lattice).

For a discussion of many aspects and applications of diffusion in condensed matter physics, an area that has sustained interest in the properties of heterogeneous and randomly microstructured materials for a long time, see the collection edited by Heitjans and Kärger [71]. One interesting diffusion problem that we do not address, since it does not manifest percolation properties, arises when one phase acts as an irreversible trap. The survival probability diffusing objects in the presence of dilute random traps decays as $\exp(-kt^{3/5})$ at large times t in three dimensions rather than the naively expected $\exp(-kt)$, due to the unexpected statistical importance of rare, large trap-free regions [58,74,80].

Lattice Problems

The lattice-based models for stochastic transport processes in random environments that we address may loosely be described as *random walks*, since they derive conceptually from the random walk problems first posed by in 1905 by Pearson [126] and in 1919 by Pólya [128,129], but the precise meaning of the phrase *random walk* varies widely in the literature. The classic account of the traditional theory of discrete time, translationally invariant random walks on $\mathbb{Z}^d$ is Spitzer [150]; substantial extensions to the theory needed for more general problems will be found in Doyle and Snell [40], Hughes [74,75], Telcs [163] and Woess [169]. The most recent comprehensive rigorous reviews of random walks in random environments are the papers by Zeitouni [172,173]. For other discussions of random walks in random environments see the books by Hughes [75] and Révész [134].

Nearest-Neighbour Stepping Discrete-Time Random Walk We consider the case in which the probability that a walker presently at site $\mathbf{s}'$ next steps to site $\mathbf{s}$ is $p(\mathbf{s}|\mathbf{s}')$, nonzero only when $\mathbf{s}$ is a nearest neighbor of $\mathbf{s}'$. The walk is called *recurrent* if there is probability 1 that the walker eventually returns to the starting site and *transient* otherwise. A recurrent walker is certain to visit all sites accessible to it, and visits its own starting site infinitely often (with probability 1). A transient walker revisits the starting site at most finitely many times.

The walk is of Pólya type if $p(\mathbf{s}|\mathbf{s}') = 1/z(\mathbf{s}')$, where $z(\mathbf{s}')$ is the coordination number of $\mathbf{s}'$, that is, all nearest-neighbor steps are equally likely. Pólya's most famous result is that for Pólya walk on $\mathbb{Z}^d$, the walk is recurrent for $d = 1$ and $d = 2$, but transient for $d \geq 3$. Translationally invariant walks are well-suited to Fourier analysis. This is how much of the classical theory is derived, and the discrete time evolution is simplified by generating function methods [74,150]. In particular, for Pólya walks on $\mathbb{Z}^d$, the position distribution of the random walker is asymptotically Gaussian (normal) at large times and the probability $p_n(\mathbf{s}_0)$ that the walker will be found at the starting site $\mathbf{s}_0$ after n steps decays in proportion to $n^{-d/2}$.

The number S_n of distinct sites visited in the first n steps is of interest in a number of contexts [74]; in the probability literature this quantity is sometimes misleadingly called the range of the random walk. For classical Pólya walks on $\mathbb{Z}^d$, the mean number of distinct sites visited has the asymptotic behavior [115]

$$\langle S_n \rangle \sim \begin{cases} (8n/\pi)^{1/2}, & d = 1\,, \\ \pi n/\ln n, & d = 2\,, \\ (1-R)n, & d \geq 3\,, \end{cases} \tag{65}$$

where R is the probability of eventual return to the starting site. The following more general results hold [74] for translationally invariant walks on periodic lattices. For transient walks, $\langle S_n \rangle \sim (1-R)n$. For recurrent walks, if $p_n(\mathbf{s}_0) \sim \text{constant} \times n^{-H/2}$ with $0 < H < 2$, or equivalently,

$$\sum_{n=0}^{\infty} p_n(\mathbf{0})\xi^n \sim \text{constant} \times (1-\xi)^{H/2-1} \quad \text{as } \xi \uparrow 1\,, \tag{66}$$

then

$$\langle S_n \rangle \sim \text{constant} \times n^{H/2}\,. \tag{67}$$

The case $H = 2$ induces logarithmic prefactors. The exponent H is variously described as the *harmonic dimension* [74], *spectral dimension* or *fraction dimension* [2], and $H = d$ for the simplest problems on $\mathbb{Z}$ or $\mathbb{Z}^2$.

For random walk problems where the generating function and Fourier analysis techniques used to derive the preceding results are not available, the important questions to addressed in other ways are as follows.

(i) Is the walk recurrent or transient?
(ii) Is the walk diffusive (large-n Gaussian or normal limiting behavior) and if so, what is the value of diffusion constant?
(iii) In non-diffusive cases, how does the mean-square displacement grow with n?
(iv) What is the value of the harmonic dimension?

Canonical Models for Random Environments Random environments for random walkers to experience may be constructed in various ways. One simple way is to assign to each bond of the lattice $\mathbb{Z}^d$ (or some other periodic lattice) an independent nonzero weight, which might be interpreted as the conductance of a random resistor. If $G_{\mathbf{r},\mathbf{s}} = G_{\mathbf{r},\mathbf{s}}$ denotes the conductance of a bond joining nearest-neighbor sites $\mathbf{r}$ and $\mathbf{s}$, then the transition probability may be defined to be

$$p(\mathbf{s}|\mathbf{s}') = \frac{G_{\mathbf{s},\mathbf{s}'}}{\sum_{\mathbf{s}''} G_{\mathbf{s}'',\mathbf{s}'}}\,, \tag{68}$$

the sum over $\mathbf{s}''$ being restricted to nearest neighbors of $\mathbf{s}'$. If all bonds emanating from a given site have equal conductance, the walk is of Pólya type.

Consider the following two problems on the infinite conducting lattice just introduced:

- a discrete-time random walk, governed by (68) starting from site $\mathbf{s}_0$;
- electrical conduction between site $\mathbf{s}_0$ and infinity.

Then it can be shown [40] that the walk is transient if and only if the total resistance between $\mathbf{s}_0$ and infinity is finite. This result, and various other relations between random walks and electric networks [40,75,117], represent one of the most important avenues for progress in the study of random walks in random environments.

Pemantle and Peres [127] have shown that for independent, identically distributed conductances, the question of recurrence or transience of this model can be settled (for all realizations of the environment save for a set of zero probability) by an analysis of Pólya random walk on the random lattice arising from the bond percolation on the same lattice. The original walk is transient if and only if for some p with $p_c < p < 1$ the Pólya walk on the infinite cluster in the bond percolation problem is transient.

Master Equations and Generalized Master Equations Differential-difference equations of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}c(\mathbf{s},t) = \sum_{\mathbf{s}'} W(\mathbf{s},\mathbf{s}')c(\mathbf{s}',t) - M(\mathbf{s})c(\mathbf{s},t) \qquad (69)$$

describe a process in which the concentration of a mobile substance at site $\mathbf{s}$ evolves by transfer of substance from other sites (gaining substance at rate $W(\mathbf{s},\mathbf{s}')c(\mathbf{s}',t)$ from site $\mathbf{s}'$) and loss of substance to other sites (total rate of loss $M(\mathbf{s})c(\mathbf{s},t)$ to all other sites). If the substance is to be globally conserved, one needs

$$M(\mathbf{s}) = \sum_{\mathbf{s}'} W(\mathbf{s}',\mathbf{s}) \qquad (70)$$

and if

$$p(\mathbf{s},t) = \frac{c(\mathbf{s},t)}{\sum_{\mathbf{s}'} c(\mathbf{s}',t)} \qquad (71)$$

denotes the normalized concentration then what physicists call the *master equation* [122],

$$\frac{\mathrm{d}}{\mathrm{d}t}p(\mathbf{s},t) = \sum_{\mathbf{s}'}[W(\mathbf{s},\mathbf{s}')p(\mathbf{s}',t) - W(\mathbf{s}',\mathbf{s})p(\mathbf{s},t)]\,, \qquad (72)$$

results. This equation is interpreted as the law governing the probability of the position of some type of random walk process in continuous time, and indeed is a standard equation from the theory of Markov processes.

To obtain a richer class of possible behaviors, physicists have also considered *generalized master equations* [84], where the transition rates $W(\mathbf{s},\mathbf{s}')$ are replaced by memory kernels $W(\mathbf{s},\mathbf{s}',t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}p(\mathbf{s},t) = \int_0^t \sum_{\mathbf{s}'}[W(\mathbf{s},\mathbf{s}',t-t')p(\mathbf{s}',t') - W(\mathbf{s}',\mathbf{s},t-t')p(\mathbf{s},t')]\mathrm{d}t'\,. \qquad (73)$$

For Eqs. (72) and(73) the transition rates or memory kernels are often nonzero only for nearest-neighbor sites. Simple models of random media arise by imposing symmetry,

$$W(\mathbf{s},\mathbf{s}') = W(\mathbf{s}',\mathbf{s}) \quad \text{or} \quad W(\mathbf{s},\mathbf{s}',t) = W(\mathbf{s}',\mathbf{s}',t),$$

and making the rate coefficients $W(\mathbf{s},\mathbf{s}')$ or parameters in the memory kernels $W(\mathbf{s},\mathbf{s}',t)$ independent random variables. If we introduce Laplace transforms, writing

$$\widehat{f}(u) = \int_0^\infty e^{-ut} f(t)\mathrm{d}t\,, \qquad (74)$$

then the transformed master and generalized master equations for a process started at the origin site $\mathbf{0}$ become

$$u\widehat{p}(\mathbf{s},u) - \delta_{\mathbf{s},\mathbf{0}} = \sum_{\mathbf{s}'} W(\mathbf{s},\mathbf{s}')[\widehat{p}(\mathbf{s}',u) - \widehat{p}(\mathbf{s},u)], \qquad (75)$$

$$u\widehat{p}(\mathbf{s},u) - \delta_{\mathbf{s},\mathbf{0}} = \sum_{\mathbf{s}'} \widehat{W}(\mathbf{s},\mathbf{s}',u)[\widehat{p}(\mathbf{s}',u) - \widehat{p}(\mathbf{s},u)]. \qquad (76)$$

In both cases, the Laplace transform equations are able to be interpreted as the equations governing a random resistor network, with all sites having an additional connection to zero potential (an earth connection) of conductance u. If we attempt to match Eq. (75), with random rates $W(\mathbf{s},\mathbf{s}')$, to an equivalent uniform system with rates W_*, then it should be anticipated that W_* is a function of the variable u, since the relative importance of the earth connection at $\mathbf{s}$ compared to its connections to other sites will fluctuate over the lattice. Consequently, as discussed in Sect. "Exactly and Approximately Solved Continuous-Time Problems", the appropriate real-time description of the system may be expected to correspond to a uniform memory kernel generalized master equation, rather than a simple, uniform transition rate master equation.

Exactly Solved Discrete-Time Problems Elegant exact results are scarce for random processes in random environments, with the strongest and most informative results limited to one dimension (that is, on $\mathbb{Z}$), where percolative problems become essentially trival: for $p < 1$ a Pólya walker is confined to a finite interval and essentially its position has a limiting distribution (to be more precise, the positions after even or odd numbers of steps have limiting distributions, which may be different). Averaging over all realizations of the environment we find that the mean-square displacement after n steps is $o(n)$. This unusual behavior is an artifact of averaging: as n increases, in more and more realizations the walker's displacements have saturated and only the effects of increasingly rare large intervals contribute to the continued growth of the mean-square displacement, and various properties of the system can be deduced by appropriate averaging over the positions of the left and right boundary sites [121].

A less trivial one-dimensional model introduced in 1972 by Temkin [164] brings out the subtle possibilities for random walks in random environments. Consider the stepping law

$$p(l|l') = A_{l'}\delta_{l,l'+1} + (1 - A_{l'})\delta_{l,l'-1}\,, \qquad (77)$$

where $\{A_l\}$ is a set of independent, identically distributed random variables. Let an overbar denote an average over all realizations of the set $\{A_l\}$. Solomon [149] has shown that: a walker initially stepping right is certain to return if and only if $\overline{\log[(1-A)/A]} \geq 0$; a walker initially stepping left is certain to return if and only if $\overline{\log[A/(1-A)]} \geq 0$, and the walk is recurrent if and only if $\overline{\log[A/(1-A)]} = 0$. In 1982, Sinai [146] produced the astonishing result that in the recurrent case, on the additional assumption that A_l is bounded away from 0 and 1, the mean-square displacement is $O(\ln^4 n)$, in place of the standard $O(n)$ for one-dimensional Pólya walkers.

Exactly and Approximately Solved Continuous-Time Problems Alexander et al. [3] have discussed exactly solvable one-dimensional master equations with independent, identically distributed random transition rates. If the probability density function for the rates is denoted by $f(w)$, they show that the mean-square displacement grows as $2W_0 t$, and the random system is equivalent in its long-time properties to a uniform system in which all transition rates are W_0, where

$$\frac{1}{W_0} = \int_0^\infty \frac{f(w)\mathrm{d}w}{w} \,, \tag{78}$$

so long as the integral on the right is finite. When this is not the case, the random motion is *sub-diffusive* : if $f(w) \to f(0) > 0$ as $w \to 0$ then the mean-square displacement grows as $t/\ln t$, while if $f(w) \propto w^{-\alpha}$ as $w \to 0$ the mean-square displacement grows as $t^{(2-2\alpha)/(2-\alpha)}$ (distribution-induced nonuniversality). There are several of other approaches to exactly solvable one-dimensional master equation problems, including asymmetric rate problems [75].

A number of authors have extended the idea of an effective medium approximation from the random resistor problem (Sect. "Effective Medium Approximations") or other contexts to the problem of random master equations. For historical details and a full account of the analysis see Sahimi et al. [142] or Hughes [74]. As noted in Sect. "Master Equations and Generalized Master Equations", in the Laplace transform domain, the master equation with random coefficients is equivalent to a random resistor network, and the uniform transition rate produced by the matching procedure is a function of the transform variable, so that the approximately equivalent uniform system is governed by a generalized master equation. In the percolative case in which the rate coefficient associated with a bond is nonzero only with probability $p < 1$, the approximately equivalent uniform system does not support motion if $p < 2/z$ (the predicted percolation threshold, where z is the coordination number of the underlying lattice on which the percolation process is realized). For $p \downarrow 2/z$, the effective diffusion constant $D(p)$ (a constant multiple of the time derivative of the mean-square displacement) is predicted to vanish linearly with $p - 2/z$ for the standard case where all nonzero rate coefficients are equal, but the exponent is predicted to change in more general cases if the average of W^{-1} is infinite, where W is the random rate associated with an arbitrary bond, conditional on the rate being nonzero.

In one dimension, the effective medium approximation reproduces a number of the exact results of Alexander et al. [3]. Its predictions for higher-dimensional systems, while not of great accuracy, do have one interesting aspect. Generalized master equations with memory kernels $W(\mathbf{s}, \mathbf{s}', t) = \phi(t)p(\mathbf{s}|\mathbf{s}')$ are naturally associated with continuous-time random walk processes in which the walker waits for a random time with probability density function $\psi(t)$ between successive steps, and then moves with the stepping law $p(\mathbf{s}|\mathbf{s}')$. The classic exponential waiting time density $\psi(t) = \kappa \exp(-\kappa t)$ arises if and only $\phi(t) = \kappa\delta_+(t)$ and the generalized master equation reduces to the ordinary master equation in this case. In general, the connection between ϕ and ψ is most simply expressed in terms of the Laplace transform: $\widehat{\psi}(u) = \widehat{\phi}(u)/[u + \widehat{\phi}(u)]$ [74,85]. Any form of nondiffusive behavior that may arise is naturally associated with a non-exponential waiting time density, and for $p < 2/z \approx p_c$ the associated waiting time is predicted to be defective, that is,

$$\int_0^\infty \psi(t)\mathrm{d}t < 1 \tag{79}$$

and the equivalent continuous time random walker takes only finitely many steps. An alternative approach to modeling nondiffusive behavior of possible relevance to percolative systems can be developed using fractional calculus ideas [107,108].

The Ant in the Labyrinth

We conclude our survey of transport and conduction in random environments, with a discussion for discrete-random walks of the problem of the ant in the labyrinth, introduced in 1975 by Brandt [24] and proposed by de Gennes [33] as a probe of the geometry of percolation. The ant, a random walker, moves through a random labyrinth generated by applying site or bond percolation to a periodic lattice [24,33,75,110,111,113]. Because the analogue of a strong ellipticity condition is not satisfied for these percolative problems, many of the more general results from rigorous analyzes of random walks in random environments [172,173] do not apply.

Models Let the original lattice from which the labyrinth is made have coordination number z. Then at a site $\mathbf{s}$ of the labyrinth, the coordination number will be $Z_\omega(\mathbf{s}) \leq z$, where $Z_\omega(\mathbf{s})$ is determined by the specific realization ω of site or bond percolation used in constructing the labyrinth. There are two canonical choices for the ant's mode of stepping through the labyrinth, the distinction between them having been clearly drawn by Mitescu and Roussenq [111] and Majid et al. [99].

myopic ant (Pólya case) —On arrival at site $\mathbf{s}$, the ant looks at the $Z_\omega(\mathbf{s})$ adjacent sites onto which it is permitted to step, assigns each of them probability $1/Z_\omega(\mathbf{s})$ and chooses one of them at random.

blind ant —On arrival at site $\mathbf{s}$, the ant attempts to move to one of the z adjacent sites on the original lattice. If this move is not allowed for the labyrinth, the ant remains at site $\mathbf{s}$.

In a given time interval, a blind ant visits fewer distinct sites than a myopic ant.

It is generally believed that the qualitative properties of myopic (Pólya) ants and blind ants are similar, but there is an important difference when the ant is introduced on a finite cluster [111]: the "equilibrium" probability distributions of the ant, approached asymptotically as the number of steps grows without bound, differ for the two models. The existence of equilibrium probability distributions on finite clusters can be discussed from the point of view of Markov chains [46], where such distributions are called "invariant measures".

Simulations suggest [41,113,125,145] that when the mean-square displacement $\langle R_n^2 \rangle$ is averaged over environments, as $n \to \infty$ we have

$$\overline{\langle R_n^2(p) \rangle} = \begin{cases} R_\infty^2 - A(p)\exp\{[-n/\theta(p)]^w\} + \cdots\,, & p < p_c\,, \\ Bn^{2k} + \cdots\,, & p = p_c\,, \\ CD(p)n + \cdots\,, & p > p_c\,, \end{cases} \tag{80}$$

where any dependence of a quantity on p or n is explicitly indicated. Here $D(p)$ is an effective diffusion constant and the exponent k, which would be 1/2 for classical diffusion, is dimension-dependent, with $k < 1/2$ in low-dimensional systems. In particular, estimates of k for the simple cubic lattice are 0.20 ± 0.01 [125] and 0.200 ± 0.002 [41]. The value of the exponent w was initially believed close to 1 [113] in three dimensions, but later evidence [125] suggests that $w \approx 0.4$ in three dimensions.

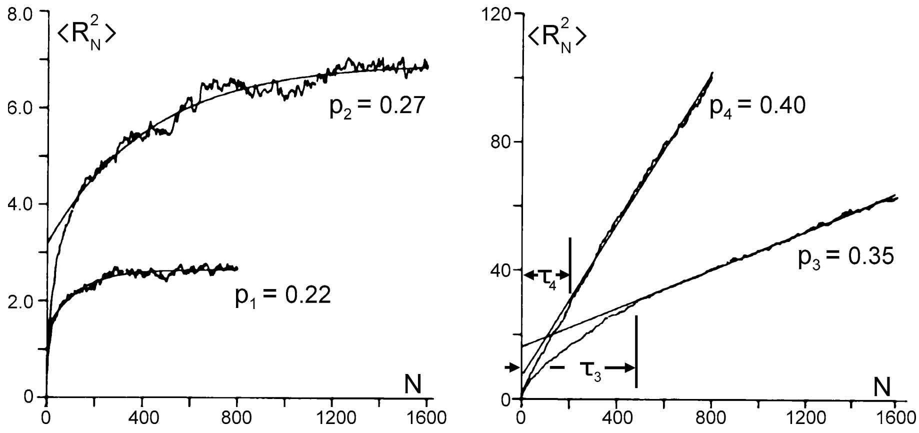
In the mathematical literature, careful distinctions are drawn between results that apply to the *annealed* case (corresponding to averaging over all realizations of the disordered system) and the *quenched* case (involving statements about realizations ω of the disorder. For $p \leq p_c$ the ant commences with probability 1 on a cluster of finite size, and in any individual realization ω of the system, the mean-square displacement will converge to a finite value (blind case) or oscillate between two finite values (myopic case for some lattices, where oscillations can occur if the cluster divides into sites accessible only after an even number of steps and sites accessible only after an odd number of steps). For $p = p_c$, at a given value of n the displacement will have saturated (that is, be close to its asymptotic value) in some environments but not in others. This partitioning of the realizations of the environments, and the tortuous structure of large clusters that are not yet fully explored by the ant, give rise to the nonclassical exponent $k < 1/2$.

Scaling Theory for the Ant in the Labyrinth Heuristic work [9,69,75,124,155] raises attractive possible connections between exponents characterizing the asymptotic behavior of the ant in the labyrinth and the geometrical exponents of percolation theory. A number of rigorous scaling relations involving various exponents that characterize random walk and electrical conduction on lattices have been derived by Telcs [163], but these are not sufficiently informative to establish rigorously the results now briefly described or to enable exponent values to be predicted.

We shall denote averages for walks, conditioned on the walks taking place on clusters of m sites, by $\langle \cdot \rangle_m$. Consider first the expected number of distinct sites visited. At the percolation threshold, if there is some kind of effective harmonic dimension for large clusters (cf. Sect. "Lattice Problems"), which one might loosely call the harmonic dimension of the incipient cluster (cf. Sect. "The Incipient Infinite Cluster") and denote by H_{iic}, then the expected number of distinct sites visited by a walker on a cluster of m sites should evolve as $\langle S_n \rangle_m \propto n^{H_{\mathrm{iic}}/2}$ for $1 \ll \langle S_n \rangle_m \ll m$. The scaling hypothesis [124] $\langle S_n \rangle_m \approx n^{H_{\mathrm{iic}}/2}\phi(m/n^{H_{\mathrm{iic}}/2})$ and the asymptotic law $P_m(p) \propto m^{-1-1/\delta}$ for the cluster size distribution gives a simple prediction [6,124] for the mean number of distinct sites visited, averaged both over cluster sizes and over walks on individual clusters:

$$\overline{\langle S_n \rangle} \approx \sum_{m=1}^{\infty} m^{-1-1/\delta}\, n^{H_{\mathrm{iic}}/2}\phi(m/n^{H_{\mathrm{iic}}/2}) \approx n^{(1-1/\delta)H_{\mathrm{iic}}/2}\,; \tag{81}$$

the last result arises from approximating the sum by an integral. Three-dimensional simulations give $\overline{\langle S_n \rangle} \propto n^{0.54 \pm 0.02}$ [124].



Conduction and Diffusion in Percolating Systems, Figure 3

Simulations of the ant in the labyrinth [113]: myopic ants (Pólya walkers) on the simple cubic lattice under site percolation ($p_c \approx 0.31$) for four values of p. Simulation data (*erratic curves*) shows the mean-square displacement after N steps, averaged over realizations of the environment when a fraction p of all sites are occupied: a shows $p < p_c$ ($p_1 = 0.22$, $p_2 = 0.27$), with the smooth curves least-squares fits to $R_\infty^2 - A(p)\exp[-n/\theta(p)]$; b shows $p > p_c$ ($p_1 = 0.35$, $p_4 = 0.40$), with the straight-line asymptotes corresponding to effectively diffusive behaviour and τ_k corresponding to the number of steps needed for diffusive behaviour to be manifest for $p = p_k$ (Figure reproduced with permission from [114])

A similar analysis may be performed for the mean-square displacement at the percolation threshold [54,75]. If the fractal dimension of the incipient infinite cluster is denoted by d_{iic} (cf. Sect. "The Incipient Infinite Cluster") one may propose that the mean-square displacement on clusters of size m is $\langle R_n^2 \rangle_m \propto n^{2\nu_{\text{iic}}}$ for $1 \ll \langle R_n^2 \rangle_m^{1/2} \ll m^{1/d_{\text{iic}}}$. The naturally associated scaling assumption $\langle R_n^2 \rangle_m \approx n^{2\nu_{\text{iic}}} \psi(m^{1/d_{\text{iic}}}/n^{2\nu_{\text{iic}}})$ predicts the mean-square displacement averaged both over cluster sizes and over walks on individual clusters to be

$$\begin{aligned}\overline{\langle R_n^2 \rangle} &\approx \sum_{m=1}^{\infty} m^{-1-1/\delta} n^{2\nu_{\text{iic}}} \psi(m^{1/d_{\text{iic}}}/n^{2\nu_{\text{iic}}}) \\ &\approx n^{2\nu_{\text{iic}}[1-d_{\text{iic}}/(2\delta)]} .\end{aligned} \tag{82}$$

The accepted scaling relations $d_{\text{iic}} = (\beta + \gamma)/\nu$ and $\gamma = \beta(\delta - 1)$ give the alternative form $\overline{\langle R_n^2 \rangle} \propto n^{2\nu_{\text{iic}}[1-\beta/(2\nu)]}$ and simulation data is consistent with this picture [68].

A conjecture of Alexander and Orbach [2] published in 1982 suggesting that $H_{\text{iic}} = 4/3$ for all dimensions d briefly raised the possibility that the conductivity exponent t could be simply related to geometrical exponents in percolation theory, and that other dynamical exponents related to random walk processes could be determined. Given the known geometrical exponents in two dimensions, the Alexander–Orbach conjecture predicted that $t = 91/72 \approx 1.2639$ in two dimensions. It is now generally accepted that the Alexander–Orbach conjecture, while an excellent approximation, is not precisely correct [9,116].

A number of other observable quantities can be estimated by scaling arguments [9,75,155]. For example, just below the percolation threshold, it is predicted [154] that the mean-square displacement (averaged over all walks and realizations) has the limiting value $R_\infty^2 \propto (p_c - p)^{2\nu-\beta}$.

Rigorous Results for Discrete Time The most important rigorous result is probably the following theorem, which shows that Pólya's simple criterion for transience $d \geq 3$ is not destroyed by the percolation process, provided the system is above the percolation threshold, and the ant starts on the infinite cluster [60].

Theorem 1 (Theorem of Grimmett, Kesten and Zhang) *For myopic ants (Pólya walkers) walking on the infinite cluster generated by bond percolation on the simple hypercubic lattice at $p > p_c$, the walk is transient with probability 1 if $d \geq 3$.*

As with most rigorous results in this area, the proof is of antisocial length. However, Rayleigh's Monotonicity Law and the correspondence between random walks and electric circuits makes it relatively straightforward to show that the recurrence of the ordinary Pólya walk on the square lattice $\mathbb{Z}^2$ ensures the recurrence of myopic ant motion on every realization of an infinite cluster in bond or site percolation on the square lattice.

For walks on the incipient infinite cluster, a nebulous concept made precise by either of the equivalent approaches of Kesten discussed in Sect. “The Incipient Infinite Cluster”, it is known [86] that the general behavior is subdiffusive ($|\mathbf{X}_n| = O(n^{1/2-\epsilon})$ for some $\epsilon > 0$) for the square lattice bond problem, but the exact exponent characterizing the growth of $|\mathbf{X}_n|$ is unknown.

Rigorous Results for Continuous Time For the natural master-equation analogue of the myopic ant (corresponding to a myopic ant executing a continuous-time random walk with an exponential waiting time density of mean waiting time 1 at all sites) several exact results that mirror known results for Pólya walks on periodic lattices are available. In stating these results, we assume that our labyrinth is generated by bond percolation on $\mathbb{Z}^d$ with $d \geq 2$ and $p > p_c$, we write C^ω_∞ to denote the (unique with probability 1) infinite cluster in the realization ω of the random environment, and we assign coordinates such that the the origin $\mathbf{0}$ is a site of the infinite cluster. We write $\Pr^\omega$ to denote probabilities associated with random walks in the fixed realization ω of the environment. The position at time t of a random walker starting from the origin at time 0 is $\mathbf{X}_t$. We say that a result holds for almost all environments if it holds with probability 1, where probability is measured with respect to the realizations of bond percolation for the given value of p.

The following results, published in 2004 [7,103], go further than results previously established only for $d = 2$ [35,36] and show that there is no anomalous behavior above the percolation threshold for the myopic ant problem for all $d \geq 2$: in the long time limit, the process strongly resembles classical diffusion.

Mathieu and Remy [103] have proved that there exists $c_1(p, d)$ (independent of time or the realization ω of the environment) such that for almost all environments, the quenched bound

$$\sup_{\mathbf{y} \in C^\omega_\infty} \Pr{}^\omega\{\mathbf{X}_t = \mathbf{y}\} \leq \frac{c_1(p,d)}{t^{d/2}} \tag{83}$$

holds. This result confirms the transience of the walk for $d \geq 2$, a result already known from the theorem of Grimmett, Kesten and Zhang [60] for the discrete-time walk, stated in Sect. “Rigorous Results for Discrete Time”. Mathieu and Remy also establish the analogous result to (83) for site percolation on the two-dimensional square lattice, and an annealed lower bound proportional to $t^{-d/2}$ for bond percolation on $\mathbb{Z}^d$.

Barlow [7] has constructed upper and lower bounds for the quenched transition density

$$q^\omega_t(\mathbf{x}, \mathbf{y}) = \frac{1}{Z^\omega(\mathbf{y})} \Pr{}^\omega\{X_{t+\tau} = \mathbf{y} \,|\, X_\tau = \mathbf{x}\} \tag{84}$$

that hold with probability 1 for sufficiently large time t:

$$\begin{aligned} &\frac{c_1(d,p)}{t^{d/2}} \exp\left[-\frac{c_2(d,p)}{t}|\mathbf{x}-\mathbf{y}|_1^2\right] \leq q^\omega_t(\mathbf{x},\mathbf{y}) \\ &\leq \frac{c_3(d,p)}{t^{d/2}} \exp\left[-\frac{c_4(d,p)}{t}|\mathbf{x}-\mathbf{y}|_1^2\right] . \end{aligned} \tag{85}$$

Here if $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_d)$, one defines

$$|\mathbf{x}-\mathbf{y}|_1 = \sum_{k=1}^{d} |x_k - y_k| \, . \tag{86}$$

Sufficiently large t is quantified by $t \geq \sup\{|\mathbf{x}-\mathbf{y}|_1, S^\omega_\mathbf{x}\}$, where the finite but realization-dependent quantities $S^\omega_\mathbf{x}$ deal with problems associated with an initial period of potentially anomalous behavior. Barlow’s results establish the essentially diffusive asymptotic behavior of the walk. For further results in the quenched case, see Berger and Biskup [13].

The above results pertain only to walks of Polyá type on the infinite cluster for $p > p_c$. For a discussion of biased random walks on the infinite cluster in two-dimensional percolation, see Berger et al. [14].

Future Directions

Problems of transport and conduction in random systems subject to the kinds of uniform ellipticity restrictions that preclude percolation phenomena have become increasingly well understood, though the address by Zeitouni [171] at the 2002 International Congress of Mathematicians summarizes a number of areas in which progress on the problem of random walk in a random environment remains to be made. Moreover, an adequate theoretical understanding does not entirely remove the tedious problem of accurately estimating overall system properties.

The situation concerning truly percolative problems remains less satisfactory. The generally classical behavior of such systems above the percolation threshold means that, for modeling purposes in science and technology, one often may reasonably replace the random system by a uniform analogue, in much the same way as one can do this for systems with ellipticity restrictions, but with the important caveat that close to the percolation threshold, the length or time scales at which the uniform treatment is accurate may in practice be unacceptably large. A fuller understanding therefore of processes at or near the percolation threshold remains the greatest challenge.

There has been spectacular progress since the late 1970s with the understanding (both heuristic and rigorous) of the geometrical side of percolation theory, especially for lattices. We have elegant results on the existence of classical or mean-field behavior for large dimension d, and the knowledge (though gaps in rigor remain) that mean-field behavior applies for $d \geq 7$ and in a slightly weaker sense for $d = 6$ but not for smaller d. More importantly, through rigorous work on conformal invariance and on Schramm–Loewner evolution processes, exact two-dimensional values of geometrical critical exponents and exact two-dimensional scaling relations are now properly established. An analogous rigorous account of dynamical exponents characterizing transport in two-dimensional percolating systems is perhaps not too much to ask for, though the way is not yet clear. It would be surprising indeed if there were significant progress in the short term on the rigorous analysis of three-dimensional systems at or near the percolation threshold.

Bibliography

1. Ahmed G, Blackman JA (1979) On theories of transport in disordered media. J Phys C 12:837–853
2. Alexander S, Orbach R (1982) Density of states on fractals: 'fractons'. J Phys Lett 43:L625–L631
3. Alexander S, Bernasconi J, Schneider WR, Orbach R (1981) Excitation dynamics in random one-dimensional systems. Rev Mod Phys 53:175–198
4. Allaire G (1992) Homogenization and two-scale convergence. SIAM J Math Anal 23:1482–1518
5. Anderson PW (1958) Absence of diffusion in certain random lattices. Phys Rev 109:1492–1505
6. Angles d'Auriac JC, Rammal R (1983) Scaling analysis for random walk properties on percolation clusters. J Phys C 16:L825–L830
7. Barlow MT (2004) Random walks on supercritical percolation clusters. Ann Probab 32:3024–3084
8. Batrouni GG, Hansen A, Larson B (1996) Current distribution in the three-dimensional random resistor network at the percolation threshold. Phys Rev E 53:2292–2297
9. Ben-Avraham D, Havlin S (2000) Diffusion and Reactions in Fractals and Disordered Systems. Cambridge University Press, Cambridge
10. Bensoussan A, Lions J-L, Papanicolaou G (1978) Asymptotic Analysis for Periodic Structures. North–Holland, Amsterdam
11. Beran MJ (1965) Statistical Continuum Theories. Trans Soc Rheol 9:339–355
12. Beran MJ (1968) Statistical Continuum Theories. Wiley, New York
13. Berger N, Biskup M (2007) Quenched invariance principle for simple random walk on percolation clusters. Probab Theory Relat Fields 137:83–120
14. Berger N, Gantert N, Peres Y (2007) The speed of biased random walk on percolation clusters. arXiv:math/0211303v3; revised version of (2003) Probab Theory Relat Fields 126:221–242
15. Bergman DJ (1978) Analytical properties of the complex effective dielectric constant of a composite medium with applications to the derivation of rigorous bounds and to percolation problems. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 46–61
16. Berlyand L, Golden K (1994) Exact result for the effective conductivity of a continuum percolation model. Phys Rev B 50:2114–2117
17. Bernasconi J, Schneider WR, Weismann HJ (1977) Some rigorous results for random planar conductance networks. Phys Rev B 16:5250–5255
18. Billingsley P (1965) Ergodic Theory and Information. Wiley, New York
19. Blumenfeld R, Meir Y, Aharony A, Harris AB (1987) Resistance fluctuations in randomly diluted networks. Phys Rev B 35:3524–3535
20. Blumenfeld R, Meir Y, Harris AB, Aharony A (1986) Infinite set of exponents describing physics on fractal networks. J Phys A 19:L791–L796
21. Bollobas B, Riordan O (2006) Percolation. Cambridge University Press, Cambridge
22. Borgs C, Chayes JT, Kesten H, Spencer J (1999) Uniform boundedness of critical crossing probabilities implies hyperscaling. Random Struct Algorithms 15:368–413
23. Borgs C, Chayes JT, Kesten H, Spencer J (2001) The birth of the infinite cluster: finite-size scaling in percolation. Commun Math Phys 224:153–204
24. Brandt WW (1975) Use of percolation theory to estimate effective diffusion coefficients of particles migrating on various ordered lattices and in a random network structure. J Chem Phys 63:5162–5167
25. Broadbent SR, Hammersley JM (1957) Percolation processes. I Crystals and mazes. Proc Camb Philos Soc 53:629–641
26. Brown WF (1955) Solid mixture permittivities. J Chem Phys 23:1514–1517
27. Bruggeman DAG (1935) Berechnung verschiedener physikalischer Konstanten von heterogen Substanzen. I Dielektrizitätskonstanten und Leitfähigkeiten der Mischkörper aus isotropen Substanzen. Annalen Phys (Leipzig) 24:636–679
28. Byshkin MS, Turkin AA (2005) A new method for the calculation of the conductivity of inhomogeneous systems. J Phys A 38:5057–5067
29. Calvert B, Keady G (1993) Braess's paradox and power-law nonlinearities in networks. J Aust Math Soc B 35:1–22
30. Chayes JT, Chayes L (1986) Bulk transport properties and exponent inequalities for random resistor and flow networks. Commun Math Phys 105:133–152
31. Chayes JT, Chayes L (1987) On the upper critical dimension of Bernoulli percolation. Commun Math Phys 113:27–48
32. Clerc JP, Podolskiy VA, Sarychev AK (2000) Precise determination of the conductivity exponent of 3D percolation using exact numerical renormalization. Eur Phys J B 15:507–516
33. de Gennes PG (1976) La percolation: un concept unificateur. Rech 7:919–927
34. de Gennes PG (1976) On a relation between percolation theory and the elasticity of gels. J Phys Lett 37:L1–L2
35. de Masi A, Ferrari PA, Goldstein S, Wick WD (1985) Invariance principle for reversible Markov processes with application to diffusion in the percolation regime. In: Durrett R (ed) Particle Systems, Random Media and Large Deviations, Contem-

porary Mathematics, Vol 41. American Mathematical Society, Providence, pp 71–85

36. de Masi A, Ferrari PA, Goldstein S, Wick WD (1989) An invariance principle for reversible Markov processes. Applications to random motions in random environments. J Stat Phys 55:787–855
37. Deng Y, Blöte HWJ (2005) Monte Carlo study of the site-percolation model in two and three dimensions. Phys Rev B 72:016126
38. Derrida B, Vannimenus J (1982) A transfer matrix approach to random resistor networks. J Phys A 15:L557–L564
39. Derrida B, Zabolitzky JG, Vannimenus J, Stauffer D (1984) A transfer matrix program to calculate the conductivity of random resistor networks. J Stat Phys 36:31–42
40. Doyle PG, Snell JL (1984) Random Walks and Electric Networks, Carus Mathematical Monograph No. 22. Mathematical Association of America, Washington
41. Duering E, Roman HE (1991) Corrections to scaling for diffusion exponents on three-dimensional percolation systems at criticality. J Stat Phys 64:851–858
42. Eggarter TP, Cohen MH (1970) Simple model for density of states and mobility of an electron in a gas of hard-core scatterers. Phys Rev Lett 25:807–810
43. Einstein A (1906) Eine neue Bestimmung der Moleküldimensionen. Annalen Phys 19:289–306
44. Einstein A (1911) Berichtigung zu meiner Arbeit: 'Eine neue Bestimmung der Moleküldimensionen'. Annalen Phys 34:591–592
45. Fatt I (1956) The network model of porous media [in 3 parts]: I—Capillary pressure characteristics; II—Dynamic properties of a single size tube network; III—Dynamic properties of networks with tube radius distribution. Trans Am Inst Min Metall Petroleum Eng, Petroleum Branch 207:144–159, 160–163, 164–177
46. Feller W (1970) An Introduction to Probability Theory and its Applications, vol 1, 3rd edn. Wiley, New York
47. Feller W (1971) An Introduction to Probability Theory and its Applications, vol 2, 2nd edition. Wiley, New York
48. Feng S, Halperin BI, Sen PN (1987) Transport properties of continuum systems near the percolation threshold. Phys Rev B 35:197–214
49. Fisher ME (1971) The theory of critical point singularities. In: Green MS (ed) Critical Phenomena: Enrico Fermi Summer School. Academic Press, New York, pp 1–99
50. Fogelholm R (1980) The conductivity of large percolation network samples. J Phys C 13:L571–L574
51. Fortuin CM (1972) On the random cluster model. II The percolation model. Physica 58:393–418
52. Fortuin CM, Kasteleyn PW (1972) On the random cluster model. I Introduction and relation to other models. Physica 57:536–564
53. Gefen Y, Aharony A, Mandelbrot BB, Kirkpatrick S (1981) Solvable fractal family and its possible relation to the backbone at percolation. Phys Rev Lett 47:1771–1774
54. Gefen Y, Aharony A, Alexander S (1983) Anomalous diffusion on percolating clusters. Phys Rev Lett 50:77–80
55. Gingold DB, Lobb CJ (1990) Percolative conduction in three dimensions. Phys Rev B 42:8220–8224
56. Golden K, Papanicolaou G (1983) Bounds for effective parameters of heterogeneous media by analytic continuation. Commun Math Phys 90:473–491
57. Grassberger P (1999) Conductivity exponent and backbone dimension in 2-d percolation. Physica A 262:251–263
58. Grassberger P, Procaccia I (1982) The long-time properties of diffusion in a medium with static traps. J Chem Phys 77:6281–6284
59. Grimmett G (1999) Percolation, 2nd edition. Springer, Berlin
60. Grimmett GR, Kesten H, Zhang Y (1993) Random walk on the infinite cluster of the percolation model. Probab Theory Relat Fields 96:33–44
61. Gu GQ, Yu KW (1992) Effective conductivity of nonlinear composites. Phys Rev B 46:4502–4507
62. Hammersley JM (1957) Percolation processes. II. The connective constant. Proc Camb Philos Soc 53:642–645
63. Hammersley JM (1957) Percolation processes. Lower bounds for the critical probability. Ann Math Stat 28:791–795
64. Hammersley JM (1961) Comparison of atom and bond percolation processes. J Math Phys 2:728–733
65. Hammersley JM (1988) Mesoadditive processes and the specific conductivity of lattices. J Appl Probab, Special vol 25A, edited by Gani J, 347–358
66. Hara T, Slade G (1994) Mean-field behaviour and the lace expansion. In: Grimmett G (ed) Probability and Phase Transition. Kluwer, Dordrecht, pp 87–122
67. Hashin Z, Shtrikman S (1962) A variational approach to the theory of the effective magnetic permeability of multiphase materials. J Appl Phys 33:3125–3131
68. Havlin S and Ben-Avraham D (1983) Diffusion and fracton dimensionality on fractals and on percolation clusters. J Phys A 16:L483–L487
69. Havlin S, Ben-Avraham D, Sompolinsky H (1983) Scaling behavior of diffusion on percolation clusters. Phys Rev A 27:1730–1733
70. Heinrichs J, Kumar N (1975) Simple exact treatment of conductance in a random Bethe lattice. J Phys C 8:L510–L516
71. Heitjans P, Kärger J (eds)(2005) Diffusion in Condensed Matter: Methods, Material, Models. Springer, Berlin
72. Herrmann HJ, Derrida B, Vannimenus J (1984) Superconductivity exponents in two- and three-dimensional percolation. Phys Rev B 30:4080–4082
73. Hong DC, Havlin S, Herrmann HJ, Stanley HE (1984) Breakdown of the Alexander–Orbach conjecture for percolation: exact enumeration of random walks on percolation backbones. Phys Rev B 30:4083–4086
74. Hughes BD (1995) Random Walks and Random Environments, vol 1: Random Walks. Clarendon Press, Oxford
75. Hughes BD (1996) Random Walks and Random Environments, vol 2: Random Environments. Clarendon Press, Oxford
76. Járai AA (2003) Incipient infinite percolation clusters in 2D. Ann Probab 31:444–485
77. Jerauld GR, Hatfield JC, Scriven LE, Davis HT (1984) Percolation and conduction on Voronoi and triangular networks: a case study in topological disorder. J Phys C 17:1519–1529
78. Jerauld GR, Scriven LE, Davis HT (1984) Percolation and conduction on the 3D Voronoi and regular networks: a second case study in topological disorder. J Phys C 17:3429–3439
79. Jikov VV, Kozlov SM, Oleinik OA (1994) Homogenization of Differential Operators and Integral Functionals. Springer, Berlin
80. Kayser RF, Hubbard JB (1983) Diffusion in a medium with a random distribution of static traps. Phys Rev Lett 51:79–82

81. Kapitulnik A, Aharony A, Deutscher G, Stauffer D (1983) Self-similarity and correlations in percolation theory. J Phys A 16:L269–L274
82. Keller JB (1964) A theorem on the conductivity of a composite medium. J Math Phys 5:548–549
83. Kenkel SW, Straley JP (1982) Percolation theory of nonlinear circuit elements. Phys Rev Lett 49:767–770
84. Kenkre VM (1982) The master equation approach: coherence, energy transfer, annihilation, and relaxation. In: Kenkre VM, Reineker P, Exciton Dynamics in Molecular Crystals and Aggregates. Springer, Berlin, pp 1–109
85. Kenkre VM, Montroll EW, Shlesinger MF (1973) Generalized master equations for continuous-time random walks. J Stat Phys 9:45–50
86. Kesten H (1986) The incipient infinite cluster in two-dimensional percolation. Probab Theory Relat Fields 73:369–394
87. Kim IC, Torquato S (1992) Effective conductivity of suspensions of overlapping spheres. J Appl Phys 71:2727–2735
88. Kirkpatrick S (1971) Classical transport in disordered media: scaling and effective-medium theories. Phys Rev Lett 27:1722–1725
89. Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
90. Kirkpatrick S (1978) The geometry of the percolation threshold. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 99–116
91. Knudsen HA, Fazekas S (2006) Robust algorithm for random resistor networks using hierarchical domain structure. J Comput Phys 211:700–718
92. Kogut PM, Straley JP (1979) Distribution-induced non-universality of the percolation conductivity exponents. J Phys C 12:2151–2159
93. Kozlov SM (1989) Geometric aspects of averaging. Russ Math Surv 44(2):91–144
94. Landauer R (1978) Electrical conductivity in inhomogeneous media. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 2–43
95. Last BJ, Thouless DJ (1971) Percolation theory and electrical conductivity. Phys Rev Lett 27:1719–1721
96. Lawler GF, Schramm O, Werner W (2002) One-arm exponent for critical 2D percolation. Electronic Journal of Probability, vol 7, paper 2 (http://www.math.washington.edu/~ejpecp/EjpVol7/paper2.html)
97. Lobb CJ, Frank DJ (1979) Large-cell renormalization group calculation of the percolation conductivity critical exponent. J Phys C 12:L827–L830
98. Lobb CJ, Frank DJ (1984) Percolative conduction and the Alexander–Orbach conjecture in two dimensions. Phys Rev B 30:4090–4092
99. Majid I, Ben-Avraham D, Havlin S, Stanley HE (1984) Exact-enumeration approach to random walks on percolation clusters in two dimensions. Phys Rev B 30:1626–1628
100. Mandelbrot BB (1982) The Fractal Geometry of Nature. W.H. Freeman, San Francisco
101. Marchant J, Gabillard B (1975) Sur le calcul d'un réseau résistif aléatoire. Comptes Rendus Acad Sci (Paris) B 281:261–264
102. Markov KZ (2000) Elementary micromechanics of heterogeneous media. In: Markov K, Preziosi L (eds) Heterogeneous Media: Micromechanics, Modeling, Methods and Simulations. Birkhäuser, Boston, pp 1–62
103. Mathieu P, Remy E (2004) Isoperimetry and heat kernel decay on percolation clusters. Ann Probab 32:100–128
104. Meester R, Roy R (1996) Continuum Percolation. Cambridge University Press, Cambridge
105. Meir Y, Blumenfeld R, Aharony A, Harris AB (1986) Series analysis of randomly diluted nonlinear resistor networks. Phys Rev B 34:3424–3428
106. Men'shikov MV (1986) Coincidence of critical points in percolation problems. Sov Math Dokl 33:856–859
107. Metzler R, Klafter J (2000) The random walker's guide to anomalous diffusion: a fractional dynamics approach. Phys Rep 339:1–77
108. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. J Phys A 37:R161-R208
109. Milton GW (2002) The Theory of Composites. Cambridge University Press, Cambridge
110. Mitescu CD, Roussenq J (1976) Une fourmi dans un labyrinthe: diffusion dans un système de percolation Comptes Rendus de l'Académie des Sciences (Paris) A 283:999–1001
111. Mitescu CD, Roussenq J (1983) Diffusion on percolation clusters. In: Deutscher G, Zallen R, Adler J (eds) Percolation Processes and Structures. Annals of the Israel Physical Society, vol 5. Adam Hilger, Bristol, pp 81–100
112. Mitescu CD, Allain M, Guyon E, Clerc JP (1982) Electrical conductivity of finite-size percolation networks. J Phys A 15:2523–2531
113. Mitescu CD, Ottavi H, Roussenq J (1978) Diffusion on percolation lattices: the labyrinthine ant. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 377–381
114. Mitescu CD, Ottavi H, Roussenq J (1978). In: AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 377–381
115. Montroll EW, Weiss GH (1965) Random walks on lattices. II J Math Phys 6:167–181
116. Nakayama T, Yakubo K, Orbach RL (1994) Dynamical properties of fractal networks: scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
117. Nash-Williams CStJA (1959) Random walks and electric currents in networks. Proc Camb Philos Soc 18:931–958
118. Normand J-M, Herrmann HJ, Hajjar M (1988) Precise calculation of the dynamical exponent of two-dimensional percolation. J Stat Phys 52:441–446
119. Normand J-M, Herrmann HJ (1990) Precise numerical determination of the superconducting exponent of percolation in three dimensions. Intern J Mod Phys C 1:207–214
120. Normand J-M, Herrmann HJ (1995) Precise determination of the conductivity exponent of 3D percolation using "Percola". Intern J Mod Phys C 6:207–214
121. Odagaki T, Lax M (1980) ac hopping conductivity of a one-dimensional bond percolation model. Phys Rev Lett 45:847–850
122. Oppenheim I, Shuler KE, Weiss GH (1977) Stochastic Processes in Chemical Physics: The Master Equation. MIT Press, Cambridge, Massachusetts

123. Palevski A, Deutscher G (1984) Conductivity measurements on a percolation fractal. J Phys A 17:L895–L898

124. Pandey RB, Stauffer D (1983) Fractal dimensionality and number of sites visited of the ant in the labyrinth. J Phys A 16:L511–L513

125. Pandey RB, Stauffer D, Margolina A, Zabolitzky JG (1984) Diffusion on random systems above, below and at their percolation threshold in two and three dimensions. J Stat Phys 34:427–450

126. Pearson K (1905) The problem of the random walk. Nature 72:294

127. Pemantle R, Peres Y (1996) On which graphs are all random walks in random environments transient? In: Aldous D, Pemantle R (eds) Random Discrete Structures, IMA Volumes in Mathematics and Its Applications, No. 76. Springer, New York, pp 207–211

128. Pólya G (1919) Quelques problèmes de probabilité se rapportant à la 'promenade au hasard'. Enseign Math 20:444–445

129. Pólya G (1921) über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. Mathematische Annalen 83:149–160

130. Puech L, Rammal R (1983) Fractal geometry and anomalous diffusion in the backbone of percolation clusters. J Phys C 16:L1179–L1202

131. Rammal R, Angles d'Auriac JC, Benoit A (1984) Universality of the spectral dimension of percolation clusters. Phys Rev B 30:4087–4089

132. Rammal R, Lemieux MA, Tremblay AMS (1985) Comment on 'ε-expansion for the conductivity of a random resistor network'. Phys Rev Lett 54:1087

133. Reš, I (2001) Corrections to scaling for percolative conduction: anomalous behavior at small *L*. Phys Rev B 64:224304

134. Révész P (2005) Random Walk in Random and Non-random Environments. World Scientific, Singapore

135. Rintoul MD, Torquato S (1997) Precise determination of the critical threshold and exponents in a three-dimensional continuum percolation model. J Phys A 30:L585–L592

136. Roman HE (1990) Diffusion in three-dimensional random systems at their percolation threshold. J Stat Phys 58:375–382

137. Sahimi M (1984) Finite-size scaling calculation of conductivity of three-dimensional conductor–superconductor networks at the percolation threshold. J Phys C 17:L355–L358

138. Sahimi M (1994) Applications of Percolation Theory. Taylor and Francis, London

139. Sahimi M (1995) Flow and Transport in Porous Media and Fractured Rock. VCH Verlagsgesellschaft, Weinheim

140. Sahimi M (2003) Heterogeneous Materials, vol 1: Linear Transport and Optical Properties. Springer, New York

141. Sahimi M (2003) Heterogeneous Materials, vol 2: Nonlinear and Breakdown Properties and Atomistic Modeling. Springer, New York

142. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Stochastic transport in disordered systems. J Chem Phys 78:6849–6864

143. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Critical exponent of percolation conductivity by finite-size scaling. J Phys C 16:L521–L527

144. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Real-space renormalization and effective-medium approximation to the percolation conduction problem. Phys Rev B 28:307–311

145. Seifert E, Suessenbach M (1984) Tests of universality for percolative diffusion. J Phys A 17:L703–L708

146. Sinai YG (1982) The limiting behavior of a one-dimensional random walk in a random environment. Theory Probab Appl 27:256–268

147. Skal AS, Shklovskii BI (1975) Topology of an infinite cluster in the theory of percolation and its relationship to the theory of hopping conduction. Sov Phys Semicond 8:1029–1032

148. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. Math Res Lett 8:729–744

149. Solomon F (1975) Random walks in a random environment. Ann Probab 3:1–31

150. Spitzer F (1976) Principles of Random Walk, 2nd edition. Springer, Berlin

151. Stanley HE (1977) Cluster shapes at the percolation threshold: an effective cluster dimensionality and its connection with critical point exponents. J Phys A 10:L211–L220

152. Stanley HE, Coniglio A (1983) Fractal structure of the incipient infinite cluster in percolation. In: Deutscher G, Zallen R, Adler J (eds) Percolation Processes and Structures, Annals of the Israel Physcal Society, vol 5. Adam Hilger, Bristol, pp 101–120

153. Stauffer D (1979) Scaling theory of percolation clusters. Phys Rep 54:1–74

154. Stauffer D (1985) Introduction to Percolation Theory. Taylor and Francis, London

155. Stauffer D, Aharony A (1994) Introduction to Percolation Theory, corrected 2nd edition. Taylor and Francis, London

156. Stinchcombe RB (1973) The branching model for percolation theory and electrical conductivity. J Phys C 6:L1–L5

157. Stinchcombe RB (1974) Conductivity and spin-wave stiffness in disordered systems: an exactly soluble model. J Phys C 7:197–203

158. Straley JP (1976) Critical phenomena in resistor networks. J Phys C 9:783–795

159. Straley JP (1977) Random resistor tree in an applied field. J Phys C 10:3009–3013

160. Straley JP (1977) Critical exponents for the conductivity of random resistor lattices. Phys Rev B 15:5733–5737

161. Straley JP, Kenkel SW (1984) Percolation theory for nonlinear conductors. Phys Rev B 29:6299–6305

162. Taitelbaum H, Havlin S (1988) Superconductivity exponent for the Sierpinski gasket. J Phys A 21:2265–2271

163. Telcs A (2006) The Art of Random Walks. Lecture Notes in Mathematics, vol 1885. Springer, Berlin

164. Temkin DE (1972) One-dimensional random walks in a two-component chain. Sov Math Dokl 13:1172–1176

165. Thorpe MF (1982) Bethe lattices. In: Thorpe MF (ed) Excitations in Disordered Systems. Plenum, New York, pp 85–107

166. Torquato S (2002) Random Heterogeneous Materials: Microstructure and Macroscopic Properties. Springer, New York

167. Torquato S, Kim IC, Kule D (1999) Effective conductivity, dielectric constant, and diffusion coefficient of digitized composite media via first-passage-time equations. J Appl Phys 85:1560–1571

168. Wiener O (1912) Die Theorie des Mischkörpers für das Feld des stationären Strömung. Abh Mathematisch-Phys Kl K Sächs Ges Wiss 32:509–604

169. Woess W (2000) Random Walks on Infinite Graphs and Groups. Cambridge University Press, Cambridge

170. Zabolitzky JG (1984) Monte Carlo evidence against the Alexander–Orbach conjecture for percolation conductivity. Phys Rev B 30:4076–4079

171. Zeitouni O (2002) Random walks in random environments. In: Daquien LI (ed) Proceedings of the International Congress of Mathematicians. vol 3. Higher Education Press, Beijing, pp 117–127
172. Zeitouni O (2004) Random walks in random environment. In: Tavaré S, Zeitouni O (eds) Lectures on Probability and Statistics (Ecole d'Eté de Probabilités de Saint-Flour XXXI), Lecture Notes in Mathematics, vol 1837. Springer, Berlin, pp 1–188
173. Zeitouni O (2006) Random walks in random environments. J Phys A: Math Gen 39:R433–R464
174. Ziman J (1968) The localization of electrons in ordered and disordered systems. I Percolation of classical particles. J Phys C 1:1532–1538

Consciousness and Complexity

ANIL K. SETH[1], GERALD M. EDELMAN[2]
[1] Department of Informatics, University of Sussex, Brighton, UK
[2] The Neurosciences Institute, San Diego, USA

Article Outline

Glossary

Thalamocortical system The network of highly interconnected cortical areas and thalamic nuclei that comprises a large part of the mammalian brain. The cortex is the wrinkled surface of the brain, the thalamus is a small walnut-sized structure at its center. An intact thalamocortical system is essential for normal conscious experience.

Theory of neuronal group selection (TNGS) A large-scale selectionist theory of brain development and function with roots in evolutionary theory and immunology. According to this theory, brain dynamics shape and are shaped by selection among highly variant neuronal populations guided by value or salience.

Neural correlate of consciousness Patterns of activity in brain regions or groups of neurons that have privileged status in the generation of conscious experience. *Explanatory* correlates are neural correlates that in addition account for key properties of consciousness.

Dynamic core A distributed and continually shifting coalesence of patterns of activity among neuronal groups within the thalamocortical system. According to the TNGS, neural dynamics within the core are of high neural complexity by virtue of which they give rise to conscious discriminations.

Neural complexity A measure of simultaneous functional segregation and functional integration based on information theory. A system will have high neural complexity if each of its components can take on many different states and if these states make a difference to the rest of the system.

Small-world networks Networks in which most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps. Small-world networks combine high clustering with short path lengths. They can be readily identified in neuroanatomical data, and they are well suited to generating dynamics of high neural complexity.

Metastability Dynamics that are characterized by segregating and integrating influences in the temporal domain; metastable systems are neither totally stable nor totally unstable.

Definition of the Subject

How do conscious experiences, subjectivity, and apparent free will arise from their biological substrates? In the mid 1600s Descartes formulated this question in a form that has persisted ever since [1]. According to Cartesian dualism consciousness exists in a non-physical mode, raising the difficult question of its relation to physical interactions in the brain, body and environment. Even in the late twentieth century, consciousness was considered by many to be outside the reach of natural science [2], to require strange new physics [3], or even to be beyond human analysis altogether [4]. Over the last decade however, there has been heightened interest in attacking the problem of consciousness through scientific investigation [5,6,7,8,9]. Succeeding in this inquiry stands as a key challenge for twenty-first century science.

Conventional approaches to the neurobiology of consciousness have emphasized the search for so-called 'neural correlates': Activity within brain regions or groups of neurons that has privileged status in the generation of conscious experience [10]. An important outcome of this line of research has been that consciousness is closely tied to neural activity in the thalamocortical system, a network of cortical areas and subcortical nuclei that forms a large part of the vertebrate brain [11,12]. Yet correlations by themselves cannot supply explanations, they can only constrain them. A promising avenue toward explanation is to focus on key properties of conscious experience and to identify neural processes that can account for these properties; we can call these processes *explanatory correlates*. This article clarifies some of the issues surrounding this approach and describes ways of characterizing quantitatively the *complexity* of neural dynamics as a candidate explanatory correlate.

Complexity is a central concept within many branches of systems science and more generally across physics, statistics, and biology; many quantitative measures have been proposed and new candidates appear frequently [13,14,15]. The complexity measures described in this article are distinguished by focusing on the extent to which a system's dynamics are *differentiated* while at the same time *integrated*. This conception of complexity accounts for a fundamental feature of consciousness, namely that every conscious experience is composed of many different distinguishable components (differentiation) and that every conscious experience is a unified whole (integration). According to the theoretical perspective described here, the combination of these features endows consciousness with a discriminatory capability unmatched by any other natural or artificial mechanism.

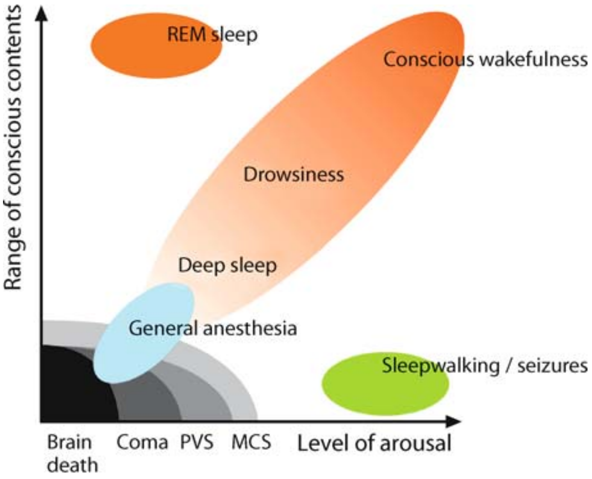
While the present focus is on consciousness and its underlying mechanisms, it is likely that the measures of complexity we describe will find application not only in neuroscience but also in a wide variety of natural and artificial systems.

Introduction

Consciousness

Consciousness is that which is lost when we fall into a dreamless sleep and returns when we wake up again. As William James emphasized, consciousness is a *process* and not a 'thing' [5]. Conscious experiences have content such as colors, shapes, smells, thoughts, emotions, inner speech, and the like, and are commonly accompanied by a sense of self and a subjective perspective on the world (the 'I'). The phenomenal aspects of conscious content (the 'redness' of red, the 'warmth' of heat, etc.) are in philosophical terminology called *qualia* [16].

It is important to distinguish between conscious *level*, which is a position on a scale from brain-death and coma at one end to vivid wakefulness at the other, and conscious *content*, which refers to composition of a conscious scene at a given (non-zero) conscious level. Obviously, conscious level and conscious content are related inasmuch as the range of possible conscious contents increases with conscious level (see Fig. 1). It is also possible to differentiate *primary* (sensory) consciousness from *higher-order* (meta) consciousness [6,17]. Primary consciousness refers to the presence of perceptual conscious content (colors, shapes, odors, etc.). Higher-order consciousness (HOC) refers to the fact that we are usually conscious of being conscious; that is, human conscious contents can refer to ongoing primary conscious experiences. HOC is usually associated with language and an explicit sense of selfhood [18] and good arguments can be made that primary consciousness can exist in principle in the absence of HOC, and that in many animals it probably does [19,20].



Consciousness and Complexity, Figure 1
Conscious level is correlated with the range of possible conscious contents. PVS = persistent vegetative state, MCS = minimally conscious state. Adapted from [21]

Consciousness as Discrimination

There are many aspects of consciousness that require explanation (see Table 1). However, one especially salient aspect that has been too often overlooked is that every conscious scene is both *integrated* and *differentiated* [22]. That is, every conscious scene is experienced 'all of a piece', as unified, yet every conscious scene is also composed of many different parts and is therefore one among a vast repertoire of possible experiences: When you have a particular experience, you are distinguishing it from an enormous number of alternative possibilities. On this view, conscious scenes reflect informative discriminations in a very high dimensional space where the dimensions reflect all the various modalities that comprise a conscious

Consciousness and Complexity, Table 1

Thirteen features of consciousness that require theoretical explanation. Items 1–6 are to some degree open to quantitative measurement whereas items 7–13 are more readily understood through logical and qualitative analysis. This list is drawn from [23] and a related list appears in [17]

1	Consciousness is accompanied by irregular, low-amplitude, fast (12–70 Hz) electrical brain activity.
2	Consciousness is associated with activity within the thalamocortical complex, modulated by activity in subcortical areas.
3	Consciousness involves distributed cortical activity related to conscious contents.
4	Conscious scenes are unitary.
5	Conscious scenes occur serially – only one conscious scene is experienced at a time.
6	Conscious scenes are metastable and reflect rapidly adaptive discriminations in perception and memory.
7	Conscious scenes comprise a wide multimodal range of contents and involve multimodal sensory binding.
8	Conscious scenes have a focus/fringe structure; focal conscious contents are modulated by attention.
9	Consciousness is subjective and private, and is often attributed to an experiencing 'self'.
10	Conscious experience is reportable by humans, verbally and non-verbally.
11	Consciousness accompanies various forms of learning. Even implicit learning initially requires consciousness of stimuli from which regularities are unconsciously extracted.
12	Conscious scenes have an allocentric character. They show intentionality, yet are shaped by egocentric frameworks.
13	Consciousness is a necessary aspect of decision making and adaptive planning.

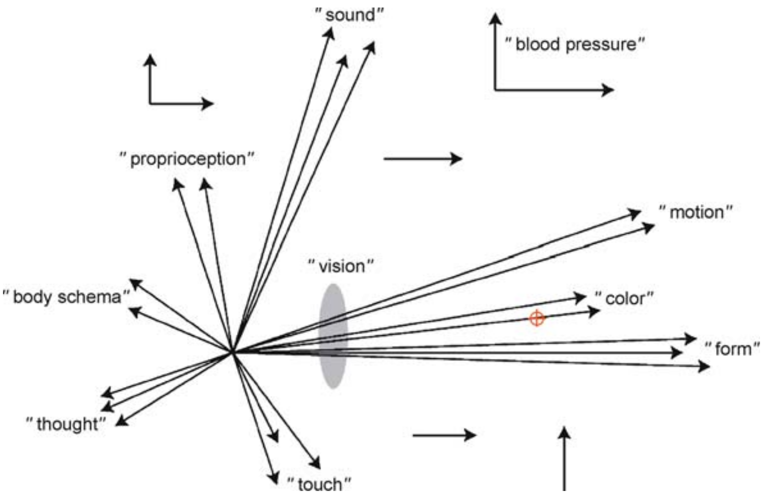
experience: sounds, smells, body signals, thoughts, emotions, and so forth (Fig. 2).

Because the above point is fundamental, it is useful to work through a simple example (adapted from [22,24]). Consider a blank rectangle that is alternately light and dark (Fig. 3a,b). Imagine that this rectangle is all there is, that you are seated in front of it, and that you have been instructed to say "light" and "dark" as appropriate. A simple light-sensitive diode is also in front of the screen and beeps whenever the screen is light. Both you and the diode can perform the task easily, therefore both you and the diode can discriminate these two states: lightness and darkness. But each time the diode beeps, it is entering into one of a total of two possible states. It is minimally differentiated. However, when you say "light" or "dark" you are reporting one out of an enormous number of possible experiences. This point is emphasized by considering a detailed image such as a photograph (Fig. 3c). A conscious person will readily see this image as distinct both from the blank rectangle and from a scrambled version of the same image (Fig. 3d). The diode, however, would classify both images and the rectangle as "light" (depending on its threshold), because it is insufficiently differentiated to capture the differences between the three.

Consider now an idealized digital camera. The electronics inside such a camera will enter a different state for the scrambled image than for the non-scrambled image; indeed, there will be a distinct state for any particular image. A digital camera is capable of much greater differentiation than the diode, but it is still not capable of discrimination because it is minimally *integrated*. In idealized form it is a collection of many independent light-sensitive diodes that must, to a good approximation, remain functionally independent from each other. From the perspective of this camera the image and the scrambled image are equivalent. *We* (as conscious organisms) can tell the difference between the two is because we integrate the many different parts of the image to form a coherent whole. We perceive each part of the image in relation to all the other parts, and we perceive each image in relation to all other possible images and possible conscious experiences that we may have. Successful discrimination therefore requires *both* integration *and* differentiation, and it can be hypothesized that it is this balance that yields the unity and diversity central to conscious experience.

Experimental evidence as well as intuition testifies to the fundamental nature of integration and differentiation in consciousness. A striking example is provided by so-called 'split brain' patients whose cortical hemispheres have been surgically separated. When presented with two independent visuospatial memory tasks, one to each hemisphere, they perform both very well [25]. In contrast, normal subjects cannot avoid integrating the independent signals into a single conscious scene which yields a much harder problem, and performance is correspondingly worse. In general, normal subjects are unable to perform multiple tasks simultaneously if they both require conscious input and they cannot make more than one conscious decision within the so-called 'psychological refractory period', a short interval of a few hundred milliseconds [26].

A loss of differentiation can be associated with the impoverishment of conscious contents following brain trauma. In 'minimally conscious' or 'persistent vegetative' states the dynamical repertoire of the thalamocortical system is reduced to the extent that adaptive behavioral responses are excluded [21]. In less dramatic cases focal cortical lesions can delete specific conscious contents; for example, damage to cortical region V4 can remove color



Consciousness and Complexity, Figure 2
The figure shows an *N*-dimensional neural space corresponding to the dynamic core (see Sect. "Consciousness and Complexity"). *N* is the number of neuronal groups that, at any time, are part of the core, where *N* is normally very large (much larger than is plotted). The appropriate neural reference space for the conscious experience of 'pure red' would correspond to a discriminable point in the space (marked by the *red cross*). Focal cortical damage can delete specific dimensions from this space

dimensions from the space of possible experiences (cerebral achromatopsia [27]; c.f., Fig. 2). Reportable conscious experience is also eliminated during generalized epileptic seizures and slow-wave sleep. Neural activity in these states is again poorly differentiated, showing hypersynchrony (epilepsy) or a characteristic synchronous 'burst pause' pattern (sleep) [22].

Consciousness and Complexity

The Dynamic Core Hypothesis

The notion that consciousness arises from neural dynamics that are simultaneously differentiated and integrated is expressed by the *dynamic core hypothesis* (DCH). This hypothesis has two parts [22,28]:

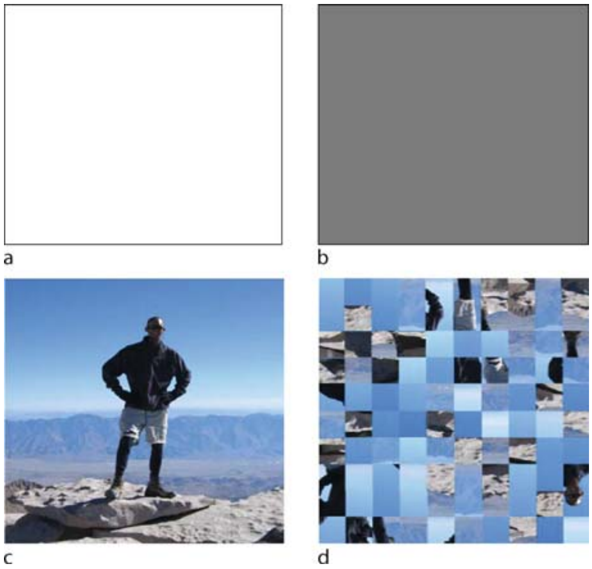
- A group of neurons can contribute directly to conscious experience only if it is part of a distributed functional cluster (the dynamic core) that, through reentrant interactions in the thalamocortical system, achieves high integration in hundreds of milliseconds.
- To sustain conscious experience, it is essential that this functional cluster be highly differentiated, as indicated by high values of complexity.

The concept of a *functional cluster* refers to a subset of a neural system with dynamics that displays high statistical dependence internally and comparatively low statistical dependence with elements outside the subset: A functional cluster 'speaks mainly to itself' [29]. Conceiving of the dynamic core as a functional cluster implies that the boundaries of the neural substrates of consciousness are continually shifting, with neuronal groups exiting and entering the core according to the flow of conscious contents and the corresponding discriminations being made. *Reentry* refers to the recursive exchange of signals among neural areas across massively parallel reciprocal connections and which in the context of the DCH serve to bind the core together. It is important to distinguish reentry from 'feedback' which refers to the recycling of an error signal from an output to an input [30,31]. The interpretation of *complexity* in the context of the DCH is the subject of Sect. "Neural Complexity"; for now we remark that it provides a quantitative measure of neural dynamics that is maximized by simultaneous high differentiation and high integration.

The Theory of Neuronal Group Selection

The DCH emerged from the theoretical framework provided by the 'theory of neural group selection' (TNGS), otherwise known as 'neural Darwinism' [18,32,33]. This section summarizes some of this essential background.

The TNGS is a biological perspective on brain processes with roots in evolutionary theory and immunology. It suggests that brain development and dynamics are *selectionist* in nature, and not instructionist, in contrast to computers which carry out explicit symbolic instructions. Four aspects of selectionist processes are emphasized: di-



Consciousness and Complexity, Figure 3

a A light-colored rectangle. b A dark-colored rectangle. c A detailed image (the summit of Mount Whitney, California). d A scrambled version of the same image. A simple light-sensitive diode would be able to discriminate a from b, but not among a, c, and d, since all these images would appear as 'light'. An idealized digital camera would enter a different state for each image a, b, c, and d, but would not discriminate between c and d because the camera does not integrate the various parts of each image to form a coherent whole. We can discriminate among all images because (i) our brain is capable of sufficient differentiation to enter a distinct state for each image, and (ii) our brain is capable of integrating the various parts of each image to form a coherent whole

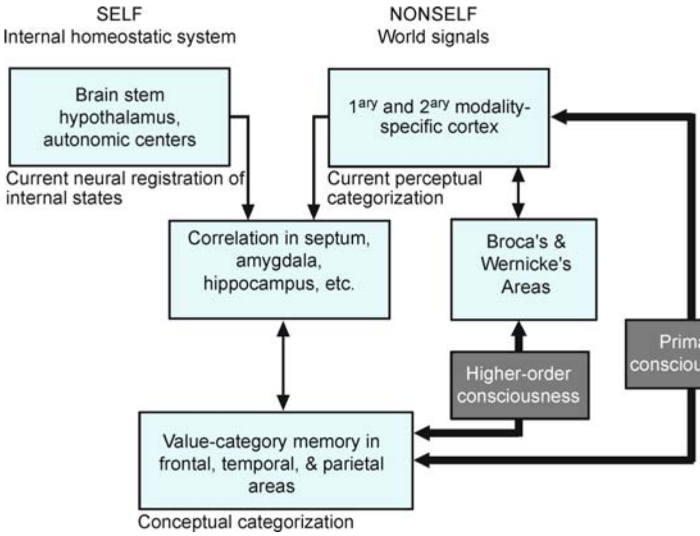
versity, amplification/reproduction, selection, and degeneracy. *Diversity* in the brain is reflected in highly variant populations of neuronal groups where each group consists of hundreds to thousands of neurons of various types. This variation arises as a result of developmental and epigenetic processes such as cell division, migration, and axonal growth; subsequent strengthening and weakening of connections among cells (synapses) via experience and behavior generates further diversity. *Amplification* and *selection* in the brain are constrained by *value*, which reflects the salience of an event and which can be positive or negative as determined by evolution and learning. Value is mediated by diffuse ascending neural pathways originating, for example, in dopaminergic, catecholaminergic, and cholinergic brainstem nuclei [34]. As a result of value-dependent synaptic plasticity, connections among neuronal groups that support adaptive outcomes are strengthened, and those that do not are weakened. Finally, *degeneracy* emphasizes that in adaptive neural systems many structurally different combinations can perform the same function and yield the same output. Degeneracy is a key feature of many biological systems that endows them with adaptive flexibility [35,36]. It is conspicuously absent in artificial systems that are correspondingly fragile (some artificial systems make use of 'redundancy' which differs from degeneracy in that specific functional units are explicitly duplicated; redundancy provides the robustness but not the flexibility of degeneracy).

According to the TNGS, primary consciousness arises when brain areas involved in ongoing perception are linked via reentry to brain areas responsible for a value-based memory of previous perceptual categorizations. On this view, primary consciousness manifests as a 'remembered present' (akin to William James' 'specious present') by which an animal is able to exploit adaptive links between immediate or imagined circumstances and that animal's previous history of value-driven behavior (Fig. 4).

The TNGS and the dynamic core hypothesis are closely related [17,18]. They share the general claim that the neural mechanisms underlying consciousness arose in evolution for their ability to support multimodal discriminations in a high-dimensional space. In addition, the reentrant interactions linking immediate perception to value-category memory are precisely those that are suggested to bind together the dynamic core. Finally, the vast diversity of neural groups is central both to the original TNGS in providing a substrate for selection and to the DCH, in providing an essential component of neural complexity.

Consciousness and the Dynamic Core

We can now summarize the DCH and its origin in the TNGS. Consciousness is entailed by extensive reentrant interactions among neuronal populations in the thalamocortical system, the so-called dynamic core. These interactions, which support high-dimensional discriminations among states of the dynamic core, confer selective advantages on the organisms possessing them by linking current perceptual categorizations to value-dependent memory. The high dimensionality of these discriminations is proposed to be a direct consequence of the rich complexity of the participating neural repertoires. Just as conscious scenes are both differentiated and integrated at the phenomenal level to yield high-dimensional discriminations, so too are the reentrant dynamics of their underlying neural mechanisms differentiated and integrated. Critically according to the TNGS, conscious qualia *are* the high-dimensional discriminations entailed by this balance of differentiation and integration as reflected in high complexity.


SELF
Internal homeostatic system
NONSELF
World signals
Brain stem hypothalamus, autonomic centers
Current neural registration of internal states
1ary and 2ary modality-specific cortex
Current perceptual categorization
Correlation in septum, amygdala, hippocampus, etc.
Broca's & Wernicke's Areas
Primary consciousness
Higher-order consciousness
Value-category memory in frontal, temporal, & parietal areas
Conceptual categorization

Consciousness and Complexity, Figure 4

Primary consciousness and HOC in the TNGS. Signals related to value and signals from the world are correlated and produce value-category memories. These memories are linked by reentry to current perceptual categorization, resulting in primary consciousness. Higher-order consciousness depends on further reentry between value-category memory and current categorization via areas involved in language production and comprehension. Reprinted from [17]

Any theory of consciousness must confront the question of whether conscious experiences have causal effects in the physical world [37]. Responding positively reflects common sense but it seems contrary to science to suggest non-physical causes for physical events. Responding negatively respects the causal closure of the physical world but appears to suggest that conscious experiences are 'epiphenomenal' and could in principle be done without (an implication that may be particularly troubling for experiences of 'free will' [38]). The TNGS addresses this quandary via the notion of *entailment*. According to the TNGS, dynamic core processes entail particular conscious experiences in the same way that the molecular structure of hemoglobin entails its particular spectroscopic properties: it simply could not be otherwise [18]. Therefore, although consciousness does not cause physical events, there exist particular physical causal chains (the neural mechanisms underlying consciousness) that by necessity entail corresponding conscious experiences: The conscious experience cannot be 'done without'.

Measuring Consciousness and Complexity

Having covered basic elements of the DCH and its origin in the TNGS, we turn now to the issue of measuring complexity in neural dynamics. To be useful in this context, candidate measures should satisfy several constraints. We have already mentioned that a suitable measure should reflect the fact that consciousness is a dynamic process [5], not a thing or a capacity. This point is particularly important in light of the observation that conscious scenes arise ultimately from transactions between organisms and environments, and these transactions are fundamentally processes [39]. (This characterization does not, however, exclude 'off-line' conscious scenes, for example those experienced during dreaming, reverie, abstract thought, planning, or imagery). A suitable measure should also take account of causal interactions within a neural system, and between a neural system and its surroundings – i. e., bodies and environments. Finally, to be of practical use, a suitable measure should also be computable for systems composed of large numbers of neuronal elements.

Obviously, the quantitative characterization of complexity can constitute only one aspect of a scientific theory of consciousness. This is true at both the neural level and at the level of phenomenal experience. At the neural level, no single measure could adequately describe the complexity of the underlying brain system (this would be akin, for example, to claiming that the complex state of the economy could be described by the gross domestic product alone). At the phenomenal level, conscious scenes have many diverse features [18,19], several of which do not appear to be readily quantifiable (see Table 1). These include subjectivity, the attribution of conscious experience to a self, and intentionality, which reflects the observation that consciousness is largely about events and objects. A critical issue nevertheless remains: how can measurable aspects of the neural underpinnings of consciousness be characterized [32,132]?

Neural Complexity

A fundamental intuition about complexity is that a complex system is neither fully ordered (e. g., a crystal) nor fully disordered (e. g., an ideal gas). This intuition is compatible with the central theme of the DCH, namely that the neural dynamics within the dynamic core should be both integrated and differentiated. The following definition of *neural complexity* (C_N), first proposed in 1994 [40], satisfies these intuitions and provides a practical means for assessing the complexity of neural and other systems.

Mathematical Definition

Consider a neural system X composed of N elements (these may be neurons, neuronal groups, brain regions, etc.). A useful description of the dynamical connectivity of X is given by the joint probability distribution of the activities of its elements. Assuming that this function is Gaussian, this is equivalent to the covariance matrix of the system's dynamics COV(X). Importantly, COV(X) captures the total effect of all (structural) connections within a system upon deviation from statistical independence of the activities of a pair of elements, and not just the effect of any direct anatomical connection linking them [41]. Given COV(X) and assuming that the dynamics of X are covariance stationary (i. e., having unchanging mean and variance over time) the entropy of the system $H(\mathrm{X})$ is given by:

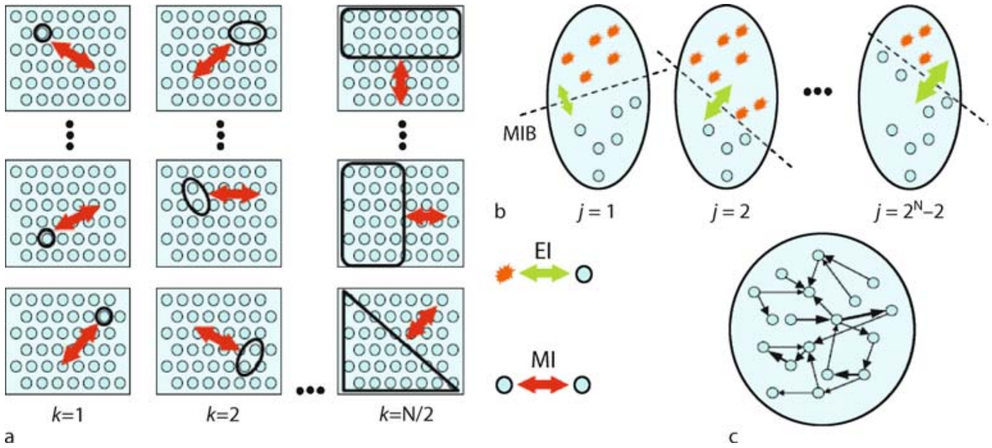
$$H(\mathrm{X}) = \tfrac{1}{2}\ln\left((2\pi \mathrm{e})^N \left|\mathrm{COV(X)}\right|\right)$$

where |.| denotes the matrix determinant [42]. $H(\mathrm{X})$ measures the overall degree of statistical independence exhibited by the system; i. e., its degree of differentiation. Knowing the entropy of a system allows calculation of the mutual information (MI) between two systems, or between two subsets of a single system. The MI between systems (or subsets) A and B measures the uncertainty about A that is accounted for by the state of B and is defined as $\mathrm{MI(A;B)} = H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{AB})$ [43].

The integration of X, $I(\mathrm{X})$, measures the system's overall deviation from statistical independence. All elements in a highly integrated system are tightly coupled in their activity. With x_i denoting the ith element of X, $I(\mathrm{X})$ can be calculated as:

$$I(\mathrm{X}) = \sum_{i=1}^{N} H(x_i) - H(\mathrm{X}) \, .$$

$I(\mathrm{X})$ is equivalent to the measure 'multi-information' which was introduced several decades ago [44]. Having ex-



Consciousness and Complexity, Figure 5
Measuring integration and differentiation in neural dynamics. a Neural complexity C_N is calculated as the ensemble average mutual information (MI) between subsets of a given size and their complement, summed over all subset sizes (k) (adapted from Fig. 2 in [46]). *Small circles* represent neuronal elements and *red arrows* indicate MI between subsets and the remainder of the system. b Information integration Φ is calculated as the effective information (EI) across the 'minimum information bipartition' (MIB). To calculate EI for a given bipartition (j), one subset is injected with maximally entropic activity (*orange stars*) and MI across the partition is measured. c Causal density c_d is calculated as the fraction of interactions that are causally significant according to a multivariate Granger casuality analysis. A weighted (and unbounded) version of causal density (c_{dw}) can be calculated as the summed magnitudes of all significant causal interactions (depicted schematically by *arrow width*). Reprinted with permission from [132]

pressions for MI, $H(\mathrm{X})$, and $I(\mathrm{X})$ allows $C_{\mathrm{N}}(\mathrm{X})$ to be expressed in two equivalent ways. First, $C_{\mathrm{N}}(\mathrm{X})$ can be calculated by summing the average MI between subsets of various sizes, for all possible bipartitions of the system:

$$C_{\mathrm{N}}(\mathrm{X}) = \sum_k \left\langle MI\left(\mathrm{X}_j^k; \mathrm{X} - \mathrm{X}_j^k\right)\right\rangle , \qquad (1)$$

where X_j^k is the jth bipartition of size k, and $\langle . \rangle$ is the average across index j (Fig. 5a). $C_{\mathrm{N}}(\mathrm{X})$ can also be expressed in terms of integration:

$$C_{\mathrm{N}}(\mathrm{X}) = \sum_k \left(\frac{k}{n} I(\mathrm{X}) - \left\langle I\left(\mathrm{X}_j^k\right)\right\rangle\right) , \qquad (2)$$

where $\langle I(\mathrm{X}_j^k)\rangle$ is the average integration of all subsets of size k. $C_{\mathrm{N}}(\mathrm{X})$ will be high if small subsets of the system show high statistical independence, but large subsets show low statistical independence. In other words, $C_{\mathrm{N}}(\mathrm{X})$ will be high if each of its subsets can take on many different states and if these states make a difference to the rest of the system.

Because the full $C_{\mathrm{N}}(\mathrm{X})$ can be computationally expensive to calculate for large systems, it is useful to have an approximation that considers only bipartitions consisting of a single element and the rest of the system. There are three mathematically equivalent ways of expressing this approximation, which is denoted $C(\mathrm{X})$:

$$\begin{aligned} C(\mathrm{X}) &= H(\mathrm{X}) - \sum_{k=1}^{N} H\left(x_i | \mathrm{X} - x_i\right) \\ &= \sum_i MI(x_i; \mathrm{X} - x_i) - I(\mathrm{X}) \\ &= (n-1)\, I(\mathrm{X}) - n\, \langle I(\mathrm{X} - x_i)\rangle , \qquad (3) \end{aligned}$$

where $H(x_i|\mathrm{X} - x_i)$ denotes the conditional entropy of each element x_i given the entropy of the rest of the system $\mathrm{X} - \mathrm{X}_i$. These three expressions are equivalent for all X, whether they are linear or non-linear, and neither $C_{\mathrm{N}}(\mathrm{X})$ nor $C(\mathrm{X})$ can adopt negative values.

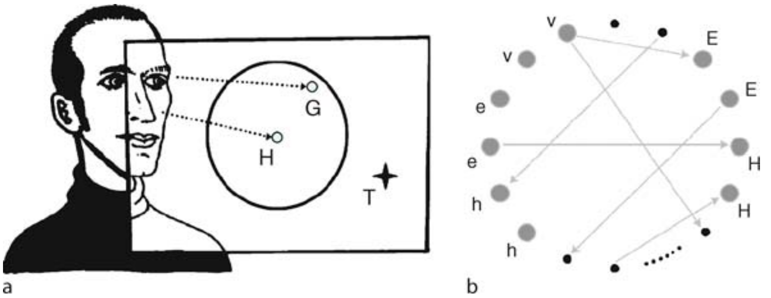
Recently, De Lucia et al. [45] have developed a different approximation to $C_{\mathrm{N}}(\mathrm{X})$ that is calculated directly from topological network properties (i. e., without needing covariance information). Their measure of 'topological $C_{\mathrm{N}}(\mathrm{X})$' is based on the eigenvalue spectrum of the connectivity matrix of a network. While topological $C_{\mathrm{N}}(\mathrm{X})$ offers substantial savings in computational expense it carries the assumption that the network is activated by independent Gaussian noise and therefore cannot be used to measure neural complexity in conditions in which a network is coupled to inputs and outputs (see Subsect. "Complexity and Behavior" below).

Connectivity and Complexity

There is a growing consensus that features of neuroanatomical organization impose important constraints on the functional dynamics underlying cognition [47,48]. Accordingly, several studies have addressed the relationship between structural connectivity and neural complexity [49,50,51,52,53].

One useful approach employs evolutionary search procedures (genetic algorithms [54]) to specify the connection structure of simple networks under various fitness (cost) functions. A population of networks $\mathrm{X}_1 \ldots \mathrm{X}_{\mathrm{N}}$ is initialized ('generation zero') with each member having random connectivity. Each network X_i is then evaluated according to a fitness function [for example, maximize $C(\mathrm{X})$] and those that score highly, as compared to the other networks in the population, are subjected to a small amount of random 'mutation' (i. e., small random changes in connectivity) and proceed to the next 'generation'. This procedure is repeated for many generations until the population contains networks that score near-optimally on the fitness function, or until the experimenter is satisfied that no further improvement is likely.

Sporns and colleagues applied a version of evolutionary search to find distinctive structural motifs associated with $H(\mathrm{X})$, $I(\mathrm{X})$, and $C(\mathrm{X})$ [49]. In this study, the initial population consisted of ten networks each with $N = 32$ nodes and $K = 256$ connections and with fixed identical positive weights w_{ij}. The fitness function was determined by the value of $H(\mathrm{X})$, $I(\mathrm{X})$, or $C(\mathrm{X})$ calculated from the covariance matrix of each network, assuming activation by covariance-stationary Gaussian noise. In each case they found that the resulting networks had distinctive structural features, as revealed both by simple visual inspection and by analysis using a variety of graph-theoretic measures. Networks optimized for $H(\mathrm{X})$ contained mostly reciprocal connections without any apparent local clustering. Networks optimized for $I(\mathrm{X})$ were highly clustered (i. e., neighboring nodes connect mainly to each other [55]) and had a long characteristic path length (i. e., a high mean separation between any two nodes in terms of number of intervening nodes). Finally, networks optimized for $C(\mathrm{X})$ had high clustering (and high reciprocal connectivity) coupled with a short characteristic path length. Strikingly, these networks were very similar to the so-called 'small world' class of network in which dense groups of nodes are connected by a relatively small number of reciprocal 'bridges' [55]. These networks also had a high proportion of 'cycles' (routes through the network that return to their starting point) and very low wiring lengths [49].



Consciousness and Complexity, Figure 6

Target fixation model. a The agent controls head-direction (H) and eye-direction (not shown) in order to move a gaze point (G) towards a target (T). b Neural network controller. The six input neurons are shown on the *left* and the four output neurons on the *right*. Each pair of inputs (v,e,h) responds to x, y displacements: 'v' neurons to displacements of G from T, 'h' neurons to displacements of H from an arbitrary origin ('straight ahead'), and 'e' neurons to displacements of H from the eye-direction. The four output neurons control head direction (H) and eye-direction relative to the head (H). For clarity only four of the 22 interneurons are shown. *Thin gray lines* show synaptic connections. Only a subset of the 256 connections are shown. Adapted from [59]

Sporns et al. extended the above findings by calculating $C(\mathrm{X})$ for networks reflecting the known cortical connectivity of both the macaque visual cortex and the entire cat cortex. In both cases covariance matrices were obtained by assuming linear dynamics, equal connection strengths, and activation by covariance-stationary Gaussian noise. They found that both networks gave rise to high $C(\mathrm{X})$ as compared to random networks with equivalent distributions of nodes and connections. Indeed, the networks seemed to be near-optimal for $C(\mathrm{X})$ because random rewiring of connections led in almost all cases to a reduction in $C(\mathrm{X})$ [49].

In a separate study using a non-linear neuronal network model including excitatory and inhibitory units, Sporns showed that regimes of high $C(\mathrm{X})$ coincided with 'mixed' connection patterns consisting of both local and long-range connections [56]. This result lines up with the previous study [49] in suggesting an association between small-world properties and complex dynamics. In addition, Sporns and Kötter found that networks optimized for the number of functional 'motifs' (small repeating patterns) had high $C(\mathrm{X})$ but those optimized for structural motifs did not [57] suggesting that high complexity reflects the presence of large functional repertoires. Finally, $C(\mathrm{X})$ seems to associate with fractal patterning, but not in a simple sense that fractal networks are optimal for complexity [53,58]. Rather, fractality seems to be one among several structural attributes that contribute to the emergence of small-world features and complex dynamics. Together, these results indicate that only certain classes of network are able to support dynamics that combine functional integration with functional segregation and that these networks resemble in several ways those found in neuroanatomical systems.

Complexity and Behavior

An important claim within the DCH is that complex dynamics provide adaptive advantages during behavior. To test this claim, Seth and Edelman examined the relationship between behavior and neural complexity in a simple agent-based computational model [59]. They evolved networks similar to those in [49] ($N = 32$, $K = 256$) by selecting for their ability to guide target fixation behavior in a simulation model requiring coordination of 'head' and 'eye' movements (Fig. 6). Networks were evolved in both 'simple' and 'complex' environments where environmental complexity was reflected by unpredictable target movement and by variation in parameters affecting head and eye movement. Consistent with the DCH, networks supporting target fixation in rich environments showed higher $C(\mathrm{X})$ than their counterparts adapted to simple environments. This was true both for dynamics exhibited during behavior in the corresponding environments ('interactive' complexity), and for dynamics evoked with Gaussian noise ('intrinsic' complexity).

Sporns and Lungarella explored the relationship between $C(\mathrm{X})$ and behavior in a different way [60]. As in [59], networks acted as neural controllers during performance of a task (in this case, control of a simulated arm to reach for a target). However, instead of evolving for successful behavior, networks were evolved directly for high $C(\mathrm{X})$. Strikingly, selecting for high $C(\mathrm{X})$ led to networks that were able to perform the task, even though performance on the task had not been explicitly selected for. Finally, Lungarella and Sporns asked how $C(\mathrm{X})$ depends on sensorimotor coupling by comparing neural dynamics of a robotic sensory array in two conditions: (i) unperturbed foveation behavior, and (ii) decoupling of sensory input and motor output via 'playing back' previously recorded motor activity [61]. They found significantly higher $C(\mathrm{X})$ when sensorimotor coupling was maintained.

Taken together, the above results suggest a strong link between high neural complexity and flexible, adaptive behavior. Of course, in none of these studies is any claim made that the corresponding networks are in any sense conscious.

Extensions and Limitations

The concept of neural complexity has been extended to characterize the selectional responses of neural systems to inputs in terms of 'matching' complexity C_{M} [62]. C_{M} measures how well the intrinsic correlations within a neural system fit the statistical structure of a sensory stimulus. Simulations show that C_{M} is high when intrinsic connectivity is modified so as to differentially amplify those intrinsic correlations that are enhanced by sensory input, possibly reflecting the capacity of a neurally complex system to 'go beyond the information given' in a stimulus [62]. Despite this possibility C_{M} has not been investigated as thoroughly as has C_{N}.

C_{N} has several limitations. In its full form it is computationally prohibitive to calculate for large networks, but in approximation it is less satisfying as a measure. Also, C_{N} does not reflect complexity in the temporal domain since functional connections are analyzed at zero-lag [23]. Finally, C_{N} does not take into account *directed* (causal) dynamical interactions for the simple reason that MI is a symmetric measure. This last point is addressed by the alternative measures described below.

Information Integration

The most prominent alternative to C_{N} is 'information integration' (Φ) [24,63]. Unlike C_{N}, Φ reflects causal interactions because it is based on 'effective information' (EI), a directed version of MI that relies on the replacement of the outputs of different subsets of the studied system with maximum entropy signals.

Mathematical Definition

Φ is defined as the effective information across the informational 'weakest-link' of a system, the so-called *minimum information bipartition* (MIB; Fig. 5b). It is calculated by the following procedure [63].

Given a system of N elements, identify all possible bipartitions of the system. For each bipartition $A|B$, replace the outputs from A by uncorrelated noise (i. e., maximally entropic activity), and measure how differentiated are the responses of its complement (B). This is the effective information (EI) between A and B:

$$\mathrm{EI}(A \to B) = \mathrm{MI}(A_{H\max}; B)\ ,$$

where $\mathrm{MI}(A_{H\max}; B)$ is the mutual information between A and B when the outputs from A have maximal entropy. $\mathrm{EI}(A{\to}B)$ measures the capacity for causal influence of partition A on its complement B (i. e., all possible effects of A on B). Given that $\mathrm{EI}(A{\to}B)$ and $\mathrm{EI}(B{\to}A)$ are not necessarily equal, one can define:

$$\mathrm{EI}(A \leftrightarrow B) = \mathrm{EI}(A \to B) + \mathrm{EI}(B \to A)\ .$$

The minimum information bipartition (MIB) is the bipartition for which the normalized $\mathrm{EI}(A \leftrightarrow B)$ is lowest. Normalization is accomplished by dividing $\mathrm{EI}(A \leftrightarrow B)$ by $\min\{H_{\max}(A); H_{\max}(B)\}$, so that effective information is bounded by the maximum entropy available. The resulting MIB corresponds to the informational weakest link of the system, and the Φ value of the system is the non-normalized $\mathrm{EI}(A{\leftrightarrow}B)$ across the MIB.

A further stage of analysis has been described [63] in which a system can be decomposed into 'complexes' by calculating Φ for different subsets of elements; a complex is a subset having $\Phi > 0$ that is not included in a larger subset with higher Φ. For a given system, the complex with the maximum value of Φ is called the 'main complex'.

Information Integration, Connectivity, and Complexity

As with neural complexity it is useful to explore what kinds of network structure lead to high values of Φ. Because of computational constraints only comparatively small networks have been investigated for their ability to generate high Φ (i. e., $N = 8$, $K = 16$ as opposed to $N = 32$, $K = 256$ as in [49]). In an initial study, networks optimized for Φ had highly heterogeneous connectivity patterns with no two elements having the same sets of inputs and outputs [63]. At the same time, all nodes tended to emit and receive the same number of connections. These two properties arguably subserve functional segregation and integration, respectively [63].

Although both Φ and C_{N} depend on a combination of functional integration and segregation, they are sensitive to different aspects of network dynamics. C_{N} reflects an average measure of integration that, unlike Φ, does not require heterogeneous connectivity. On the other hand, unlike C_{N}, Φ is determined by the value of an informational measure (EI) across only a single bipartition (the MIB) and is not modified by dynamical transactions across the remainder of the network. Finally, as mentioned above, Φ but not C_N is sensitive to causality.

Limitations and Extensions

As with C_N, Φ does not measure complexity in the temporal domain [23]. There are also substantial limitations attending measurement of Φ for non-trivial systems. First, it is not possible in general to replace the outputs of arbitrary subsets of neural systems with uncorrelated noise. An alternative version of Φ can be envisaged in which 'transfer entropy' (TE) [64], a directed version of MI, is substituted for EI. TE can be calculated from the dynamics generated by a neural system during behavior and therefore does not require arbitrary perturbation of a system; it measures the *actual* causal influence across partitions whereas EI measures the *capacity* for causal influence. However, a version of Φ based on TE does not in general find the informational 'weakest link' (MIB) of a system since the MIB depends on capacity and not on transient dynamics.

Second, unlike C_N there is presently no well-defined approximation for Φ that removes the need to examine all possible bipartitions of a system. However, it may be possible to make use of some informal heuristics. For example, bipartitions for which the normalized value of EI will be at a minimum will be most often those that cut the system in two halves, i. e., midpartitions [63]. Similarly, a representative rather than exhaustive number of perturbations may be sufficient to obtain at least an estimated value of Φ [63].

The Information Integration Theory of Consciousness

Φ occupies a central place in the 'information integration theory of consciousness' (IITC, [24]). According to this theory, consciousness *is* information integration as measured by Φ. The nature of the conscious content in a system with high Φ is determined by the particular informational relationships within the main complex (the complex with the highest Φ). While there are many similarities between the DCH and the IITC, most obviously that both make strong appeal to a measure of complexity, there are also important differences of which we emphasize two:

(i) Because Φ measures the capacity for information integration, it does not depend on neural activity per se. The IITC predicts that a brain where no neurons were active, but in which they were potentially able to react, would be conscious (perhaps of nothing). Similarly, a brain in which each neuron were stimulated to fire as an exact replica of your brain, but in which synaptic interactions had been blocked, would *not* be conscious [24]. The DCH has neither of these implications.

(ii) On the IITC Φ is an adequate measure of the 'quantity' of consciousness, therefore any system (biological or artificial) with sufficiently high Φ would necessarily be conscious. According to the DCH, high C_N is necessary but not sufficient for consciousness.

Point (ii) is particularly important in view of the finding that an arbitrarily high Φ can be obtained by a system as simple as a Hopfield network, which is a fully connected network with simple binary neuronal elements [23]. By choosing the synaptic strengths according to an exponential rule it can be shown that the corresponding Φ value scales linearly with network size, such that $\Phi(\mathrm{X}) = N$ bits for a network X of size N nodes. On the IITC this result leads to the counterintuitive conclusion that a sufficiently large Hopfield network will be conscious. Another challenge for the IITC in this context is the fact that the probability distributions determining entropy values (and therefore by extension Φ values) depend on subjective decisions regarding the spatial and temporal granularity with which the variables in a system are measured ([23,65] but see [24]).

Causal Density

A balance between dynamical integration and differentiation is likely to involve dense networks of causal interactions among neuronal elements. Causal density (c_d) is a measure of causal interactivity that captures both differentiated and integrated aspects of these interactions [23,66]. It differs from C_N by detecting causal interactions, differs from Φ by being sensitive to dynamical interactions across the whole network, and differs from both by being based not on information theory but instead on multivariate autoregressive modeling.

Mathematical Definition

Causal density (c_d) measures the fraction of interactions among neuronal elements in a network that are causally significant (Fig. 5c). It can be calculated by applying 'Granger causality' [67,68], a statistical concept of causality that is based on prediction: If a signal x_1 causes a signal x_2, then past values of x_1 should contain information that helps predict x_2 above and beyond the information contained in past values of x_2 alone [67]. In practice, Granger causality can be tested using multivariate regression modeling [69]. For example, suppose that the temporal dynamics of two time series, $x_1(t)$ and $x_2(t)$ (both of length T), can be described by a bivariate autoregressive model:

$$x_1(t) = \sum_{j=1}^{p} A_{11,j} x_1(t-j) + \sum_{j=1}^{p} A_{12,j} x_2(t-j) + \xi_1(t)$$
$$x_2(t) = \sum_{j=1}^{p} A_{21,j} x_1(t-j) + \sum_{j=1}^{p} A_{22,j} x_2(t-j) + \xi_2(t) \tag{4}$$

where p is the maximum number of lagged observations included in the model (the model order, $p < T$), A contains the coefficients of the model, and ξ_1, ξ_2 are the residuals (prediction errors) for each time series. If the variance of ξ_1 (or ξ_2) is reduced by the inclusion of the x_2 (or x_1) terms in the first (or second) equation, then it is said that x_2 (or x_1) *G-causes* x_1 (or x_2). In other words, x_2 G-causes x_1 if the coefficients in A_{12} are jointly significantly different from zero. This relationship can be tested by performing an F-test on the null hypothesis that $A_{12,j} = 0$, given assumptions of covariance stationarity on x_1 and x_2. The magnitude of a significant interaction can be measured either by the logarithm of the F-statistic [70] or, more simply, by the log ratio of the prediction error variances for the restricted (R) and unrestricted (U) models:

$$gc_{2\to 1} = \begin{cases} \log \dfrac{\operatorname{var}(\xi_{1R(12)})}{\operatorname{var}(\xi_{1U})}\,, & \text{if } gc_{2\to 1} \text{ is significant} \\ 0\,, & \text{otherwise,} \end{cases}$$

where $\xi_{1R(12)}$ is derived from the model omitting the $A_{12,j}$ (for all j) coefficients in Eq. (4) and ξ_{1U} is derived from the full model, where $gc_{2\to 1}$ refers to G-causality from x_2 to x_1.

Importantly, G-causality is easy to generalize to the multivariate case in which the G-causality of x_1 is tested in the context of multiple variables $x_2 \ldots x_N$. In this case, x_2 G-causes x_1 if knowing x_2 reduces the variance in x_1's prediction error when the activities of all other variables $x_3 \ldots x_n$ are also included in the regression model. Both bivariate and multivariate G-causality have been usefully applied to characterizing causal interactions in simulated [71,72] and biological [73] neural systems.

Following a Granger causality analysis, both normalized (c_d) and non-normalized (c_{dw}) versions of causal density of a network X with N nodes can be calculated as:

$$c_d(\mathrm{X}) = \frac{\alpha}{N(N-1)}\,,$$
$$c_{dw}(\mathrm{X}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} gc_{j\to i}$$

where α is the total number of significant causal interactions and $N(N-1)$ is the total number of pairwise interactions among elements. While normalized causal density is bounded to the range [0,1] the non-normalized version is unbounded.

High causal density indicates that elements within a system are both globally coordinated in their activity (in order to be useful for predicting each others activity) and at the same time dynamically distinct (reflecting the fact that different elements contribute in different ways to these predictions). Therefore, as with both C_N and Φ, c_d reflects both functional integration and functional segregation in network dynamics.

Conditions Leading to High Causal Density

In terms of connectivity, computational models show that both fully connected networks (having near-identical dynamics at each node) and a fully disconnected networks (having independent dynamics at each node) have low c_d and c_{dw}; by contrast, randomly connected networks have much higher values [71]. More detailed connectivity studies remain to be conducted.

An initial attempt to analyze behavioral conditions leading to high causal density was made by [66] revisiting the model of target fixation described previously [59]. To recapitulate, in this model networks were evolved in both 'simple' and 'complex' environments where environmental complexity was reflected by unpredictable target movement and by variation in parameters affecting head and eye movement (Fig. 6). Causal density in this model was calculated from first-order differenced time series of the ten sensorimotor neurons and it was found that highest values of causal density occurred for networks evolved and tested in the complex environments. These results mirrored those obtained with C_N, indicating an association between a high value of a complexity measure and adaptive behavior in a richly structured environment.

Extensions and Limitations of Causal Density

A practical problem for calculating causal density is that multivariate regression models become difficult to estimate accurately as the number of variables (i. e., network elements) increases. For a network of N elements, the total number of parameters in the corresponding multivariate model grows as pN^2, and the number of parameters to be estimated for any single time series grows linearly (as pN), where p is the model order (Eq. (4)). We note that these dependencies are much lower than the factorial dependency associated with Φ and C_N, and may therefore be more readily circumvented. One possible approach may involve the use of Bayesian methods for limiting the number of model parameters via the introduction of prior constraints on significant interactions [74]. In neural systems, such prior constraints may be derived, for example, on the basis of known neuroanatomy or by anatomically based clustering procedures.

Several other extensions to causal density are suggested by enhancements to the statistical implementation of Granger causality:

- Non-linear G-causality methods based, for example, on radial-basis-function kernels allow causal density to detect both linear and non-linear causal interactions [75,76].
- Partial G-causality (based on partial coherence) enhances robustness to common input from unobserved variables, supporting more accurate estimates of causal density in systems that cannot be fully observed [77].
- G-causality has a frequency-dependent interpretation [70,73] allowing causal density to be assessed in specific frequency bands.

Causal Density and Consciousness

Although causal density is not attached to any particular theory of consciousness, it aligns closely with the DCH because it is inherently a measure of process rather than capacity. Causal density cannot be inferred from network anatomy alone, but must be calculated on the basis of explicit time series representing the dynamic activities of network elements during behavior. It also depends on all causal interactions within the system, and not just on those interactions across a single bipartition, as is the case for Φ. Finally, causal density incorporates the temporal dimension more naturally than is the case for either C_N or Φ; while the latter measures functional interactions at zero-lag only, causal density incorporates multiple time lags as determined by the order parameter p (Eq. (4)).

The foregoing descriptions make clear that although existing formal measures may have heuristic value in identifying functional integration and functional segregation in neural dynamics, they remain inadequate in varying degrees. C_N can reflect process, can be computed for large systems in approximation, but does not capture causal interactions. Φ captures causal interactions, is infeasible to compute for large neural systems, and can be shown to grow without bound even for certain simple networks. Also, Φ is a measure of capacity rather than process but this is a deliberate feature of the IITC. c_d reflects all causal interactions within a system and is explicitly a measure of process, but it also is difficult to compute for large systems. An additional and important practical limitation of C_N, Φ, and c_d is that they apply only to statistically stationary dynamics.

Empirical Evidence

We turn now to empirical evidence relevant to the DCH. Much of this evidence comes from patients with focal brain lesions and neuroimaging of healthy subjects using functional magnetic resonance imaging (fMRI), electroencephelography (EEG) and magnetoencephalography (MEG). Although current experimental methods are not sufficiently powerful to confirm or refute the DCH, their application, separately and in combination, yields much useful information. A detailed review appears in [17]; below we select some pertinent features.

Involvement of the Thalamocortical System

A wealth of experimental evidence attests to thalamocortical involvement in consciousness, as demonstrated by both clinical studies and by experiments using normal subjects. Clinical studies show that damage to non-specific (intralaminar) thalamic nuclei can abolish consciousness in toto [78], whereas damage to cortical regions often deletes specific conscious features such as color vision, visual motion, conscious experiences of objects and faces, and the like [79]. No other brain structures show these distinctive effects when damaged.

Conscious functions in normal subjects are usefully studied by comparison with closely matched controls who perform the function unconsciously, an approach known as 'contrastive analysis' [80,81,82]. An emerging consensus among contrastive studies is that conscious contents correspond to widespread thalamocortical activation as compared to unconscious controls [81]. For example, Dehaene and colleagues have shown widespread fMRI activation peaks in parietal, prefrontal, and other cortical regions for conscious perception of visual words, as compared to unconscious inputs which activated mainly primary visual cortex [83]. Along similar lines, a recent fMRI study of motor sequence learning showed a shift from widespread cortical involvement during early learning (when conscious attention is required) to predominantly subcortical involvement during later learning phases (when skill production is comparatively 'automatic') [84].

Dynamical Correlates: Binocular Rivalry

A classical example of contrastive analysis makes use of the phenomenon of binocular rivalry, in which different images are projected to each eye [85]. Despite the integrative nature of consciousness these images, if sufficiently different, are not combined into a single composite; rather, conscious experience alternates between them. Srinivasan and colleagues used magnetoencephalography to measure brain responses to flickering visual stimuli under rivalrous conditions [86,87]. A vertical grating flickering at one frequency was presented to one eye and a horizontal grating flickering at another frequency was presented to the other; these different frequencies allowed stimulus-specific brain responses to be isolated in the neuromagnetic signal, a technique known as 'frequency tagging' [88]. As expected for such stimuli, subjects perceived only one grating at any given time. It was found that the power of the frequency-tag of a stimulus was higher by 30–60% across much of the cortical surface when that stimulus was perceptually dominant compared to when it was perceptually suppressed. Moreover, there was a large increase in coherence among distant brain regions, consistent with the idea that conscious perception is associated with widespread integration of neural dynamics mediated by reentry. Although this coherence increase is not a direct test of the DCH it is consistent with the theory and underscores the value of looking for *dynamical* correlates of consciousness.

Cosmelli and colleagues have used a similar paradigm to show that development of a perceptual dominance period arises in neural terms as an extended dynamical process involving the propagation of activity throughout a distributed brain network beginning in occipital regions and extending into more frontal regions [89]. Such a 'wave of consciousness' might reflect underlying integrative processes that lead to the formation of a dynamic core [90]. Chen and colleagues modified the rivalry paradigm so that subjects saw both gratings with both eyes but had to differentially *pay attention* to one or the other [91]. Power increases but not coherence increases were found for the attended stimulus, suggesting that attention may not involve the same global integrative processes implicated in consciousness. Finally, Srinivasan has shown that coherence increases during dominance are due partly to increased phase locking to the external stimulus and partly to increased synchrony among intrinsic network elements, again in line with the idea that consciousness involves coalescence of a distributed functional cluster within the brain [92].

Sleeping, Waking, and Anesthesia

Binocular rivalry involves constant conscious level and changing conscious *content*. Experimental evidence relevant to conscious *level* comes from studies involving transitions between sleeping and waking, anesthesia, epileptic absence seizures and the like. Many studies have tracked changes in endogenous activity across these various transitions but direct assessments of specific complexity measures are mostly lacking. Nonetheless, current findings are broadly consistent with the DCH. As noted in Subsect. "Consciousness as Discrimination", absence seizures and slow-wave sleep (but not rapid-eye-movement sleep) are characterized by hypersynchronous neural activity that may correspond to reduced functional segregation [22]. Anesthesia has particular promise for further experimental study because global anesthetic states can be induced via a wide variety of pharmacological agents having diverse physiological effects. Moreover, proposed unifying frameworks, such as Mashour's 'cognitive unbinding' theory [93], share with the DCH the idea that loss of consciousness can arise from diminished functional integration. In line with Mashour's proposal, John and colleagues have observed at anesthetic loss-of-consciousness (i) functional disconnection along the rostrocaudal (front-to-back) axis and across hemispheres (measured by coherence changes), and (ii) domination of the EEG power spectrum by strongly anteriorized low frequencies [94].

The development of transcranial magnetic stimulation (TMS) has opened the possibility of studying effective (causal) connectivity in the brain. TMS non-invasively disrupts specific cortical regions by localized electromagnetic induction. Massimini and colleagues combined high-density EEG with TMS to test whether effective connectivity among distant brain regions is diminished during sleep [95]. Applying a TMS pulse to premotor cortex during quiet wakefulness led to a sequence of waves propagating throughout the cortex, but this widespread propagation was mostly absent during non-rapid eye movement (slow wave) sleep. While these results do not allow calculation of any specific complexity measure they are consistent with both the IITC and the DCH.

In summary, it is clear that enhanced experimental and analytical methods are needed in order to test adequately whether C_N (or other specific measures) are modulated as predicted by the DCH. Initial attempts to calculate C_N directly from neural dynamics have not been successful (see [96] for a review) although a link to complexity is suggested by the discovery of small-world networks in functional brain dynamics [96,97].

Related Theoretical Proposals

Dynamical Systems Theory and Metastability

We have already mentioned that the measures of complexity discussed in this article apply only to statistically stationary dynamics (Subsect. "Causal Density and Consciousness"). This restriction contrasts sharply with an alternative tradition in theoretical neuroscience that focuses on non-stationary brain dynamics and which emphasizes the tools of dynamical systems theory. This alternative tradition can be traced back to early suggestions of Turing [98] and Ashby [99] and was concisely expressed by Katchalsky in 1974: "… waves, oscillations, macrostates emerging out of cooperative processes, sudden transitions, patterning, etc., seem made to order to assist in the understanding of integrative processes in the nervous system" [100]. More recently the dynamical systems approach has been championed in neuroscience by, among others, Haken [101] under the rubric 'coordination dynamics' and Freeman who has produced a steady stream of papers exploring dynamical principles in brain activity [102,103]. Valuable reviews of work in this tradition can be found in [104,105,106].

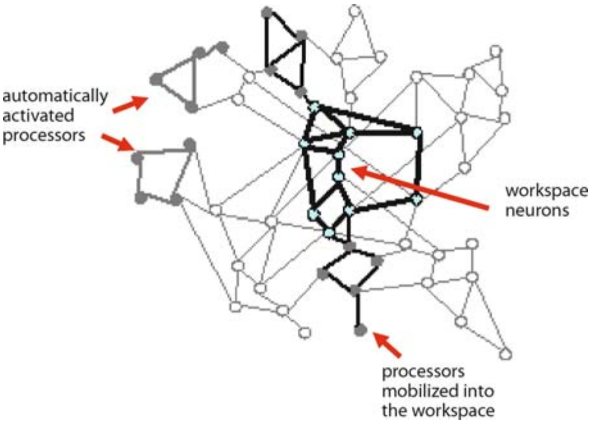
A key concept in the dynamical systems approach is 'metastability' which describes dynamics that are "distinguished by a balanced interplay of integrating and segregating influences" (see p. 26 in [107]). While this definition is obviously similar to the intuition driving neural complexity, metastability has been fleshed out, not in the concepts of information theory or time-series analysis, but instead in the language of attractor dynamics. A dynamical system inhabits a metastable regime when there are no stable fixed points but only partial attraction to certain phase relationships among the system variables. At the level of neural dynamics metastability may reflect the ongoing creation and dissolution of neuronal assemblies across distributed brain regions [105,107,108]. A now classical experimental example of metastability comes from a study in which subjects were asked to flex a finger in response to a periodic tone, initially in a syncopated manner [107]. As the frequency of the tone increases, the syncopated response becomes harder to maintain until a critical point is reached at which the subject switches to a synchronous mode of response. Strikingly, this behavioral phase transition is accompanied by a corresponding transition in the patterning of neuromagnetic cortical signals. At the critical point, where there is partial attraction to both syncopated and synchronous response modes, both behavioral and neural dynamics are dominated by metastability. For other evidence of metastability in the brain see [109].

Metastability characterizes an important aspect of conscious experience, namely that conscious events are rapidly adaptive and fleeting [17]. Consciousness is remarkable for its present-centeredness [5,6]. Immediate experience of the sensory world may last at most a few seconds and our fleeting cognitive present is surely less than half a minute in duration. This present-centeredness has adaptive value for an organism by allowing time enough to recruit a broad network of task-related neural resources while permitting neural dynamics to evolve responses to subsequent events. Thus, conscious experience can be described by an interplay of segregating and integrating influences in both the temporal (metastability) and spatial (complexity) domains. A key theoretical challenge is to work out in greater detail the relationship between these two concepts.

Global Workspace Theory

Beginning in 1988 [80] Baars developed a cognitive theory of consciousness under the rubric 'global workspace (GW) theory' [80,81,110]. The cornerstone of GW theory is the idea that consciousness involves a central resource (the GW) which enables distribution of signals among numerous otherwise informationally encapsulated and functionally independent specialized processors. GW theory states that mental content becomes conscious mental content when it gains access to the GW such that it can influence a large part of the brain and a correspondingly wide range of behaviors. A key aspect of GW theory is that conscious contents unfold in an integrated, serial manner but are the product of massively parallel processing among the specialized processors. The integrated states of the GW follow each other in a meaningful but complex progression that depends on multiple separate processes, each of which might have something of value to add to the ongoing constitution of the GW. Although these notions are compatible with the DCH, they do not by themselves specify dynamical properties to the same level of detail.

A dynamical approach to GW theory has been pursued by Wallace [112] and separately by Dehaene, Changeux and colleagues [111,113,114]. Wallace adopts a graph-theoretic perspective proposing that the GW emerges as a 'giant component' among transient collections of otherwise unconscious processors. The formation of a giant component in graph theory denotes a phase transition at which multiple sub-networks coalesce to form a large network including the majority of network nodes [115]. Dehaene and colleagues have built a series of computational models inspired by GW theory, to account for a variety of psychological phenomena including 'attentional blink' [111], 'inattentional blindness' [114], and effortful cognitive tasks [113]. These models are based on the concept of a 'neuronal global workspace' in which sensory stimuli mobilize excitatory neurons with long-range cortico-cortical axons, leading to the genesis of global activ-



Consciousness and Complexity, Figure 7
A schematic of the neuronal global workspace. A central global workspace, constituted by long-range cortico-cortical connections, assimilates other processes according to their salience. Other automatically activated processors do not enter the global workspace. Adapted from [111]

ity patterns among so-called 'workspace neurons' (Fig. 7). This model, and that of Wallace both, predict that consciousness is 'all or nothing' – i. e., a gradual increase in stimulus visibility should be accompanied by a sudden transition (ignition) of the neuronal GW into a corresponding activity pattern. As with Wallace, although some dynamic properties of the neuronal GW have been worked out and are compatible with the DCH, a rigorous account of how the model relates to neural complexity has not been attempted.

Neuronal Synchrony and Neuronal Coalitions

The association of neural synchrony with consciousness arose from its proposed role as a mechanism for solving the so-called 'binding problem', which in general terms refers to the problem of coordinating functionally segregated brain regions. The binding problem is most salient in visual perception for which the functional and anatomical segregation of visual cortex contrasts sharply with the unity of a visual scene. Since the 1980s a succession of authors have proposed that the binding problem is solved via neuronal synchronization [116,117] and both experimental evidence [118,119] and computational models have borne out the plausibility of this mechanism [31]. In the 1990s, starting with an early paper by Crick and Koch [7], this proposal grew into the hypothesis that consciousness itself is generated by transient synchronization among widely distributed neuronal assemblies, with particular emphasis on oscillations in the gamma band ($\sim$ 40 Hz) [39,120]. In support of this idea we have already seen that conscious perception correlates with increased synchrony of a (non-gamma-band) visual 'frequency tag' [87], and several studies have reported associations between gamma-band synchrony and consciousness [121,122,123]. However, synchrony-based theories of binding (and by extension consciousness) remain controversial [124] and there is not yet evidence that disruptions of gamma-band synchrony lead to disruptions of conscious contents [12].

From the perspective of the DCH a deeper concern with the synchrony hypothesis is that it accounts only for integration, and not for the combination of integration and differentiation that yields the discriminatory power of consciousness. In a recent position paper [125], Crick and Koch reversed their previous support for gamma-band synchrony as a sufficient mechanism for consciousness, favoring instead the notion of competition among 'coalitions' of neurons in which winning coalitions determine the contents of consciousness at a given time. Such neuronal coalitions bear similarities to the decades-old notion of Hebbian assemblies [126] on a very large and dynamic scale. They also suggest that unconscious processing may consist largely of feed-forward processing whereas consciousness may involve standing waves created by bidirectional signal propagation, a proposal advanced as well by Lamme [127]. Crick and Koch note that the 'coalition' concept is similar to the dynamic core concept [125] although lacking in the detailed formal specification of the latter.

A possible role for gamma-band synchrony in both the DCH and in Crick and Koch's framework is that it may facilitate the formation but not the ongoing activity of the core (or a coalition) [125]. In this light it is suggestive that correlations between gamma-band synchrony and consciousness tend to occur at early stages of conscious perception [121,122].

Outlook

Scientific accounts of consciousness continue to confront the so-called 'hard problem' of how subjective, phenomenal experiences can arise from 'mere' physical interactions in brains, bodies, and environments [128,129]. It is possible that new concepts will be required to overcome this apparent conceptual gap [130]. It is equally likely that increasing knowledge of the mechanisms underlying consciousness will lead these philosophical conundrums to fade away, unless they have empirical consequences [81,125]. In short, to expect a scientific resolution to the hard problem as it is presently conceived may be to misunderstand the role of science in explaining nature. A scientific theory cannot presume to replicate the experience that it describes or explains; a theory of a hurricane is not a hurricane [18]. If the phenomenal aspect of experience is irreducible, so is the fact that physics has not explained why there is something rather than nothing, and this ontological limit has not prevented physics from laying bare many mysteries of the universe.

The approach described in this article is one of developing *explanatory correlates* of consciousness, namely properties of neural dynamics that are experimentally testable and that account for key properties of conscious experience. Thanks to accelerating progress in experimental techniques and increased attention to theory, the outlook for this approach is healthy. We close by suggesting some key areas for further study:

Development of a large-scale model of a dynamic core Although progress in large scale neural network modeling has been rapid [131], we currently lack a sufficiently detailed model of environmentally coupled thalamocortical interactions needed to test the mechanistic plausibility of the DCH. Having such a model should also allow substantive connections to be drawn between the DCH and GW theory.

Development of new experimental methods New methods are needed to track neuronal responses at sufficient spatio-temporal resolutions to support accurate estimation of C_N and other complexity measures during different conscious and unconscious conditions. Among current methods fMRI has poor time resolution and measures neural activity indirectly, while MEG/EEG lacks spatial acuity and is unable to record details of thalamic responses.

Complexity and metastability New theory is needed to relate the class of complexity measures described in this article to metastability, which analyzes functional segregation and integration in the temporal domain.

Emergence and 'downward causality' New theory is also needed to better understand how global dynamical states arise from their basal interactions and how these global states can constrain, enslave, or otherwise affect properties at the basal level [39]. Applied to consciousness and to cognitive states in general, such downward causality can suggest functional roles and may even help reconcile the phenomenology of free-will with physiological fact.

Bibliography

Primary Literature

1. Haldane E, Ross G (1985) The philosophical work of Descartes. Cambridge University Press, Cambridge
2. Popper K, Eccles JF (1977) The self and its brain. Springer, New York
3. Penrose R (1994) Shadows of the mind: A search for the missing science of consciousness. Oxford University Press, Oxford
4. McGinn C (1991) The problem of consciousness. Blackwell, Oxford
5. James W (1904) Does consciousness exist? Philos J Psychol Sci Methods 1:477–491
6. Edelman GM (1989) The remembered present. Basic Books, New York
7. Crick F, Koch C (1990) Towards a neurobiological theory of consciousness. Semin Neurosci 2:263–275
8. Dalton TC, Baars BJ (2003) Consciousness regained: The scientific restoration of mind and brain. In: Dalton TC, Evans RB (eds) The lifecycle of psychological ideas: Understanding the prominence and the dynamics of intellectual change. Springer, Berlin, pp 203–247
9. Koch C (2004) The quest for consciousness: A neurobiological approach. Roberts, Greenwood Village
10. Metzinger T (2000) Neural correlates of consciousness: Empirical and conceptual questions. Press MIT, Cambridge
11. Llinás R, Ribary U, Contreras D, Pedroarena C (1998) The neuronal basis for consciousness. Philos Trans Soc R Lond Biol B Sci 353:1841–1849
12. Rees G, Kreiman G, Koch C (2002) Neural correlates of consciousness in humans. Nat Rev Neurosci 3(4):261–70
13. Crutchfield JP, Young K (1989) Inferring statistical complexity. Phys Rev Lett 63:105–108
14. Zurek WH (1990) Complexity, entropy, and the physics of information. Addison-Wesley, Redwood City
15. Adami C (2002) What is complexity? Bioessays 24:1085–1094
16. Tye M (2007) Qualia. In: Zalta E (ed) The Stanford Encyclopedia of Philosophy (Summer 2008 Edition). http://plato.stanford.edu/archives/sum2008/entries/qualia/
17. Seth AK, Baars BJ (2005) Neural Darwinism and consciousness. Conscious Cogn 14:140–168
18. Edelman GM (2003) Naturalizing consciousness: A theoretical framework. Proc Natl Acad Sci USA 100(9):5520–5524
19. Seth AK, Baars BJ, Edelman DB (2005) Criteria for consciousness in humans and other mammals. Conscious Cogn 14(1):119–139
20. Edelman DB, Baars BJ, Seth AK (2005) Identifying the hallmarks of consciousness in non-mammalian species. Conscious Cogn 14(1):169–187
21. Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. Trends Cogn Sci 9:556–559
22. Tononi G, Edelman GM (1998) Consciousness and complexity. Science 282:1846–1851
23. Seth AK, Izhikevich E, Reeke GN, Edelman GM (2006) Theories and measures of consciousness: An extended framework. Proc Natl Acad Sci USA 103(28):10799–10804
24. Tononi G (2004) An information integration theory of consciousness. Neuroscience BMC 5(1):42
25. Gazzaniga MS (2005) Forty-five years of split-brain research and still going strong. Nat Rev Neurosci 6:653–659
26. Pashler H (1994) Dual-task interference in simple tasks: data and theory. Psychol Bull 116:220–244
27. Zeki S (1990) A century of cerebral achromatopsia. Brain 113(6):1721–1777
28. Edelman GM, Tononi G (2000) A universe of consciousness: How matter becomes imagination. Basic Books, New York

29. Tononi G, McIntosh AR, Russell DP, Edelman GM (1998) Functional clustering: identifying strongly interactive brain regions in neuroimaging data. Neuroimage 7:133–149
30. Tononi G, Sporns O, Edelman GM (1992) Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. Cerebral Cortex 2(4):31–35
31. Seth AK, McKinstry JL, Edelman GM, Krichmar JL (2004) Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. Cerebral Cortex 14:1185–99
32. Edelman GM (1987) Neural Darwinism. Basic Books, New York
33. Edelman GM (1993) Selection and reentrant signaling in higher brain function. Neuron 10:115–125
34. Friston KJ, Tononi G, Reeke GN, Sporns O, Edelman GM (1994) Value-dependent selection in the brain: Simulation in a synthetic neural model. Neuroscience 59:229–243
35. Tononi G, Sporns O, Edelman GM (1999) Measures of degeneracy and redundancy in biological networks. Proc Natl Acad Sci USA 96:3257–3262
36. Edelman GM, Gally J (2001) Degeneracy and complexity in biological systems. Proc Natl Acad Sci USA 98(24):13763–13768
37. Kim J (1998) Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation. Press MIT/Bradford Books, Cambridge
38. Wegner D (2003) The illusion of conscious will. Press MIT, Cambridge
39. Thompson E, Varela F (2001) Radical embodiment: Neural dynamics and consciousness. Trends Cogn Sci 5:418–425
40. Tononi G, Sporns O, Edelman GM (1994) A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proc Natl Acad Sci USA 91:5033–5037
41. Vanduffel W, Payne BR, Lomber SG, Orban GA (1997) Functional impact of cerebral connections. Proc Natl Acad Sci USA 94:7617–7620
42. Papoulis A, Pillai SU (2002) Probability, random variables, and stochastic processes, 4th edn. McGraw-Hill, New York
43. Jones DS (1979) Elementary information theory. Clarendon Press, Oxford
44. McGill WJ (1954) Multivariate information transmission. Trans IEEE Inform Theory 4:93–111
45. de Lucia M, Bottaccio M, Montuori M, Pietronero L (2004) A topological approach to neural complexity. Phys Rev E 71:016114
46. Tononi G, Edelman GM, Sporns O (1998) Complexity and coherency: Integrating information in the brain. Trends Cogn Sci 2:474–484
47. Bressler SL (1995) Large-scale cortical networks and cognition. Brain Res Brain Res Rev 20:288–304
48. Friston K (2002) Beyond phrenology: what can neuroimaging tell us about distributed circuitry? Annu Rev Neurosci 25: 221–250
49. Sporns O, Tononi G, Edelman GM (2000) Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. Cerebral Cortex 10:127–141
50. Sporns O, Tononi G (2002) Classes of network connectivity and dynamics. Complexity 7(1):28–38
51. Seth AK, Edelman GM (2004) Theoretical neuroanatomy: Analyzing the structure, dynamics and function of neuronal networks. In: Ben Naim E, Fraunfelder H, Toroczkai Z (eds) Complex networks. Lecture Notes in Physics. Springer, Berlin, pp 487–518
52. Sporns O, Chialvo D, Kaiser M, Hilgetag C (2004) Organization, development and function of complex brain networks. Trends Cogn Sci 8:418–425
53. Buckley CL, Bullock S (2007) Spatial embedding and complexity: The small-world is not enough. In: Almeida e Costa F (ed) Proceedings of the Ninth European Conference on Artificial Life. Springer, Berlin, pp 986–995
54. Mitchell M (1997) An introduction to genetic algorithms. Press MIT, Cambridge
55. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small world' networks. Nature 393:440–442
56. Sporns O (2004) Complex neural dynamics. In: Jirsa VK, Kelso JAS (eds) Coordination dynamics: Issues and trends. Springer, Berlin, pp 197–215
57. Sporns O, Kötter R (2004) Motifs in brain networks. PLoBiol S 2:e369–e369
58. Sporns O (2006) Small-world connectivity, motif composition, and complexity of fractal neuronal connections. Biosystems 85:55–64
59. Seth AK, Edelman GM (2004) Environment and behavior influence the complexity of evolved neural networks. Adapt Behav 12:5–21
60. Sporns O, Lungarella M (2006) Evolving coordinated behavior by maximizing information structure. In: Rocha L, Yaeger L, Bedau MS, Floreano D, Goldstone RL, Vespignani A (eds) Proceedings of the 10th European Conference on Artificial Life. Press MIT, Cambridge, pp 323–330
61. Lungarella M, Sporns O (2006) Mapping information flow in sensorimotor networks. PLoComput S Biol 2:e144–e144
62. Tononi G, Sporns O, Edelman GM (1996) A complexity measure for selective matching of signals by the brain. Proc Natl Acad Sci USA 93:3422–3427
63. Tononi G, Sporns O (2003) Measuring information integration. Neuroscience BMC 4(1):31
64. Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85(2):461–4
65. Werner G (2007) Perspectives on the neuroscience of cognition and consciousness. Biosystems 87:82–95
66. Seth AK (2005) Causal connectivity of evolved neural networks during behavior. Netw Comput Neural Syst 16:35–54
67. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438
68. Seth AK (2007) Granger causality. Scholarpedia, p 15501
69. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
70. Geweke J (1982) Measurement of linear dependence and feedback between multiple time series. J Amer Stat Assoc 77:304–13
71. Seth AK (2008) Causal networks in simulated neural systems. Cogn Neurodyn 2:49–64
72. Seth AK, Edelman GM (2007) Distinguishing causal interactions in neural populations. Neural Comput 19:910–933
73. Ding M, Chen Y, Bressler S (2006) Granger causality: Basic theory and application to neuroscience. In: Schelter S, Winterhalder M, Timmer J (eds) Handbook of Time Series Analysis. Wiley, Wienheim, pp 438–460

74. Zellner A (1971) An introduction to Bayesian inference in econometrics. Wiley, New York
75. Ancona N, Marinazzo D, Stramaglia S (2004) Radial basis function approaches to nonlinear granger causality of time series. Phys Rev E 70:056221
76. Chen Y, Rangarajan G, Feng J, Ding M (2004) Analyzing multiple nonlinear time series with extended Granger causality. Phys Lett A 324:26–35
77. Gao S, Seth AK, Kendrick K, Feng J. Partial granger causality: Eliminating exogenous input. J Neurosci Methods 172(1)79–93
78. Bogen JE (1995) On the neurophysiology of consciousness: I An overview. Conscious Cogn 4:52–62
79. Kolb B, Whishaw IQ (1996) Fundamentals of human neuropsychology, 4th edn. Freeman WH, New York
80. Baars BJ (1988) A cognitive theory of consciousness. Cambridge University Press, New York
81. Baars BJ (2002) The conscious access hypothesis: Origins and recent evidence. Trends Cogn Sci 6:47–52
82. Lau HC, Passingham RE (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. Proc Natl Acad Sci USA 103:18763–18768
83. Dehaene S, Naccache L, Cohen L, Bihan DL, Mangin JF, Poline JB, Rivière D (2001) Cerebral mechanisms of word masking and unconscious repetition priming. Nat Neurosci 4:752–758
84. Floyer-Lea A, Matthews PM (2004) Changing brain networks for visuomotor control with increased movement automaticity. Neurophysiol J 92:2405–2412
85. Blake R, Logothetis N (2002) Visual competition. Nat Rev Neurosci 3:13–21
86. Tononi G, Srinivasan R, Russell DP, Edelman GM (1998) Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. Proc Natl Acad Sci USA 95:3198–3203
87. Srinivasan R, Russell DP, Edelman GM, Tononi G (1999) Increased synchronization of magnetic responses during conscious perception. J Neurosci 19:5435–5448
88. Silberstein RB, Schier MA, Pipingas A, Ciorciari J, Wood SR, Simpson DG (1990) Steady-state visually evoked potential topography associated with a visual vigilance task. Brain Topogr 3:337–347
89. Cosmelli D, David O, Lachaux J-P, Martinerie J, Garnero L, Renault B, Varela F (2004) Waves of consciousness: Ongoing cortical patterns during binocular rivalry. Neuroimage 23:128–140
90. Nunez P, Srinivasan R (2006) A theoretical basis for standing and traveling brain waves measured with human EEG with implications for an integrated consciousness. Clin Neurophysiol 117:2424–2435
91. Chen Y, Seth AK, Gally JS, Edelman GM (2003) The power of human brain magnetoencephalographic signals can be modulated up or down by changes in an attentive visual task. Proc Natl Acad Sci USA 100:3501–3506
92. Srinivasan R (2004) Internal and external neural synchronization during conscious perception. Int Bifurcat J Chaos 14:825–842
93. Mashour GA (2004) Consciousness unbound: Toward a paradigm of general anesthesia. Anesthesiology 100:428–433
94. John ER, Prichep LS, Kox W, Valdés-Sosa P, Bosch-Bayard J, Aubert E, Tom M, di Michele F, Gugino LD (2001) Invariant reversible QEEG effects of anesthetics. Conscious Cogn 10:165–183
95. Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, Tononi G (2005) Breakdown of cortical effective connectivity during sleep. Science 309:2228–2232
96. Stam CJ, Jones BF, Nolte G, Breakspear M, Scheltens P (2007) Small-world networks and functional connectivity in Alzheimer's disease. Cereb Cortex 17:92–99
97. Stephan KE, Hilgetag CC, Burns GA, O'Neill MA, Young MP, Kötter R (2000) Computational analysis of functional connectivity between areas of primate cerebral cortex. Philos Trans Soc R Lond Biol B Sci 355:111–126
98. Turing A (1950) Computing machinery and intelligence. Mind 59:433–460
99. Ashby WR (1952) Design for a brain: The origin of adaptive behaviour. Chapman Hall, London
100. Katchalsky A, Rowland V, Hubermann B (1974) Neurosci Res Prog Bull 12
101. Haken H (2002) Brain dynamics. Springer, New York
102. Freeman WJ (1975) Mass action in the nervous system. Academic Press, New York
103. Freeman WJ (2005) A field-theoretic approach to understanding scale-free neocortical dynamics. Biol Cybern 92(6):350–359
104. Kelso JAS (1995) Dynamic patterns: The self-organisation of brain and behavior. Press MIT, Cambridge
105. Werner G (2007) Metastability, criticality and phase transitions in brain and its models. Biosystems 90:496–508
106. Izhikevich EM (2006) Dynamical systems in neuroscience: The geometry and excitability of bursting. Press MIT, Cambridge
107. Bressler S, Kelso J (2001) Cortical coordination dynamics and cognition. Trends Cogn Sci 5:26–36
108. Friston KJ (1997) Transients, metastability, and neuronal dynamics. Neuroimage 5:164–171
109. Fingelkurts A (2004) Making complexity simpler: Multivariability and metastability in the brain. Int Neurosci J 114:843–862
110. Baars BJ (2005) Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. Prog Brain Res 150:45–53
111. Dehaene S, Sergent C, Changeux J-P (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc Natl Acad Sci USA 100:8520–8525
112. Wallace R (2005) Consciousness: A mathematical treatment of the neuronal global workspace model. Springer, New York
113. Dehaene S, Kerszberg M, Changeux J-P (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci USA 95:14529–14534
114. Dehaene S, Changeux J-P (2005) Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. PLoBiol S 3:e141–e141
115. Bollobás B (1985) Random graphs. Academic Press, London
116. von der Malsburg C (1995) Binding in models of perception and brain function. Curr Opin Neurobiol 5:520–526
117. Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 18:555–586

118. Fries P, Reynolds JH, Rorie AE, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291:1560–1563
119. Steinmetz PN, Roy A, Fitzgerald PJ, Hsiao SS, Johnson KO, Niebur E (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404:187–190
120. Engel AK, Fries P, König P, Brecht M, Singer W (1999) Temporal binding, binocular rivalry, and consciousness. Conscious Cogn 8:128–151
121. Melloni L, Molina C, Pena M, Torres D, Singer W, Rodriguez E (2007) Synchronization of neural activity across cortical areas correlates with conscious perception. Neurosci J 27: 2858–2865
122. Palva S, Linkenkaer-Hansen K, Näätänen R, Palva J (2005) Early neural correlates of conscious somatosensory perception. Neurosci J 25:5248–5258
123. Meador KJ, Ray PG, Echauz JR, Loring DW, Vachtsevanos GJ (2002) Gamma coherence and conscious perception. Neurology 59:847–854
124. Shadlen MN, Movshon JA (1999) Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron 24:67–77 111
125. Crick F, Koch C (2003) A framework for consciousness. Nat Neurosci 6:119–126
126. Hebb DO (1949) The organization of behavior. Wiley, New York
127. Lamme VAF (2006) Towards a true neural stance on consciousness. Trends Cogn Sci 10:494–501
128. Chalmers DJ (1996) The conscious mind: In search of a fundamental theory. Oxford University Press, Oxford
129. Bennett MR, Hacker PMS (2003) Philosophical foundations of neuroscience. Blackwell, Oxford
130. Thompson E (2004) Life and mind: A tribute to Francisco Varela. Phenomenol Cogn Sci 3:381–98
131. Izhikevich E, Gally JS, Edelman GM (2004) Spike-timing dynamics of neuronal groups. Cerebral Cortex 14:933–944
132. Seth AK, Dienes Z, Cleeremans A, Overgaard M, Pessoa L (2008) Measuring consciousness: relating behavioural and neurophysiological approaches. Trends Cogn Sci 12:314–21

Books and Reviews

Baars DJ, Banks WP, Newman JB (2003) Essential sources in the scientific study of consciousness. Press MIT, Cambridge

Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: from biological networks to the Internet and WWW. Oxford University Press, Oxford

Edelman GM (2004) Wider than the sky: The phenomenal gift of consciousness. Yale University Press, New Haven

Edelman GM, Tononi G (2000) A universe of consciousness: How matter becomes imagination. Basic Books, New York

Koch C (2004) The Quest for Consciousness: A Neurobiological Approach. Roberts, Greenwood Village

Metzinger T (2000) Neural correlates of consciousness. Press MIT, Cambridge

Tononi G (2004) An information integration theory of consciousness. Neuroscience BMC 5(1):42

Seth AK, Izhikevich E, Reeke GN, Edelman GM (2006) Theories and measures of consciousness: An extended framework. Proc Natl Acad Sci USA 103(28):10799–10804

Continuum Percolation

ISAAC BALBERG
The Racah Institute of Physics, The Hebrew University,
Jerusalem, Israel

Article Outline

Glossary

Site percolation Assuming a lattice in which the sites can be occupied, with a probability p^{s}, two sites are assumed locally connected if they are nearest neighbors and occupied. A group of sites in which each of its' sites is occupied and locally connected to at least one other site in the group is called a connected cluster. If there is a cluster that connects the edges of an "infinite" lattice, the cluster is called the percolation cluster. The lowest p^{s} for which such a cluster is found is called the percolation threshold, $p_{\mathrm{c}}^{\mathrm{s}}$.

Bond percolation Assuming a lattice, the segment between nearest neighbor sites is called a bond. A bond can be occupied with a probability p^{b}. Two bonds are locally connected if they are both occupied and have a common site. A group of bonds in which each is connected to at least one other bond is called a connected cluster of bonds. If the span of this cluster is infinite we have a "percolation cluster" and the lowest p^{b} for which such a cluster is formed is called the bonds percolation threshold, $p_{\mathrm{c}}^{\mathrm{b}}$.

Critical behavior Percolation can be modeled as a phase transition and thus various properties have a power-low dependence on $|p - p_{\mathrm{c}}|$ for small values of $|p - p_{\mathrm{c}}|/p_{\mathrm{c}}$. The exponent that describes this dependence is called the critical exponent.

Universal and non universal behavior If the critical exponent depends only on the dimensionality of the system, we say that the critical behavior is universal. If the exponent depends on other parameters of the system, we say that the behavior is non universal.

Continuum percolation In a system of objects one can define a local connectivity criterion such as the overlap of pores in porous media. The objects are assumed to be randomly distributed in space, they may have various and variable shapes and sizes and they may interact with each other. A group of objects such that each object is locally connected to at least one other object in the group is a connected cluster. The cluster that has an infinite span is the "percolating" cluster in the continuum and the lowest concentration of objects that yield such a cluster defines the percolation threshold.

Critical fractional volume The content of the objects at the onset of global connectivity, i. e. the percolation threshold, is usually characterized in the continuum by the measurable fractional volume content of the objects of interest in the system. This volume at the percolation threshold is the critical fractional volume.

Excluded volume In the case where the local connectivity is determined by a partial overlap of the volumes of two equal objects, the volume in space in which the two centers of the two objects can be, and must be, in order to have such an overlap is defined as the excluded volume of the object. If the objects are not equal one has to define a corresponding average.

Average bonds per object The average number of objects that are locally connected to a given object in the system. This quantity, at the percolation threshold, B_c, is the quantity that characterizes topologically the onset of global connectivity i. e. the onset of percolation.

"Pointedness" The parameter that can be defined qualitatively as the degree of deviation from sphericity of a given object. The manifestation of this property for different objects, that have the same excluded volume, is that with its increase the value of B_c decreases.

Critical behavior of dynamical properties The critical behavior of a property that has to do with flow in the system, such as electrical conductivity and fluid permeability. In the continuum, it is related to the distribution of the local values of the dynamical parameter associated with a given bond. If the average of this parameter effects the global critical behavior of the system, its contribution is added to the "universal" critical behavior that is found in lattices.

The random void and the inverted random void systems A system of pores or topologically similar systems which are reminiscent of the "Swiss Cheese". In the random void (RV) case one is concerned with the network excluding the pores. The mirror image of this system, i. e. when one is concerned with the network that consists of the pores (or particles that can coalesce) is called the inverted random void (IRV) system. In these systems the "neck" formed by the separation (RV) or the overlap (IRV) of two adjacent pores determines the local dynamical property.

Tunneling percolation The conduction process in a system where the "local connectivity" is not determined by a geometrical contact but is determined electronically. This is in particular by inter-object tunneling. The corresponding connected system exhibits a percolation-like critical behavior of the dynamical properties.

"Physically controlled percolation" When an externally applied quantity, such as an electric field or mechanical pressure, affects the parameters that characterize the percolation-like behavior, one may call the corresponding phenomena "physically controlled percolation".

Definition of the Subject

In this review we will describe the two main features that are very different in lattice percolation and in continuum percolation and which bring about new concepts that are not encountered in the "more traditional" and well developed theory of percolation in lattices [121]. The origin of the difference is that in continuum percolation we are dealing with "real" objects that have sizes and shapes, and that are randomly distributed in space, while in lattices we are dealing with abstract mathematical objects such as dots and line segments in a priori defined locations. The result is that in the continuum we have to use different quantities to describe the percolation threshold and we obtain a much richer variability of the corresponding behaviors of the dynamical properties, such as the electrical transport. To equip the reader with the basic concepts needed for the discussion of continuum percolation we also include a short introduction to the relevant principles of lattice percolation and then we discuss in some detail the very basic concepts and achievements in continuum percolation. Finally, we discuss open questions and possible future developments associated with this field.

Percolation theory is concerned with the effects of local and global connectivity on the geometrical and dynamical properties of systems. Within the limited scope of this review we cannot discuss the numerous systems and many properties associated with continuum percolation and we deal only with the principles of this field. However, in order to illustrate the great variety of systems and properties that constitute this area of research, we mention now single, typical, examples of the most conspicuous of those.

Continuum percolation encompasses all the many body systems from the smallest, the elementary particles systems [104], to the largest, the galactic-cosmological systems [109]. Correspondingly continuum percolation was applied to discuss and describe structures and phenomena in physics, chemistry, earth sciences and biology, as well as communication and traffic networks. As some typical examples, let us mention, the electrical properties of numerous composites [128], porous media [125], microemulsions [36,69], disordered semiconductors [137], disordered superconductors [91], molecular [75] and macromolecular [39] liquids, nano-tubes in composites [45,64] and suspensions [64], quantum dot composites [27], thin metal films [118], layered materials [47], quasi-crystals [98], chemical networks [92] and transport in them [96], biological networks [134] and flow in them [117], bioinformatics [99], and ecological systems [132]. Then there are the properties associated with the flow of charge carriers, such as electrical conductivity [10], the flow of liquids such as the permeability [88], as well as corresponding processes [77] such as displacements [133], drainage [6], dispersion [107], hydrological flows [31] and diffusion [130]. Very closely related is the class of rheological properties [57] such as the viscosity [84]. The possible variation in the local ("bond strength") parameters makes the concepts of continuum percolation also applicable to information and traffic management [135]. Finally, there are the related areas of the elastic [81], the dielectric [67] and the magnetic [122] properties, and their relation to electrical properties [28,85,93].

In this review we will limit ourselves to a discussion of the percolation threshold and electrical conductivity, as they are the most studied properties in "real systems", with the hope that the principles that are understood to govern this property will be sufficient for approaching the understanding of the various systems and properties such as those mentioned above. In particular this review is expected to be also helpful for the analysis of related behaviors such as the electrical noise [127], piezoresistivity [71,128], magneto-transport [29], alternating current conductivity [43], and the effects of interplay between electrical and elasto-thermal properties [8]. An example of a generalization of the concepts used in the discussion of the electrical conductivity to other properties is the extension of the theory [127] and the experimental analysis [102] of the electrical conductivity to the electrical noise.

Turning to the historical highlights of the development of lattice percolation theory and continuum percolation theory we realize that by the very nature of this limited and subjective review there will be many important works that will not be mentioned here and need to be summarized in a more expanded review. We point out however that our previous reviews [11,17,26,32] cover the basic understanding of the present subject for the corresponding years.

The history of percolation theory is usually considered to begin with the works of Flory in 1941 [60] and Stockmayer in 1943 [123] on gelation, and, by definition, as a flow problem in the work of Broadbent and Hammersley in 1957 [33], who also coined the name of the corresponding processes as percolation phenomena [136]. These works were followed by very many works that consolidated the concepts associated with the theory of percolation on lattices. Excellent general reviews on lattice percolation were given by Zallen in 1983 [136], Stauffer and Aharony in 1992 [121], and Sahimi in 1994 [105] and some aspects of it were considered in more specialized reviews such as by Kirkpatrick in 1973 [80], Shklovskii and Efros in 1984 [114], Isichenko in 1992 [77] and Sahimi in 1998 [106].

The principal concepts, that were developed for lattices were the site and bond occupation probabilities, p, and the resulting clusters, that are groups of occupied sites or bonds that are connected by having pairs of occupied sites (or bonds) when the latter are nearest neighbors. The corresponding theory was concerned with the properties of the clusters that enable this connectivity, and in particular those of the infinite-spanning cluster, when it exists. The onset of this cluster at a given occupation probability, p_c, is associated with the onset of global connectivity as well as with the possible onsets of "flows" through the systems. The major developments of percolation theory were then along two lines that will be emphasized in this review: the determination of the onset of global connectivity, and the variation of geometrical and dynamical properties of such systems as the percolation threshold is approached. The corresponding developments in the 1960's culminated in the 1969 work of Kasteleyn and Fortuin (that is detailed in Fortuin and Kasteleyn [62]). They were able to show that the p transition through p_c can be described as a second order phase transition [119], thus justifying a posteriori the previous findings that the above various properties behave as power laws of $|p - p_c|$, with powers that are only dimensionally dependent. The latter behavior as $p \to p_c$ is known as the critical behavior, and the independence of the exponents on other system parameters is referred to as the universal critical behavior of the quantities in percolation systems. While the initial motivation to consider percolation problems was to account for the behavior of "real" systems, the considerations of such systems and the realization that they cannot be simply discussed using the terms of lattice systems was postponed to the 1970's. Since "real" systems are made of objects and these are randomly dispersed in the continuum, the corresponding research field of "real" systems became synonymous with the later coined concept of "continuum" percolation.

The first major step in trying to account generally for more realistic systems can be considered the finding of Scher and Zallen (hereafter S&Z) in 1970 [108]. They found that for properly placed hard core spheres in lattice points, the onset of global connectivity will take place at a volume fraction of the objects in the system, τ_c, that is nearly a constant for all possible lattices in a given dimension D. This is in contrast with the relatively wide range of p_c values for the various lattices in the same D, thus suggesting that at τ_c there is an onset of a fundamental topological connectivity. Similar invariants were found by Powell in 1979 [97] who essentially used a model in which "conducting" and "non-conducting" spheres fill a box under "gravitation". Another approach, which relates to problems of flow via porous-like media or impurity band conduction-like systems, was taken by Holcomb and Rehr in 1969 [76], and Pike and Seager in 1974 [95]. They considered essentially the fractional occupied volume of spherical pores, when the local connectivity is defined by partial overlap of the pores, at the onset of global connectivity. They also found invariants in the case for systems for which no volume can be attached to the object such as widthless sticks in a two dimensional system [25]. A very significant insight into the problem of percolation thresholds of systems in the continuum was provided by Shante and Kirkpatrick (hereafter S&K) in 1971 [112]. They confirmed that the underlying topology of the connected objects, i. e. the average number of connected permeable spheres, per given sphere, at the onset of percolation, is the same as that of the "fine grid" lattice. It took then more than a decade until Balberg and co-workers in 1983 [14,20] and 1984 [21,22] provided the generalization of the above concepts to zero-volume and "soft-core" non-parallel objects. They systematically established the concepts of the average excluded volume, $\langle V_{\text{ex}} \rangle$, and the number of bonds per object, B_c. Consequently the quantity $B_c = \rho_c \langle V_{\text{ex}} \rangle$, where ρ_c is the number of objects per unit volume at the onset of percolation, became the guide for *the phenomenological theory* of the percolation threshold in the continuum [131] and its application [30]. For example, these concepts of $\langle V_{\text{ex}} \rangle$ and B_c have enabled the explanation of the nearly zero volume of the conducting phase needed for the onset of percolation in many systems [9] and the analysis of the various cases where the objects are not entirely permeable [17].

The first step towards a rigorous derivation of results in continuum percolation is probably the suggestion of Coniglio, DeAngelis and Forlani who developed, in 1977 [44], a series expansion in the objects density, ρ, to describe the average size of the finite clusters in fluid systems. This utilization of liquid theory [74] was translated to the prediction of the ρ_c values for partially permeable spheres by DeSimone, Stratt and Demoulini in 1986 [49], and to an analytic prediction of trends, in the B_c dependence on the type of the objects, by Bug, Safran and Webman in 1985 [34]. Extension of the latter approach and the application of extrapolation methods (by fitting the results to the known average cluster size dependence on $|\rho - \rho_c|$) finally yielded very accurate predictions for B_c [3,53]. In particular the work of Drory et al. in 1997 [56] has lent a firm basis for a rigorous theory for percolation thresholds in the continuum, which followed Drory's works in 1996 [50,51] and 1997 [52] that proved that continuum percolation is also a second order-like phase transition. An insight concerning the variation of B_c (that is much less straightforward than the variation of ρ_c) from a system of one type of objects to a system of another type of objects was suggested by Alon, Balberg and Drory who introduced the concept of "*pointedness*" in 1991 [4]. Recently, some works on permeable systems in high dimensions have shown that applying intuition for the prediction of various specific behaviors is not always justified [63,68]. The case at hand was the finding of Wagner, Balberg and Klein, in 2006 [131], that for $D \to \infty$ the value of B_c can become smaller than the intuitively expected unity [136].

Turning to the critical behavior of percolation systems in the continuum, we note that the study of the dynamical properties, and in particular the study of the electrical resistivity, became an intensive area of research in the 1970's. The first work to test experimentally the predicted properties on "real" systems can be considered to be the study of Abeles, Pinch and Gittleman in 1975 [2] regarding the critical behavior of the electrical resistivity of granular metals [1] at the onset of metallic conduction. They found that the conductivity exponent that characterizes the critical behavior, t, was 1.9 in accordance with the universal values predicted for lattices. This was followed by numerous works that yielded the same conclusion on other electrical properties, such as the electrical noise [40]. In fact various experiments on "real" systems have also confirmed the lattice-like geometrical-structural exponents, such as the correlation length exponent, ν [18,79] and the fractal dimension of the percolation cluster, D_f [113]. In parallel, the developments in computers and consequent computer simulations in the 1980's have enabled quite accurate simulations for the determination of the critical exponents in continuum systems, yielding again very good agreement with the universal predictions [20]. On the other hand, the many experimental results that did not conform with the universal predictions [7,94] were largely ignored by the mainstream of the research. Even the "toy" model of non-universal behavior of the electrical resistivity, proposed by Kogut and Straley (hereafter K&S) in 1979 [82], was bypassed for a few years, until it became the cornerstone of the understanding of non universal behavior of the dynamical properties. This overlooking of the theory of K&S was apparently for a good reason, since all that time there was no known link to connect the features of the continuum system with the parameters of their "toy" model. This "missing" link was provided by the work of Halperin, Feng and Sen in 1985 [72]. They showed that non universality can be a fundamental property of dynamical properties in continuum systems. By finding the distribution function of the pore sizes, in sedimentary rock-like systems, they were able to show its immediate consequence regarding the resistors value distribution, and thus the appearance of non universality in accordance with the model of K&S. Their finding opened the way then to the understanding of the many dynamical properties in the various continuum systems that were mentioned above. In particular, soon after these findings, in 1987, Balberg showed [10] that tunneling conduction in random systems can also be mapped onto a K&S-like resistor distribution. This realization has enabled, at least, the qualitative interpretation of the numerous experimental results in the broadest group of studied continuum systems, i. e. composite materials, and the broadest group of studied properties, i. e. the various electrical transport properties. The very simple theoretical predication that was derived for the latter case yielded, however, critical exponents that were much larger than the many experimentally observed values that were reported in the literature [26]. The other difficulty with the corresponding physical picture was that it did not explain how is it that systems in which all the particles are electrically connected on the one hand, and in which there is no geometrical connectivity, between any pair of them, on the other hand, yield behaviors that are identical to those obtained in bona fide percolation systems [13]. The two problems have been resolved in principle only very recently, the first by Grimaldi and Balberg in 2006 [70], and the other by Toker et al. in 2003 [124]. These developments will be described in more detail below.

Introduction

Percolation phenomena are associated with the consequences of the connectivity in systems composed of elementary members that are distributed in some random manner, after a criterion for the connectivity between two such elements is defined. The corresponding theory that deals with those phenomena tries to describe geometrical structures of connected parts in those systems as well as their dynamical properties. The most interesting and consequently the most studied manifestations of percolation phenomena are those taking place in the close proximity to the onset of global connectivity in the system. This onset is known as the percolation threshold, and the behavior of the geometrical and dynamical properties in the vicinity of this threshold is known as the critical behavior of the percolation system. It was shown that this critical behavior can be described as a second-order phase transition and thus, close enough to the percolation threshold, the various properties are given by power laws [121] that are reminiscent of the dependencies found in the most well-known phase transitions [119] such as the simplest magnetic transitions [116].

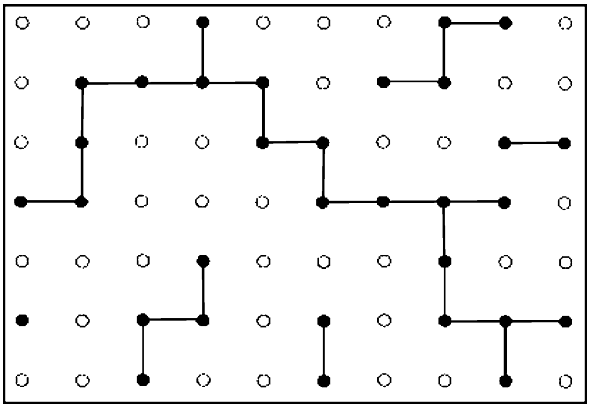
There are thousands of papers that have been written on the theory and the applications of percolation in the continuum. These applications are found in very many fields of physics, chemistry, biology, geophysics and communications, and there are very many relevant properties that are discussed in terms of the concepts of this theory. Since all these cannot be summarized within the framework of the present review we will concentrate here on the theory and the understanding of the principal concepts and on the major experimental manifestations of the theory and its consequences. In particular we will focus on the central essential concept of the *percolation threshold* in the continuum and on the dynamical property that was studied more than any other property in corresponding systems, i. e., the *electrical conductivity*.

Since the fundamental concepts that are used for the description of percolation in the continuum are based on the great advancements made in the theory of lattice percolation, we will start this review by briefly outlining the principles of the latter theory. Following the fact that there are excellent textbooks [106,121,136] and many review articles on lattice percolation, this background will be given very briefly in Sect. "Lattice Percolation". It is expected that this brief background will be sufficient for the understanding of the basic ideas of continuum percolation that will be discussed then in more detail here, in Sects. "Principal Issues in Continuum Percolation", "Percolation Thresholds in the Continuum", and "The Critical Behavior of the Dynamical Properties in the Continuum". In those sections we will put emphasis on the properties of continuum percolation that cannot be extrapolated simply from their lattice counterparts. For example, the almost zero-percolation threshold as well as the non universal behavior of various physical properties do not follow the scenarios of, and expectations from, lattice percolation. Finally, in Sect. "Future Directions" we consider the major issues in continuum percolation the solution of which seem to be important in order to deepen our understanding of this and related fields of research. In Sect. "Bibliography" we provide a rather very limited list of papers relevant to the subjects considered here with the hope that this will be sufficient as a guide for the application of continuum percolation theory in various and many related research areas.

Lattice Percolation

To introduce the above mentioned basic concepts we start by considering the very simple model of the two dimensional square lattice that is illustrated in Fig. 1. Let us assume a probability, p, that a lattice site is occupied, and denote the occupied sites by full circles. The step that is very important here, and will be very crucial when we discuss continuum percolation, is the definition of the local connectivity between members in a given system. In the present lattice these members are the occupied sites and two such sites are defined to be connected if they are first nearest neighbors. In the particular case shown in Fig. 1 the number of nearest neighbors of a given site (the coordination number of the lattice), z, is 4. We can now define a (connected) cluster size as the number of sites such that each of them is connected to at least one (same or other) occupied site in the cluster. In Fig. 1 we see then clusters of size 1, size 2, size 4 and a larger cluster that connects opposite edges of the system. The latter is called the spanning cluster. In percolation theory we are concerned mainly with "infinite" systems, i. e. with systems where the details (e. g. the sites themselves and the intersite distances) are much smaller than the size of the system. Correspondingly, in an "infinite" system we call the spanning cluster, the percolation, or the infinite, cluster. The illustration given in Fig. 1 reflects then a finite (small) portion of the infinite system. Such small portions are used below for the description and definitions of local and global quantities of the system, but it has to be remembered that quantitative characterizations of the various properties are considered meaningful only in the (statistically sufficient) "infinite" system limit.

From Fig. 1 it is easy to appreciate that the number of clusters and their sizes will grow as the site occupation, p, increases, as can be derived quantitatively by various models and simulations [121,136]. The most important consequence, however, is that there will be a value of p, that is known in the literature as p_c, such that for $p < p_c$ there will be no percolation cluster, while for $p \geq p_c$ there will be such a cluster. This p_c is known as the *percolation threshold* of the lattice. The behavior of the various geometrical or physical properties for p values in the close proximity of p_c is known as the *critical behavior*. The above mentioned considerations of the percolation as a phase transition yield then that these properties will be well described by power laws of the proximity to the threshold $|p - p_c|/p_c$. For example, the average cluster size, S, is defined usually [121] as the average number of sites in a cluster, per an occupied site in the lattice. The average that is thus found for the finite clusters in the lattice, behaves, for both, $p > p_c$ and $p < p_c$, as



Continuum Percolation, Figure 1
A two dimensional illustration of a portion of an infinite lattice. The open circles represent lattice sites and the full circles represent occupied lattice sites. The segments connecting two nearest neighbor occupied sites represent a local connection or a bond. In this illustration there are finite clusters of connected sites (of size 1, 2 and 4) and a "spanning" cluster that connects two opposite edges of the system. In a system of infinite size the latter is known as the "infinite" or the "percolation" cluster. (From [32])

$$S \propto |p - p_c|^{-\gamma} \,. \tag{1}$$

In particular, the phase transition-like behavior yields that exponents, such as γ, depend only on the lattice dimension D, and are thus independent of the details (e. g. the lattice type) of the system. This property is known as the *universality* and the corresponding dependence on $|p - p_c|$ as the *universal behavior*. The universal values of γ are, for example, 2.4 for $D = 2$ and 1.8 for $D = 3$. The other important property of the clusters is their geometrical extent, i. e. their effective (e. g. gyration) radius (say, in units of the lattice spacing). The corresponding average for the finite clusters in the lattice (e. g., per sites connected to a given site in the cluster; Stauffer and Aharony [121]) defines then the "connectivity" length in the system, ξ. This length, as to be expected from the above discussion, behaves as

$$\xi \propto |p - p_c|^{-\upsilon} \,, \tag{2}$$

where ν is a "universal" exponent. The quantity ξ, known also as the *correlation length*, is the most important parameter of the system since it characterizes its connectivity and provides the basic length scale for dealing with the behavior of the properties of interest in the system. The well known values of υ are 1.33 for $D = 2$ and 0.88 for $D = 3$.

We note in passing that in Fig. 1 we may also consider the occupancy of the bonds (i. e. the segments between the sites) for which the local bond connectivity criterion is the occupation of two bonds that share a site [129]. It is obvious that for a given lattice there is a simple numerical factor that relates the site occupancy, p^s, and the bond occupancy, p^b [136]. Below, we will use then the more general p concept to describe both quantities except when the difference between the two types of occupancy needs to be emphasized.

Turning to the dynamic properties associated with the connectivity of the system in lattices we consider now the global conductance of a system, G. This is the dynamic property that we will discuss in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum". for continuum systems. Viewing then the system shown in Fig. 1 as consisting of "bonds" (between two occupied sites), let us assume that each bond is a resistor bar of a given resistance value. In such a system there will be electrical conduction from right to left or from top to bottom, only if there is a spanning cluster. Examining the system illustrated in Fig. 1 from this point of view, we see that only a fraction of the 18 resistor-bonds in this cluster (15 from left to right or 11 from top to bottom) will participate in the conduction in the respective direction. On the other hand, there will be a few resistance-bonds (3 from left to right, or 7 from top to bottom) that do not participate in the conduction process. The bonds of the first type belong to the *backbone* of the percolation cluster while the others are known as the *"dead ends"* of the percolation cluster.

Let us consider now the structure of the backbone net. We can imagine the backbone as (say, a cubic) net of "links" and "nodes". These "links" intersect then at the "nodes" that define the corresponding net. The average length of the links is easily appreciated to be of the order of ξ, since the "holes" in the backbone net can encapsulate only finite clusters, and these, as we saw above, have their diameter distributed around ξ. Turning to the structure of the links, we can look at it from a geometrical point of view or an electrical point of view. In the first picture let us consider those bonds in the link that connect two sites in the link such that there is no other "indirect" connection between them within the distance ξ (i. e. when only the members of the link) are included. We call the corresponding bond a "*singly connected bond*". From the electrical point of view the latter bonds are simply those that carry the same current as the link (say, through a plane that intersects the link, but only through members of the link). The parts of the backbone that are confined between two adjacent "singly connected bonds" are known as the "blobs" of the backbone. Hence, since the system will conduct only for $p > p_c$, the backbone can be envisioned as consisting of, say, a square or a cubic, network of "links" that are made of series connections of "singly connected bonds" and "blobs". The model described here is known as the *links-nodes-blobs* (the LNB) model [32,114,120,121].

Considering the expected resistance of a cubic (or hypercubic) LNB network let us assume that the resistance of a link is R_ξ. If the sample size is L we have then, on the average, L/ξ links that connect one edge of the sample with its counterpart. Correspondingly, there will be $(L/\xi)^{D-1}$ parallel links between these "opposite-edges", where D is the dimensionality of the system. Since the link's resistance is R_ξ, the resistance of the whole network $\mathbf{R}_L$ will be given by

$$\mathbf{R}_L = R_\xi (\xi/L)^{D-2} \,. \tag{3}$$

The crucial step here is then "only" to estimate the value of R_ξ [32,114]. The corresponding easiest approach (that can be termed also the lattice approach) is to assume that all the bond-resistors have the same resistance value $\mathbf{r}_0$ [121,122]. In passing we note that when the resistors' values come from a distribution, even if a lattice structure is assumed as the platform of the resistors network, the problem is classified as a problem in continuum percolation. This follows of course the fact that in "real" (or continuum) systems a distribution in the resistance values of the local resistors is usually expected (see Sect. "The Local Resistors and Their Distributions"). The deviation from the simple $\mathbf{r}_0$ assumption that is used in lattice models will be shown in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum" to lead to the very different and rich behaviors that are encountered in continuum percolation.

However, with the above assumption of all resistors having the same value $\mathbf{r}_0$, we can estimate the value of R_ξ in the LNB model as follows. We saw that the link consists of singly connected bonds and blobs. Let us determine first how many singly connected bonds we expect to have in a link of length ξ. Assuming a lattice with $p > p_c$ and assuming that there are L_1 such bonds in this link [114], let us cut out $(p - p_c)/p$ of the bonds in the whole lattice. The number of singly connected bonds that will be cut in a link will be then $L_1(p - p_c)/p$. If, upon the increase of L_1 as $p \to p_c$, this quantity reaches unity, there will be on the average one singly connected bond missing in a link and "all" the links will not conduct. In the limit of $\xi \to \infty$ (i. e. $(p - p_c)/p_c \ll 1$) we approach then the onset of "no percolation" with $L_1 \propto (p - p_c)^{-1}$. If we neglect the blobs (that, following the above, are aggregates of parallel resistors and thus have lower resistance than the same length chain of singly connected resistors), the resistance of the link can be estimated by its lower bound of $R_\xi = \mathbf{r}_0 L_1 \propto (p - p_c)^{-1}$. Since the blobs do contribute to the resistance of the link, one expects that $R_\xi > \mathbf{r}_0 L_1$ and that R_ξ will behave as $(p - p_c)^{-\zeta}$ where ζ is an exponent that is dimensional-dependent. While the calculation of the value of ζ is not straightforward [122], it can be well approximated by fractal models [48]. That the value of ζ is larger, but not much larger, than unity is to be expected, since, as in the "dilution" of the network of links (when $p \to p_c$), there is also a dilution in the local structure of the blobs, and the links will approach the limit of a chain of singly connected bonds. In other words as $p \to p_c$ the resistance $\mathbf{R}_L$ will be affected by both the dilution of the links in the network Eq. (3) and the dilution of the resistors in the blobs. The latter effect means a "stronger" (in comparison with the one that would be encountered in a link that is made only of "singly connected bonds") divergence of the sample resistance as $p \to p_c$ and thus it is manifested by ζ that are, for $D < 6$, larger than 1. Indeed, it was found that $\zeta \approx 1.3$ for $D = 2$ and $\zeta \approx 1.1$ for $D = 3$ [121]. All the above results can be summarized now by presenting the critical behavior of the global resistance, $\mathbf{R}_L$, in a percolation system of size L, by

$$\mathbf{R}_L (\equiv 1/G) \propto (p - p_c)^{-t} , \qquad (4)$$

where t is the critical exponent of the global conductance G, that is given, in view of the above, by

$$t = (D - 2)\nu + \zeta . \qquad (5)$$

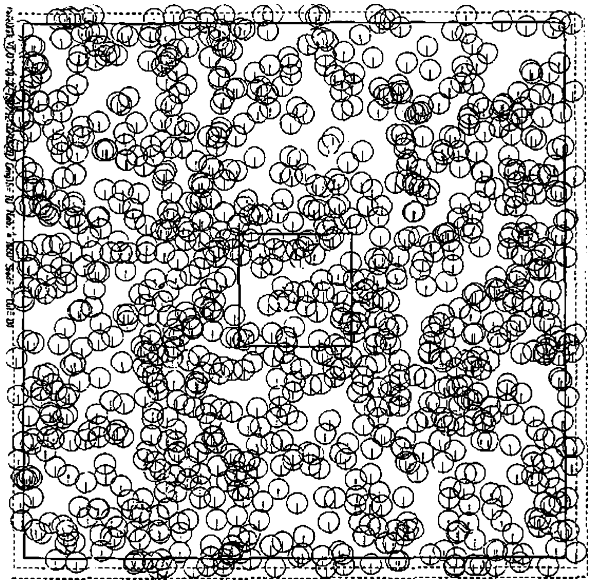
The numerical t values that were derived by various corresponding analytical approximations, or Monte Carlo simulations, are known as the universal values of the conductivity exponent, t_{un}. These values are, in particular [121,136], 1.3 for $D = 2$ and ≈ 2 for $D = 3$. The deviation from these "lattice" values that represent the behavior of the "global network" in lattices will be discussed in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum".

Principal Issues in Continuum Percolation

In the previous section we have briefly reviewed the principal concepts of lattice percolation where the systems have sites (or bonds) on lattices and these can be empty or occupied with a probability $p^s(p^b)$. This yields a percolation threshold p_c that depends on the particular lattice of interest. The physical properties on the other hand, all of which are associated with the existence of the corresponding bonds in the system, are all characterized by the same single-valued physical parameters. In continuum percolation the system is composed of objects (or structures) that are *randomly placed in space*, that may be of *various sizes and various shapes* and, if non-spherical, may have a *distribution of their orientations*. Correspondingly, the physical parameters that characterize the bonds may vary from one bond to another in a manner that may or may not be determined solely by the local geometry and environment of the bond. The two most significant issues that are of concern in continuum percolation are then the connectivity of the system, as reflected by the percolation threshold, in systems of objects, and the possibility of a different critical behavior of the geometrical and physical properties in comparison with that of lattices.

To make this general discussion more specific, we start by an illustration of a well known prototype of a two dimensional continuum percolation system which consists of randomly distributed circles, as shown in Fig. 2. The circles in this model can be described as permeable, overlapping, interpenetrating, or of "soft core". Such a system can represent a metal sheet in which holes have been drilled at random. In this case the electrical conduction between opposite edges of the sample is carried by the "background" (the sample without the holes) of the sheet. We call this "interpretation" of the system, the random void (RV), or the "Swiss Cheese" model. The system shown in Fig. 2 can be also considered as representing a mirror image of the RV system, i. e. to be taken to represent pores (filled, say, with a conducting liquid) in a porous medium, or a collection of isolated and fused metallic grains embedded in an insulating matrix as in granular metals. In these cases, the conducting parts are the pores or the grains, and their network provide the global conductivity of the system. The transport in the latter system can be of course of a liquid or an electrical charge. Correspondingly, this view of the system is known as the inverted random void (IRV) model.

In the latter model the criterion for local connectivity (that is equivalent to the simple bond connection in lattice percolation) is rather simple; two circles are connected if they partially overlap. While (as we will see later) the local



Continuum Percolation, Figure 2
A two dimensional illustration of a continuum system. The circles are distributed randomly and two of them are considered connected if they partially overlap. There are finite clusters here as well as an "infinite" cluster of connected circles. A small finite portion of the system is enclosed in the square shown at the center. (From [130])

connectivity criterion is not always so simple in continuum systems, this one is easy to understand intuitively and thus it is the most abundant model in continuum percolation. In what follows, and in particular, for the discussion of the "connecting criteria" and their interpretation in the continuum, we will usually illustrate our arguments by figures that show small portions of the infinite system, such as the area enclosed in the square that confines the center of Fig. 2.

It is already the first examination of the system shown in Fig. 2 that makes one appreciate the fundamental questions of continuum percolation: how to define quantitatively the percolation threshold in such a system (there is no apparent meaning to p), and how the local degree of objects overlap determines the local and global geometrical and dynamical properties of the system. We will discuss these two issues in Sects. "Percolation Thresholds in the Continuum" and "The Critical Behavior of the Dynamical Properties in the Continuum" of our review. In contrast, we will not describe here the numerous natural and artificial systems for which the consequences of these issues are of great significance, leaving these for more specific reviews to come. We provided, however, a small representative list of such systems in Sect. "Definition of the Subject".

Percolation Thresholds in the Continuum

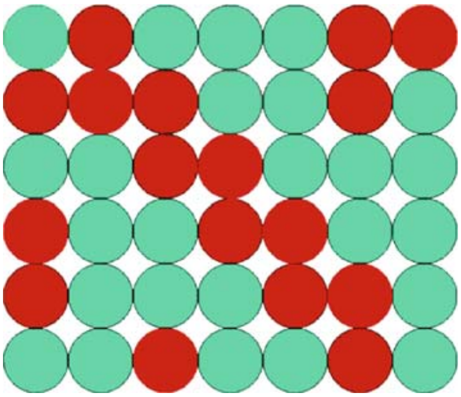
The Basic Concepts

The first attempt to characterize the percolation threshold in the continuum, which is usually credited as the beginning of the field of continuum percolation, was the 1970 [108] work of Scher and Zallen (S&Z), who based their argument on the "behavior" of a regular lattice, such as the one shown in Fig. 3. There, the red (say, the conducting) spheres occupy lattice sites, with a probability p^{s}, as in the lattice models described in Sect. "Lattice Percolation". The green (insulating) spheres correspond to empty sites in the lattice. The local connectivity criterion here is simply the red spheres occupation of nearest neighbor sites, as in the lattice, but here, if we consider the spheres as objects, the connectivity between two adjacent spheres, can be looked upon, as defined by a *single point contact* between two such spheres. Hence, if the diameter of the hard core spheres is equal to the lattice spacing the connectivity of the network stays the same as in the corresponding lattice. We can consider then the fractional volume of the conducting (red) phase which is simply given by $\phi = p^{\mathrm{s}} f$, where f is the filling factor of the system (the maximum fraction of the volume in a given lattice that can be filled by hard spheres of equal volume). Correspondingly, one can define the percolation threshold by the total occupied volume fraction of the conducting spheres, i. e. $\phi_{\mathrm{c}} = p_{\mathrm{c}}^{\mathrm{s}} f$. It is obvious then that $\phi - \phi_{\mathrm{c}} \propto p^{\mathrm{s}} - p_{\mathrm{c}}^{\mathrm{s}}$, and that the critical behavior of a given property Ξ in this case will be determined by

$$\Xi \propto (\phi - \phi_{\mathrm{c}})^{\chi} \,, \tag{6}$$

where χ is the corresponding-lattice critical exponent. S&Z noticed that ϕ_{c}, unlike $p_{\mathrm{c}}^{\mathrm{s}}$ (or $p_{\mathrm{c}}^{\mathrm{b}}$), is nearly a dimensional invariant, yielding values very close to 45 (area %) for all 2D lattices and very close to 16 (vol.%) for all 3D lattices. This brought them to the important realization that if we "shake" the system shown in Fig. 3 to yield a disordered system of touching spheres, its topology will not be significantly modified, and following the above close invariance, both ϕ and ϕ_{c} will be as good parameters for the characterization of the disordered arrangement of spheres, as p^{s} and $p_{\mathrm{c}}^{\mathrm{s}}$ are for lattices. The latter system can be also envisioned as a result of randomly dropping red and green spheres "gravitationally" into a container so that the "gravitation" secures the (single point) contacts between adjacent spheres [97]. Following the above concluded topology, it was of no surprise then that the latter approach confirmed the ideas and the results of S&Z.

We note of course that the above systems is by no means general or a prototype of a continuum system.



Continuum Percolation, Figure 3

A small portion of a square lattice that is comprised by "conducting" (red) and "insulating" (green) circles. The local connectivity is defined here by the "single point" contact of two red circles, and the global connectivity is determined by the spanning cluster of touching red circles

Rather, it represent an example of simple mapping of a continuum problem onto a lattice problem. While *the use of the fractional volume as a parameter*, which is equivalent in a way to p^s in lattices, was an important step in launching a theory for continuum percolation, their ϕ_c values were misused later, and to this date, for the characterization of continuum systems that are very different than the ones for which these, very particular values (secured single-point contacts), were calculated. For example, consider a system of red spheres that are embedded randomly in a continuous insulating matrix. Such a system represents many common composites (see Sect. "The Critical Behavior of the Dynamical Properties in the Continuum"). It can be easily appreciated that $\phi_c = 16$ vol.% in such a system of hard spheres yields a very sparse composite [1] and that the metallic (touching grain) conduction will be set at a much higher vol.% (of the order of 50 vol.% in the granular metals that constitute in fact an IRV system; Fonseca and Balberg [61]). Moreover, in practice there is no random system in which the "contact" or the "bond" between non permeable elements is "automatically" provided (if at all, it is approached only at the very high close packing limit of $\phi > 64$ vol.% [136,138]. Hence, there is no resemblance between the abundant common composite (with a continuous insulating matrix) and the models of S&Z and Powell, and thus the value of $\phi_c = 16$ vol.% is by no means a general approximate-invariant. This problem was discussed in more detail by Balberg and Binenbaum in 1987 [17].

In 1971 [112] Shante and Kirkpatrick (S&K) approached the percolation problem from the opposite end, i. e., considering permeable objects as in the case shown in Fig. 2. Like S&Z, it was obvious to S&K that the most promising approach to continuum systems was to try and map them onto a lattice problem, for which the theory was quite developed in the 1970's. They took what may be considered the most important single step in correlating lattice percolation and continuum percolation, proving that, at least in a specific systems, the topology of the two can be mapped onto each other. Their approach took into account an earlier empirical observation of Dalton, Domb and Sykes in 1964 [46] that in lattices, where z (the coordination number of the sites) approaches infinity, the value of zp_c^s extrapolates to 4.5 in 2D and to 2.8 in 3D. The quantity zp_c^s is, however, the average number of near-neighbors that are connected to a given site and it can be considered then as the number of occupied "bonds" per site. Hence, the quantity zp_c^s can be interpreted simply as the number of the occupied neighbor sites of a given site at the onset of percolation in the $z \to \infty$ limit. On the other hand, if one considers a system of permeable objects one can "implant" them anywhere in space. In particular for permeable spheres, the density of the possible centers of spheres which overlap a given sphere is also unlimited.

Following this observation, S&K conjectured that the topology of the $z \to \infty$ lattice and the continuum of permeable spheres is the same at the onset of percolation, i. e., that the number of "occupied" bonds per site, or the number of overlapping spheres for a given sphere, will be the same. Topologically, then, these two systems are expected to have the same onset of global connectivity. To test this idea, S&K considered the fact that the probability of randomly choosing a geometrical point that will be out of a given sphere of a volume v, in a unit size system, is $(1 - v)$. The probability that this point will be out of all the ρ_c spheres in this unit volume system, at the percolation threshold, is $(1 - v)^{\rho_c}$. In a large system that is statistically sufficient (i. e., $v \ll 1$ but $\rho_c \to \infty$) we can write the latter quantity as $e^{-v\rho_c}$. On the other hand, the probability of a point to be in any of the spheres (at the onset of percolation) is simply ϕ_c, the fractional volume that is covered by the ρ_c spheres. Hence, $\phi_c = 1 - e^{-v\rho_c}$. For a permeable sphere of radius a, in order not to overlap "another" sphere, its center has to be at a distance of $2a$ from the center of the "first" sphere. Correspondingly, the volume in which the center of the second sphere is not "allowed" to be (if no overlap is permitted) is not v but, rather, $8v$, or for hyperspheres in a D dimensional system it is, $v2^D$. On the other hand, the average number of centers of spheres that overlap a given sphere is simply $\rho_c v 2^D$. We have then that the quantity $\rho_c v 2^D$, is the critical number of bonds per sphere (or center, or site), B_c. Applying now their above conjecture, i. e. that $B_c = z p_c^s(z \to \infty)$ one finds that

$$\phi_{Dc} = 1 - \exp(-B_c/2^D)\,, \tag{7}$$

for a system of hyperspheres of dimension D.

Using the above mentioned "lattice" values of B_c for $D = 2$ and $D = 3$, with the prediction of Eq. (7), S&K found an excellent agreement with the Monte Carlo findings of $\phi_{2c} = 68$ area % and $\phi_{3c} = 29$ vol.%. In fact the B_c values were confirmed directly for permeable spheres later (in 1987 [16]) by Balberg and Binenbaum. The most important step of S&K was then the recognition of the fact that the number of bonds per site is the proper topological parameter to describe the connectivity in continuum as well as in lattice systems. In passing we note that following the above argument for any system of parallel aligned "regular objects" (such as boxes or cylinders) the relation $B_c = \rho_c v 2^D$ is expected to be obeyed [21,115].

In 1984 [22] Balberg et al. realized that the ideas of S&K can be generalized to non spherical objects, and even to objects that have no "volumes" of their own. Such is the case of two dimensional "widthless sticks" [20,25], where the local connectivity criterion is the intersection of two sticks, and the onset of percolation does not require any area of the sticks. This is in contrast with the systems mentioned so far. The way this generalization was conducted [9] followed the procedure of S&K with the topological expectation that for any system (regardless of the particular objects) B_c is the meaningful general topological parameter, and as such, it is expected to be a finite invariant. Following the B_c values in lattices of various dimensions [136], their values are intuitively expected to be between 1 and 5. In particular, B_c is always the number of the "other" objects centers (or sites) within the "excluded volume" V_{ex} of a given object, where the latter is the volume in which the centers of two objects must be in order for them to overlap. As such, the quantity V_{ex} is well defined for any permeable (or partially permeable object, see below). This yields that, in general, $B_c = V_{ex}\rho_c$. Naturally, the simple relation we had above for spheres Eq. (7) can be generalized now to all dimensions and for all types of permeable objects of a finite volume. This is done by noting that we can write ϕ_c by using, instead of $v\rho_c$ as above, the relation, $v\rho_c V_{ex}/V_{ex} = vB_c/V_{ex}$ for an object of hypervolume v and a hyper-excluded volume, V_{ex}, thus yielding that

$$\phi_c = 1 - e^{-B_c(v/V_{ex})}\,. \tag{8}$$

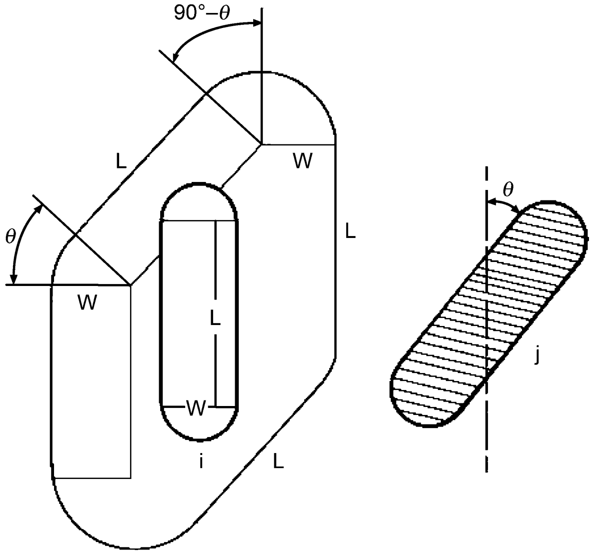
In 1986 [9] Balberg noted that Eq. (8) has far-reaching consequences, since it says that we can have a connected system with very minute content of the "percolating" or "conducting" phase ($v\rho_c$ or ϕ_c) involved. This follows from the fact, seen in Eq. (8), that ϕ_c can be as small as desired provided that v/V_{ex} is very as small as requireds. Indeed, this result was able to explain the $\phi_c \to 0$ values in porous systems, such as sedimentary rocks [7,110] and geological cracks [58], as well as in cermets such as cellular composites [94,128] in which the conducting phase is arranged along lines or planes [101]. Moreover, the fact that for lattices we have always finite p_c^s (or p_c^b) values that are larger than 0.3 for 2D and larger than 0.1 for 3D, made this result, of ϕ_c values as small as desired, quite surprising, and demonstrated that the mapping of the continuum onto lattices is not as simple as might have been conjectured from the pioneering works of S&Z and S&K (who were concerned only with systems of spheres).

The above generalization is also very useful for systems with distributions of object sizes, shapes and orientations, when a proper determination of the average excluded volume of the system, $\langle V_{ex}\rangle$, is carried out [22,53]. Hence, the above $B_c = \rho_c V_{ex}$ relation can be generalized to:

$$B_c = \rho_c\langle V_{ex}\rangle\,. \tag{9}$$

This result is very significant and important from the point of view of applications since if B_c is nearly an invariant (as to be expected topologically in view of the above, i. e. its value changes only in a very limited range, see also Sect. "The Origin of the Different **B**$_c$ Values in Systems of Different Objects" below) for various objects, one can predict the variation of ρ_c with the variation of the $\langle V_{ex}\rangle$ values in a group of objects. Hence, by assuming a constant B_c, trends in the behavior of percolation thresholds, ρ_c, as a function of the object's parameters, can be evaluated readily [9,37,83,103]. Moreover, one can predict the actual vol.% of the conducting phase (ϕ_c or $\rho_c v$) by just knowing the values of B_c and $\langle V_{ex}\rangle$ that can be well approximated by models (see below).

It turns out that the application of the concepts exhibited by Eqs. (8) and (9), beyond the very simple solution of $V_{ex} = 2^D v$ suggested for spheres (and for parallel convex shaped objects by Pike and Seager [95], and by Skal and Shklovskii, in 1974 [115]), has enabled the application of the above considerations to very many continuum systems. This was done in particular by the use and/or modification of the most celebrated model of non trivial systems, i. e., that of the three dimensional continuum system of capped cylinders [22]. These capped cylinders (of length L, radius W/2 and two capped spheres of radius W/2) span objects from spheres to cylinders with a single "close to an invariant" (see Sect. "The Origin of the Different B_c Values in Systems of Different Objects"), value of B_c. The basis of this model has been applied recently by Berham



Continuum Percolation, Figure 4
The volume and the excluded volume of two adjacent capped cylinders. The latter is for two cylinders (*i* and *j*) the angle between which is θ. (From [22])

and Sastry in 2007 [30] even to rather non-regular "wavy" shaped objects. The system of capped cylinders became then the prototype of continuum percolation systems of "non trivial" (non spherical) permeable and non-permeable objects. The volume of this capped cylinder is simply $v = (4\pi/3)(W/2)^3 + \pi(W/2)^2 L$, but its average excluded volume, shown in Fig. 4, is given by

$$\langle V_{\text{ex}}\rangle = (32\pi/3)(W/2)^3 + 8(W/2)^2 L + 4(W/2)L^2\langle\sin\theta\rangle\,, \tag{10}$$

where θ is the angle between the two of possible intersecting cylinders, and the average is over the particular distribution of the θ values [22,89]. The "non trivial" aspect is manifested by the fact that $\langle V_{\text{ex}}\rangle$ is not proportional to v, and that in the large aspect ratio case (i. e. $L \gg W$) we have that $v/\langle V_{\text{ex}}\rangle \propto W/L$, while for the small aspect ratio case we recover the above mentioned, trivial $v/V_{\text{ex}} = 1/8$ result for spheres. The system described here can be shown to represent very many natural and artificial systems where the objects are elongated [38] or have a negligible volume [101]. We note also that another system that is very helpful for the description of various systems is that of disks with a radius a, and a thickness t. These disks have a volume $v = \pi a^2 t$, while their excluded volume, in the isotropic orientation case, is given by $V_{\text{ex}} = \pi^2 a^3$ [37]. It is apparent that in the thin disk limit ($t \ll a$), we get that $v/V_{\text{ex}} \propto t/a$, explaining the very small ϕ_c values observed, as mentioned above, in sedimentary rocks, in systems of geological cracks and in cellular composites. In passing we note that the low dimensionality of the conducting phase inclusions in these systems explains the correlation between low percolation thresholds and the high critical conductivity-exponents (that are associated with low dimensional tunneling systems, see Sect. "The Average Resistance in the Case of Tunneling-Percolation").

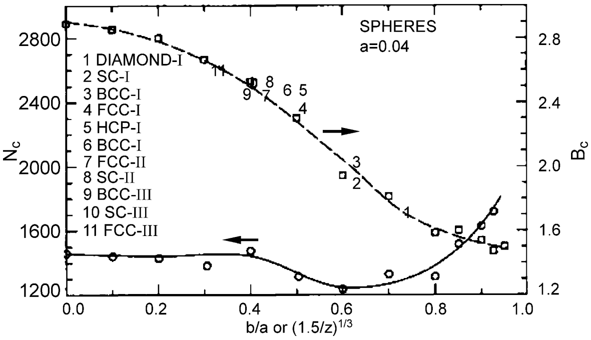
Following the above, let us try and guess now the possible variations of B_c as one goes from systems that consist of permeable (soft-core) spheres to quasi-permeable spheres with a hard core of radius b and a wrapping soft shell with a thickness of $a - b$. We mentioned already that B_c is 2.8 in the 3D soft core limit. We also know then, as found directly [17], that the value of B_c decreases, in the transition with the increase b, towards a value of 1.5. What we have available empirically are, those limits, the fact that what influences the behavior are volume quantities, and the above expectation that B_c will decrease between the two limits as $b/a \to 0$. Hence, we can simply expect empirically that

$$B_c = 1.5 + 1.3\left[1 - (b/a)^3\right]\,. \tag{11}$$

Considering the above mentioned topological meaning of B_c we recall that for lattices we realized already that in the $z \to \infty$ limit, $B_c = 2.8$, and we note that for the other limit we can apply the conjecture of Ziman in 1979 [138] (following the fact that zp_c^b is an invariant in lattices such that $zp_c^b = 1.5$ in 3D) that the smallest z (existing or imaginary) needed for percolation in 3D lattices, is 1.5. In this case we expect then that the approximation:

$$B_c = 1.5 + 1.3\left[1 - (1.5/z)\right]\,, \tag{12}$$

will be reasonable for lattices. If the connecting topology in the continuum and in lattices is the same (as found to be the case in the $b/a \to 0$ limit by S&K), B_c should scale equally well with $(b/a)^3$ or $1.5/z$. Indeed, as is clearly seen in Fig. 5, Monte Carlo simulations on a series of systems that represent the "soft core" to the "hard core" transition [17] have confirmed this mapping of the continuum system onto the lattice system. This establishes that, indeed, the underlying connectivity is the same in lattices and in continuum systems of partially permeable spheres (i. e., beyond the soft core S&K limit). A reproduction of the corresponding B_c values themselves can be achieved and explained phenomenologically as we will see in the following section. We noted already that from the practical point of view, one can generally predict the percolation threshold ρ_c once one knows the value of B_c (or an approximate of it) and the value of V_{ex}. In the present case we have the value of B_c (Eq. (11)) and



Continuum Percolation, Figure 5
Monte Carlo results for the dependence of the critical number of spheres in the system, N_c (denoted by circles, solid curve) and B_c (denoted by squares, dashed curves) on the spheres' b/a ratio in a continuum system of semi-permeable spheres. Also shown is the dependence of B_c on $(1.5/z)^{1.3}$ for three dimensional lattice systems. (From [16])

$V_{ex} = (32\pi/3)(a^3 - b^3)$. Hence, as the hard core limit is approached, and B_c gets the value of 1.5, we expect that ρ_c ($= N_c$ for a system of a unit volume) will diverge (of course up to the close packing limit that is possible). This is well exhibited by the Monte Carlo determination of N_c that is shown in Fig. 5. We note that the above derived phenomenological behavior does not tell us why the value of B_c itself decreases in the $b/a \to 1$ and $z \to 1.5$ limits (see Sect. "The Origin of the Different B_c Values in Systems of Different Objects").

A somewhat different view of the hard core-soft shell problem leads us to examine continuum percolation properties that can be considered to be even more remote from those of lattice systems. A conspicuous case of such a property is the effect of objects interaction due to a potential associated with the objects [41,90]. To introduce this scenario, we can consider the above hard core-soft shell sphere as being composed of a region of infinite repulsive interaction at distances r (from its center) smaller than $2b$, and a shell of zero potential for $r > 2b$. We may further look at the problem as a one in which the density of the objects is not uniform around an object (here e. g. $\rho = 0$ for $r < 2b$ and $\rho = \rho_s$ for $r > 2b$). If we use then the corresponding radial distribution function $g_R(r)$ of the particles (or objects' centers), the average number of bonds per site will be given by $B = \langle\rho\rangle \int_0^{2a} g_R(r) \mathrm{d}^3 r$, where $\langle\rho\rangle$ is the average density of objects in the whole system and a is the radius of the soft shell (that determines the local connectivity of the system). If we have a more complicated potential (i. e. when $g_R(r)$ has values different from 0 or 1 as above), B, and thus B_c, will depend on the parameters of this distribution, such as on the interaction potential that determines the $\rho(r)$ dependence [4]. A case that was considered in detail and is relevant, for example, to microemulsions [69], is that of an attractive potential shell that "wraps" the hard core ($b \leq r \leq (1 + \lambda)b$) and has a constant, non-zero value, in the interval of λb. In that case, it has been shown first by Bug et al. [35] (and later by Drory, Balberg and Berkowitz [54,55]) that interactions can raise or lower the percolation thresholds, i. e., the value of the global ρ_c value, according to the details of the magnitude of the potential and the ratios between b, λ and a.

Physically, this competition is not too surprising, since the attractive potential will cause the increase of the density of objects in the attractive potential region, but for a globally given ρ, it will cause a reduction of the density outside this region. The two effects will be manifested by an increased probability for the formation of the connected clusters, on the one hand, but by the reduction of the connectivity between different clusters, on the other hand. The interesting observations are then the non-trivial results of the interplay between these two effects that is determined by the system parameters, i. e. by the dependence of the values of B_c and ρ_c on them.

From the point of view of continuum percolation theory, the important conclusion is that even in this explicit interaction case we can characterize the system by its B_c value, correlate it with system parameters, and apparently generalize it to systems where the mapping onto lattice models is, conceptually and/or computationally, not obvious. On the other hand, it is apparent that such derivations can account for general connectivity problems where ρ or other "bonding strength" parameters are not uniform in space or in the network.

The Generalized Thresholds and the Critical Behavior in the Continuum

Let us turn now to see how the above mentioned quantities play a role in the determination of the most important parameter in the study of the critical behavior, i. e. the proximity to the percolation threshold. Following the fact that $B_c = \rho_c \langle V_{ex} \rangle$, we simply have the relation $B - B_c \propto \rho - \rho_c$. Hence, for B very close to B_c and ϕ very close to ϕ_c (which is the interesting "critical" range), we can further generalize this relation (by using Eq. (8)) to

$$\phi - \phi_c \propto B - B_c \propto \rho - \rho_c \,. \tag{13}$$

This relation is a very important one from the practical point of view, since ϕ is the only parameter that one can determine readily in experimental studies. However, the requirement for small $\phi - \phi_c$ values must be kept then in mind when the critical behavior is studied in continuum systems. In fact, another very important precaution has to be taken into account when studying the critical behavior as a function of $\phi - \phi_c$. We saw above that in many (e. g. sedimentary rocks and cellular composites) systems, ϕ_c can be very small. Hence, a small $\phi - \phi_c$ (with respect to unity) does not actually tell us about the proximity to the threshold. The parameter that one should consider is a normalized proximity to the threshold, such as $(\phi - \phi_c)/\phi_c$. We note that the latter problem is not severe in lattices of the "practical" ($D = 2$ or $D = 3$) dimensions because p_c^s and p_c^b for all these lattices are larger than 0.12 [136]. On the other hand, as noted by Grimaldi and Balberg in 2006 [70], the overlooking of the above two precautions is one of the reasons that when considering experimental results, improper comparisons with lattice-like critical behavior have been suggested. Following the above discussion on the soft core-hard core transition we also note that in the small $B - B_c$ limit, Eq. (13) covers all systems, including those where ϕ represents te hard core fractional volume in partially permeable (hard-core/soft-shell) objects. This is easy to appreciate since at the close vicinity of the percolation threshold the ratio of the hard core volume to the entire (hard core and soft shell) volume can be considered a constant and since we know that the relation given by Eq. (13) holds in limit as shown by the S&Z case. Indeed, many computer simulations on many continuum systems [15,65,66], where the systems are well defined (and ϕ_c can be approached much more closely than in the experimental studies on "real" systems), have revealed the critical behavior of lattices. This is quite an important realization when one tries to explain the critical behavior in composites, considering the fact that the rigorous development that we described above was given for permeable objects.

From our above discussion we can expect then *the same critical behavior* for all types of objects, and may further conjecture that this is the case when other local connectivity criteria are applied. In fact, thus far all Monte Carlo simulations have confirmed these expectations. We note, however, that a high precision study of ϕ_c and the critical exponents of a system of permeable spheres, was claimed to indicate a small variations with respect to the corresponding lattice values of the same exponents [100]. This conclusion can be contrasted by the fact that rigorous exact mapping of the permeable circles system onto the fine grid lattice system was demonstrated by McCarthy already in 1987 [86]. As of now, however, further work will be needed in order to determine whether the above small deviation is a matter of accuracy or proximity to ϕ_c (see the discussion following Eq. (13) above), or if there is still something more fundamental here that has not been satisfactorily studied thus far.

The Origin of the Different B_c Values in Systems of Different Objects

So far we have taken B_c to be an empirical parameter, as are the p_c^s and p_c^b quantities in lattice percolation, and the application of the relation given by Eq. (9) was proven extremely useful for the determination of trends in the variation of measurable percolation thresholds (i. e., ρ_c) when a constant (system independent, for a class of materials, e. g. boxes with a different aspect ratio) B_c value was assumed. However, the various values of B_c, while being confined to the 1–5 range (see above) are not the same for systems of objects of different shapes. These differences are important to understand since, as we discussed above, this is the only parameter that accounts for the global topological connectivity of percolation systems in the continuum. We note in passing that while the B_c values can be derived rigorously (see Sect. "Rigorous Determination of the Percolation Threshold"), their relation to the particular objects in the system under consideration is not a priori transparent from the corresponding derivations. At this point we concentrate then on the reasons for the different B_c values for different "types" of objects. For example, B_c *is 2.8 for permeable spheres and is only 2.6 for parallel permeable cubes*. While this difference seems minute (in comparison with the larger differences to be mentioned below), it is a convenient case for the illustration of the concept of "pointedness" that enables to explain the observed differences in the B_c values.

The reason for the above mentioned variations in the B_c values had hardly been considered prior to the work of Alon, Balberg, and Drory in 1991 [4] that yielded an insight into the apparent "different topology" associated with different objects. They derived this understanding using a heuristic or "semi-rigorous" argument, which will be illustrated here by the different B_c values for spheres and cubes, as well as by the variation of B_c in the soft core to hard core transition.

Let us consider then a system of permeable spheres of radius r_0 and a system of parallel permeable cubes with an edge a_0. Choosing their size such that they have the same excluded volume $\langle V_{ex} \rangle$ (in this case, also the same volume), we note that $\langle V_{ex} \rangle = (4\pi/3)(2r_0)^3 = (2a_0)^3$. Now let us consider the maximum distance between the centers of two spheres that can overlap. This distance is $2r_0$, and thus the possible distances, l, between the centers of two partially overlapping spheres is in the range $0 \le l \le 2r_0$. In contrast, for three dimensional cubes that are also connected by a partial overlap (and have the same $\langle V_{\mathrm{ex}} \rangle$), the corresponding range is $0 \leq l \leq \sqrt{3}a_0$, i. e., $0 \leq l \leq 2.7r_0$. Hence, since the density of the centers of the objects is uniform in both cases, on the average, two partially overlapping cubes can "on the average span" a larger distance than that of partially overlapping spheres. Correspondingly, in order to span an "infinite" cluster, we will need a smaller number of cubes than spheres, i. e., the value of ρ_c will be smaller in the case of cubes compared with the case of spheres. Since we have chosen V_{ex} to be the same in the above two systems, the corresponding value of ρ_c and thus the value of $B_c (= \rho_c V_{\mathrm{ex}})$ will be lower in the case of the cubes. Of course, the exact ratio between the B_c values of the two systems cannot be derived from the qualitative-simple illustration given here, but a more refined argument has been shown to heuristically yield such ratios exactly [4].

Since the apparent difference between the cube and the sphere is that the former has corners ("points"), one can say that the "pointedness" (that was defined by Alon, Balberg and Drory accurately) brings about the lower value of B_c. We can correspondingly conclude that since the average "covered" l "per excluded volume" will be larger with a higher aspect ratio of elongated objects, such as boxes that are isotropically distributed in space, we can further expect that for elongated boxes, of the same $\langle V_{\mathrm{ex}} \rangle$, the value of ρ_c will decrease with this ratio. For these, as mentioned above, the corresponding trend of the decrease of B_c with the aspect ratio was confirmed both rigorously and computationally [53,131] to decrease from about 2.7 to 1.2 for aspect ratios between 1 and 500. The latter trend was confirmed also for other types of permeable elongated objects [30,103].

A similar argument can apply for spheres of a hard core of radius b and a soft shell with a wrapping thickness of $a - b$. The "local span" of the two objects' in the $b = 0$ case is in the $0 \leq l \leq 2a$ range, while for the hard-core sphere with a soft shell it is $2b \leq l \leq 2a$, so that the average l value is larger in the latter case. Correspondingly, for the same average ρ in the system, the "span" of larger-hard core spheres will be larger, or the value of ρ_c will be smaller in the hard core-soft shell case, suggesting the decrease of B_c from 2.8 to 1.5, as we saw already in Fig. 5. The above result is of great conceptual importance, since it shows that while the phenomenological theory of the excluded volume theory was developed for permeable objects, *all consequences regarding the trends* in the values of the percolation thresholds (such as the effect of the aspect ratio) *apply also upon the approach to the hard core limit*. This can be illustrated simply by replacing W by $W - 2b$, where b is the hard core radius, in Eq. (10), and noting that as the aspect ratio increases the last term will eventually become the dominant term for any $2b/W$ ratio. In fact from the above arguments concerning the aspect ratio and the soft core to hard core transition, we conclude that the two effects can be compounded. This is also important from the practical point of view, since one can be guided, for example, for "real" composites that are composed of hard core objects [19], by the many works and the numerous simulations [21] that were concerned with the corresponding (easier to study) systems of permeable objects. A case in point is the decrease of the percolation threshold in carbon-black-polymer composites as the conducting particles become elongated (or, as better known, of "higher structure" [10,11,102]).

Rigorous Determination of the Percolation Threshold

Thus far we have considered the derivation of the concepts and the empirical trends in the behavior of the percolation thresholds as a function of the correlation between, and the properties of, the individual objects, as well as the corresponding global averages such as the anisotropy of a system. However, as we saw above, the particular fundamental values of the thresholds were either derived from lattice quasi-invariants (notably $p_c^s z$ for $z \to \infty$), heuristic arguments or computer simulations. Noting that a "complete" theory with a firm basis requires a rigorous derivation of the fundamental quantities from first principles, it is always a challenge to achieve such a derivation and the theoretical justification for the experimental and computational observations and their trends. This challenge was recognized also for the problem at hand, i. e., the rigorous finding of the "fundamental-topological" percolation threshold B_c (e. g., $B_c = 4.5$ for circles and $B_c = 2.8$ for spheres), which we encountered in the previous sections.

The above challenge was appreciated in 1977 [73] when Haan and Zwanzig considered the existence of bonds, defined by the overlap of permeable objects, and applied a graph theory for the analogous lattice-bond percolation problem. This attempt, which involved a lengthy cluster enumeration and high-order overlap integrals, has not been followed however by others, probably because of the latter reasons. The breakthrough in trying to find a rigorous route which is physically more transparent can be attributed then to the work of Coniglio, DeAngelis and Forlani in that year#160;[44]. They applied for the above purpose the formalism of the connectedness functions that is well known in the theory of liquids [74]. For the sake of brevity we will describe here their approach by the following simple "intuitive" arguments.

If we define a local (direct) connectedness criterion (e. g., overlap between permeable spheres) and a global connectedness (e. g., being on the same cluster), we can derive a direct connectedness function $C(\boldsymbol{x}, \boldsymbol{y})$, such that if $\boldsymbol{r}$ is the vector $\boldsymbol{x} - \boldsymbol{y}$ connecting "directly" the two objects at $\boldsymbol{x}$ and at $\boldsymbol{y}$, $\rho(\boldsymbol{r})C(\boldsymbol{r})\mathrm{d}^3 r$ is the probability that given a particle at $\boldsymbol{r} = \boldsymbol{0}$, there will be a particle, in the volume $\mathrm{d}^3 r$ around $\boldsymbol{r}$, that is locally connected to the particle at $\boldsymbol{r} = \boldsymbol{0}$. Here, $\rho(\boldsymbol{r})$ is the particles' (or objects') density in the system. The quantity $C(\boldsymbol{x}, \boldsymbol{y})$ accounts, in addition to the direct connection, for all possible additional "indirect" connections between the "directly connected" these two particles, and thus it can be presented by corresponding "closed diagrams" [56]. Similarly, one can define a "total" pair connectedness function $g(\boldsymbol{z}, \boldsymbol{y})$ between two objects (say, at $\boldsymbol{z}$ and at $\boldsymbol{y}$) which involves all the connecting (direct and indirect) routes between them. The function $g(\boldsymbol{r})$ (where $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{z}$) can be defined then by the probability, $\rho(\boldsymbol{r})g(\boldsymbol{r})\mathrm{d}^3 r$, of finding a particle in the volume element $\mathrm{d}^3 r$ around $\boldsymbol{r}$, which is connected (i. e., belonging to the same cluster) to the particle that is given to exist at $\boldsymbol{r} = \boldsymbol{0}$. Since $g(\boldsymbol{x}, \boldsymbol{y})$ is composed of "direct" connectivity steps such as $C(\boldsymbol{x}, \boldsymbol{z})$ and $C(\boldsymbol{z}, \boldsymbol{y})$, the "total" $g(\boldsymbol{x}, \boldsymbol{y})$ function can be expressed by the corresponding (Ornstein-Zernike) "chain" rule

$$g(\mathbf{x}, \mathbf{y}) = C(\mathbf{x}, \mathbf{y}) + \int \rho(\mathbf{z})C(\mathbf{x}, \mathbf{z})g(\mathbf{z}, \mathbf{y})\mathrm{d}^3 z\,, \tag{14}$$

where $\rho(\boldsymbol{z})$ is the density of particles (or centers of objects) in a volume of $\mathrm{d}^3 z$ around $\boldsymbol{z}$. For simplicity we will assume in the following that $\rho(\mathbf{z}) = \rho$, where ρ is a constant-uniform density in our problem. The crucial step that connects relation (14) with the present continuum percolation problem, i. e. with the average cluster size, is the consideration of its Fourier transform. For a "wave vector" $\boldsymbol{k}$ this relation yields that for the transformed functions $C(\boldsymbol{k})$ and $g(\boldsymbol{k})$ we will have that $g(\mathbf{k}) = C(\mathbf{k})/(1 - \rho C(\mathbf{k}))$. If we are interested in the integral $\int g(\boldsymbol{r})\mathrm{d}^3 r$, i. e. in the Fourier transform $g(\boldsymbol{0}) \equiv g(\boldsymbol{k} = \boldsymbol{0})$, we get in particular that

$$g(\mathbf{0}) = C(\mathbf{0})/(1 - \rho C(\mathbf{0}))\,. \tag{15}$$

The new important step made in the work of Coniglio, DeAngelis and Forlani in 1977 [44] was that, as can be appreciated intuitively, the average cluster size S in the system is simply given by

$$S = 1 + \int \rho g(\boldsymbol{r})\mathrm{d}^3 r\,, \tag{16}$$

where the unity here represents the given initial object (with a center at, say, $\boldsymbol{r} = \boldsymbol{0}$), $\rho\mathrm{d}^3 r$ is the probability of finding an object center in the space element $\mathrm{d}^3 r$, and $g(\boldsymbol{r})$ is essentially the percolation correlation function [121], i. e., the probability that, if there are objects centers at $\boldsymbol{r} = \boldsymbol{0}$ and $\boldsymbol{r}$, they will be on the same cluster. From Eq. (16) we have that $\rho g(\boldsymbol{0}) = S - 1$ and thus using Eq. (15), we obtain that

$$S = 1/(1 - \rho C(\mathbf{0}))\,. \tag{17}$$

Hence, knowing that at the threshold $S \to \infty$, we get that $\rho_c = 1/C(\mathbf{0})$, and thus, in principle, "all that is left" is to calculate $C(\mathbf{0})$.

There are basically two approaches for the calculation of $C(\mathbf{0})$. The first one is immediately called for by Eq. (14), from which it is apparent that if another relation between $g(\boldsymbol{x}, \boldsymbol{y})$ and $C(\boldsymbol{x}, \boldsymbol{y})$ is known, one can solve both functions and then find the value of S. Such a relation can be derived from corresponding approximations. One of these approximations, known as the Percus–Yevick closure [74]), was used for the solution of the problem at hand. Since the application of this closure was of a limited success [49]), and since it would require a deviation from the simple outline of our main argument regarding the rigorous derivation of ρ_c, it will not be detailed here.

The conceptually simpler and physically more transparent approach to the problem can be attributed to the suggestion of Bug, Safran and Webman in 1985 [34]. They were able to determine the initial terms in a (diagrammatic) series expansion of the form $C(\mathbf{0}) = \sum c_n \rho^n$, and to relate the initial coefficients to the concept of the excluded volume. This already enabled them to derive the trends in the behavior of the thresholds. However, this use, of what one may call an order-by-order (or series) approach, did not yield (in that work) actual values for the thresholds. The full utilization of this order-by-order approach came in the work of Drory and coworkers in 1990 [3], 1991 [53] and 1997 [56]. In passing we note that Drory also got in 1996 [51] and 1997 [52] a firm basis for the conjecture that continuum percolation is a phase transition, as proven previously [62] to be the case of lattice percolation.

Turning to the utilization of the latter approach we noticed already that in principle, $C(\mathbf{0}) = 1/\rho_c$. However, in practice, this does not provide directly an accurate enough value for ρ_c for the rather short series that were obtained thus far. On the other hand a more successful, albeit biased, approximation has yielded, as mentioned below, very accurate results. For the description of the latter we start then from Eq. (17) using the well understood relation that we had considered before, i. e., replacing ρ by $B = \rho V_{\mathrm{ex}}$. Hence, the aim of the procedure is to determine the, by now, well understood and well-characterized parameter B_c. Following the above, we can write S in terms of a power series of the form $S = \sum a_n B^n$, which is similar to the form used for the derivation of S by a power series of p in lattice percolation [121] In the present case, having the quantity $C(\mathbf{0})$ expressed as a power series of the form $C(\mathbf{0}) = \sum b_n B^n$ enables the comparison of the coefficients on both sides of Eq. (17), yielding then relations between the coefficients a_n and b_n. Here, for example, $a_0 = 1$, $b_0 = 0$, $a_1 = b_1$, $a_2 = b_2 + b_1 a_1$, $a_3 = b_3 + b_2 a_1 + b_1 a_2$, and $a_4 = b_4 + b_3 a_1 + b_2 a_2 + b_1 a_3$. Having these relations and implementing the fact (see Eq. (1)) that S has a critical behavior of the form

$$S \propto |B - B_c|^{-\gamma} , \tag{18}$$

where γ is the well-established critical exponent for lattices and, as expected from universality, also for the continuum (see above), one can fit the series of S to the behavior described by Eq. (18), in order to find the value of B_c. Of course the "only" input needed for the implementation of this procedure is then the determination of, as many b_n coefficients as possible and a good extrapolation of S. However, the derivation of the coefficients b_n is not trivial and becomes more and more complicated with increasing n [53]. So far these values of b_n have been determined up to $n = 5$ for permeable hyperspheres, yielding, however, very accurate values for B_c that are only a few % off the Monte Carlo estimates for corresponding $D = 2$ to $D = 6$ systems [3]. Similar successes for systems of elongated boxes [53] and hard core-soft shell cubes [56] have been already obtained. In spite of those achievements, further basic and computational developments in the study of this subject are called for. These will be outlined in Sect. "Future Directions".

The Critical Behavior of the Dynamical Properties in the Continuum

As we pointed out above, within the limited framework of this review, it will not be possible to consider all the many various dynamical properties (mentioned in Sect. "Definition of the Subject") that are determined by the added locally variable features in continuum systems and compare their behaviors with those derived for lattices. Correspondingly, we will confine our discussion to the conceptually most conspicuous representative, which is also the most studied, dynamical property, i. e., the electrical conductivity. It can be shown, as done by Rubin et al. in 1999 [102] for the electrical noise, that the basic arguments associated with the critical behaviors of the other properties are similar in principle, and that the specific details of those behaviors can be derived along the same lines that are used here for the electrical conductivity.

In the previous sections, we saw that one property that makes continuum percolation an area of a much wider diversity and application, in comparison with what one finds in lattice percolation, is the percolation threshold. The global properties of a percolation system were own there to depend on the details of the systems, where these may vary in many respects from one system to another. On the other hand, as far as the critical behavior is concerned, we have seen that the global behaviors of the geometrical-statistical properties are expected to be exactly the same as those found in lattices, provided that one uses the topological concept of bonds per site. In those cases a simple mapping, which is intuitively straightforward, of the "percolating phase" content onto the lattice occupation probability, can be made. Following the fact that for a given lattice there is a constant ratio between p^{b} and p^{s} [136], we will simply use here p and p_c as a generalized "conducting phase" content and describe the critical behavior of the dynamical properties in the continuum using these parameters. When needed however, the relation between these and the commonly measurable quantities, such as the corresponding partial volume contents of the conducting phase, τ and τ_c, will be emphasized.

Does the above behavior of the geometrical-statistical properties give us a hint as to the critical behavior of the dynamic (or "physical") properties of the system? To answer this question, let us consider the two possible features that are fixed in lattices, and may be random (or follow another distribution) in the continuum. This is easy to do by reexamining the system shown in Fig. 2 as a system of permeable spheres. In this figure we can see that the continuum randomness is manifested by the positions of the centers of the spheres and the random-like degrees of overlap between two adjacent spheres. If these spheres are envisioned as pores filled with a conducting fluid, the latter feature determines the local resistance between a pair of them. This will yield a distribution in the resistance values of the actual resistors that connect pairs of pores in the system. The simplest approach to evaluate the contribution of the latter effect (that has no counterpart in lattice percolation) to the global conductance of the system is to assign a resistor, with a resistance value that is taken from the corresponding distribution, to the occupied bonds on a given lattice [111]. A more direct simulation of the system that involves also the random "implantation" of the spheres has also been shown [23] to yield the predicted results that will be given below. The purpose of the following discussion is then to show how corresponding distributions of the local resistance values can guide us to the determination of the critical behavior in "real" continuum systems.

The simplest system to realize then is the lattice system of Fig. 1, when one assumes that the "bond" or the conducting "bar" between two occupied nearest neighbor sites corresponds to a resistor, the value of which is taken from a given distribution. A priori, one does not foresee too much of a surprise here, and a "universal" (lattice-like) behavior of the conductivity is expected, except that the individual local resistance value, $\mathbf{r}_0$ (associated with R_ξ; see Eq. (3)), will be replaced now by an average value of the local resistors, $\langle \mathbf{r} \rangle$. In fact, the finding of a universal value for t in many "real" systems [2,19,25] suggested that this distribution-effect will not yield a critical behavior that is different from the behavior in lattices, as appears to be the case for the geometrical-statistical properties. Considering this problem Kogut and Straley (K&S) developed in 1979 [82] a simple "toy" model that suggested that for some distributions, the above universality will hold, while for others, it will not. It took six years, however, until the missing link of the "toy" distribution and "real", or continuum, systems was made by Halperin, Feng and Sen in 1985 [72], and attention was given finally to the results of K&S. The work of K&S, which can be considered as the cornerstone of the field of "non-universal" behavior of dynamical properties in the continuum, will be reviewed below, after a modified-generalized account of the basic physics that is involved in their model will be given.

What apparently took the above-mentioned six years in order to understand, even in principle, the emerging numerous experimental observations of non-universal t values in the 1970's and 1980's (i. e., $t \neq t_{\text{un}}$; [10,11,94]), and to appreciate the theory of K&S, was *the unsatisfied need to find a distribution function* of the resistors values in a given "real", natural or artificial, system. This breakthrough came then with the work of Halperin, Feng and Sen in 1985 [72], who were able to show that in sedimentary rocks the values of the local resistors are essentially distributed according to the "toy model" function of K&S. In fact, it will be emphasized all along this part of the review that the determination of the resistors distribution function is the crucial step for the understanding of the critical behavior of the electrical conductivity (or other dynamical properties) in a given continuum system. The basic physics of this key issue, which has not been reviewed before, will be described below in some detail.

The Basic Physics of the Non-universal Behavior of the Conductivity

The physics of the critical behavior of the global resistance in continuum systems can be described as follows. Suppose we have a distribution $f(g)$ of the values of the individual conductors (the resistors that occupy the bonds) in our infinite lattice. Let us assume that the system has a fractional bond occupation p and that the percolation threshold is p_c such that $p > p_c$.

We have seen that in lattice percolation (Sect. "Lattice Percolation") the resistance of the system can be written as

$$\mathbf{R}_L = \mathbf{r}_0 A_R L_1^{\xi} (L/\xi)^{2-D} \, , \tag{19}$$

where A_R is some constant (i. e., ξ, or $p - p_c$, independent) parameter that accounts for the local and global geometry of the system and $\mathbf{r}_0$ is the resistance value of the individual occupied bond in the system. The question of interest here is then the behavior of $\mathbf{R}_L$, when we have a distribution of the individual resistance values (as expected in continuum systems) rather than a single $\mathbf{r}_0$ value. Let us denote the conductance of the individual resistor by g, so that the local value of the resistance is given by g^{-1}. Now suppose that the g values are taken from a distribution $f(g)$ that is confined to the range $g_1 \leq g \leq g_2$. Let us also assume that in our system there is a largest possible g_2 value (e. g., a zero distance tunneling or a complete overlap of IRV spheres; see Sect. "The Local Resistors and Their Distributions", below). The most important assumption in the following consideration is however that there is no correlation between the location of the resistor in space and its value, and thus one would expect that if there is such a distribution, $\mathbf{r}_0$ in Eq. (19) should be simply replaced by $\langle \mathbf{r} \rangle$ where

$$\langle \mathbf{r} \rangle = \langle g^{-1} \rangle = \int_{g1}^{g2} g^{-1} f(g) \mathrm{d}g \, . \tag{20}$$

We note in passing that such an average is meaningful if it is taken on $1/g$ (rather than on g), since it is the series connection of resistors that determines the resistance of the link (see Eq. (3)).

In principle, g_1 can be any value in the interval $0 < g_1 \leq g_2$ (a $g_1 = 0$ value corresponds to an unoccupied bond), and it is apparent that for a well defined finite-constant $g_1 > 0$ value, the critical behavior described by Eq. (19), with $\langle \mathbf{r} \rangle$ replacing $\mathbf{r}_0$, will be the same as in Eqs. (3)–(5), i. e., the universal (lattice-like) critical behavior will be obtained. What if $g_1 \to 0$ (but $g_1 \neq 0$; see above)? In this case it is apparent that the result will depend on $f(g)$. If $f(g)$ decreases as $g \to 0$, the value of the average $\langle \mathbf{r} \rangle$, as defined by Eq. (20), will be finite, as above. On the other hand, if $f(g)$ is a constant or it increases towards $g = 0$ (i. e. $f(g)$ diverges at $g = 0$), the value of $\langle \mathbf{r} \rangle$ will be ∞. In other words, if there are "enough" $g \to 0$

conductors in the system, the average will be determined by them and we will also get that $\mathbf{R}_L \to \infty$.

To consider however the behavior of $\mathbf{R}_L$ in more detail (i. e. the way in which $\langle \mathbf{r} \rangle$ diverges) for a non-decreasing $f(g)$ distribution as $g \to 0$, we use the following estimate of $\langle \mathbf{r} \rangle$ for the $g_1 \to 0$ case as follows. Let us list the values of the local conductances in the system in a descending order. Now, electrical conductance in the system is only possible if $p > p_c$, and thus we can choose the fraction (or subset), p_c of the p local conductors in the system, that contains the "top" g, p_c-values from that list. Following our basic assumption that there is no correlation between the location of the bond and its attached g value, we can use the fact that any randomly selected subset of p_c conductors is, by definition, a percolating cluster, and conclude that the above chosen p_c subset of the top value conductors constitute a percolation cluster. Of course, the largest conductance value in the system, g_2, is also the largest possible g value in the so chosen subset. For convenience let us choose now the value of g_2, by the relation $F(g_2) = 1$, where $F(g)$ is the indefinite integral of the function $f(g)$. This is done without loss of generality since all other g values in the system can be normalized accordingly. Then, the lowest g value of the conductors in the above subset p_c of p has the corresponding normalized value g_c. Mathematically, this g_c value is given then by

$$p\left[\int_{\mathrm{gc}}^{g2} f(g)\mathrm{d}g\right] = p_c \,. \tag{21}$$

This yields that $F(g_c) = (p - p_c)/p$, and thus that $g_c = F^{-1}[(p - p_c)/p]$, where F^{-1} is the inverse function of F. We have then a very significant result that connects the conductors' values in the system with the proximity to the percolation threshold. In particular, it is the nature of the F^{-1} function that determines the "pace" at which $g_c \to 0$, as p is made to become closer to p_c. The very important observation here is that for $p = p_c$ we have to exhaust all the resistors in the system and thus $g_c = 0$. The question that we would like to resolve is, following the variation of g_c as p approaches p_c, how will $\langle \mathbf{r} \rangle$ change as a consequence of this variation.

As above, we can distinguish here between three cases, but now we can also get some physical insight as to the corresponding critical behaviors. If $f(g)$ is a constant (in the $0 \leq g \leq 1$ interval) g_c will approach 0 at exactly the same "pace" as $p - p_c$. If $f(g)$ increases as $g \to 0$ (i. e. $f(g)$ diverges as $g \to 0$), g_c will approach $g = 0$ "faster" than $p - p_c$, and if $f(g)$ decreases toward $g = 0$, g_c will approach $g = 0$ at a "slower pace" than p approaches p_c. In other words, $1/g_c$ which is a measure of the resistors that determine the value of $\langle \mathbf{r} \rangle$ (see below), will not diverge in the latter case when $p \to p_c$. On the other hand, if $1/g_c$ diverges as $p \to p_c$ there is a possibility that a dynamic property that depends on g will also diverge then. Correspondingly then the above $p - p_c$ dependence of $1/g_c$ on $p - p_c$ dependence of $\langle \mathbf{r} \rangle$ may contribute to a non universal behavior of $\mathbf{R}_L$.

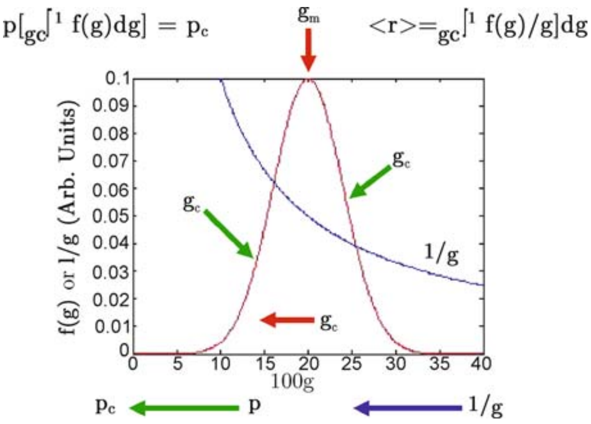
In order to derive a quantitative determination of the critical behavior of $\mathbf{R}_L$ let us try now to evaluate the behavior of $\langle \mathbf{r} \rangle$. Above, we have selected a subset of the conductors in the system (i. e. a subset of the resistors that participate in the conduction process). The average resistance of this subset is

$$\langle \mathbf{r_c} \rangle = \int_{\mathrm{gc}}^{g2} g^{-1} f(g)\mathrm{d}g \,. \tag{22}$$

Any other subset of p_c resistors will have an $\langle \mathbf{r} \rangle$ which is larger than $\langle \mathbf{r_c} \rangle$ (or equal to it), and thus $\langle \mathbf{r_c} \rangle$ yields the lowest estimate of $\langle \mathbf{r} \rangle$ that can be inserted in Eq. (19) instead of the $\mathbf{r}_0$ value in Eq. (19). Considering the fact that we are concerned with $p \to p_c$ and that $g_c \to 0$, the contribution of the very low g-values (with values of $g < g_c$) to the "measured" network resistance becomes negligible as $g_c \to 0$. This is in particular so if the values of g are changing by orders of magnitude as $g \to 0$ [87]. Hence, this $\langle \mathbf{r_c} \rangle$ of the "bypassing" network becomes a "more and more" accurate estimate for the value of $\langle \mathbf{r} \rangle$ in the $g_c \to 0$ limit, and consequently a reasonable replacement for $\mathbf{r}_0$ in Eq. (19).

Let us try and get a physical feeling for the correspondingly expected critical behavior, i. e., for the dependence of $\mathbf{R}_L$ on $(p - p_c)$. For that purpose we follow the dependence of $\langle \mathbf{r_c} \rangle$ on $(p - p_c)$ by considering the behavior of $1/g_c$ when $p \to p_c$. The above mentioned "list" of conductors, as p is made to decrease (i. e., some of these occupied bonds are eliminated as p_c is approached), we have to go "down the list" of the conductors in the system, in order to choose p_c bonds from the new p, so that the smallest g_c involved is smaller (or at most equal) in comparison with the g_c value of the larger (i. e. "former") p. Of course, it is not enough that g_c will just become smaller as $p \to p_c$, since if the availability of these low-value g_c conductors becomes scarce as $p \to p_c$, they will not contribute significantly to the network (or to the average $\langle \mathbf{r_c} \rangle$) and there will be an effective constant $\langle \mathbf{r_c} \rangle$ value (which is independent of $p - p_c$) that will determine the resistance $\langle \mathbf{r} \rangle$ and thus the global resistance of the system (see examples below, and in the following sections). If this is not the case, the effect of $\langle \mathbf{r} \rangle$ on $\mathbf{R}_L$ is clear then; not only changes of the network connectivity (Eq. (19)) take place as $p \rightarrow p_c$, but also, the average resistor value that contributes to the conduction can diverge, yielding an additional $p - p_c$ divergence to the "universal" contribution of Eq. (19). Hence, a compounded, non-universal (or non lattice-like) behavior can be obtained.

For a more quantitative illustration of the above discussion let us consider the simple case of $f(g) = 1$ in the interval $0 \leftarrow g \leq 1$. In this case we have from (Eq. (22)) that $\langle \mathbf{r_c} \rangle \propto \log(1/g_c)$ and from (Eq. (21)) that $g_c \propto p - p_c$. Hence, substituting $\langle \mathbf{r_c} \rangle$ instead of $\mathbf{r}_0$ in Eq. (19) will provide a logarithmic divergence to $\mathbf{R}_L$, as $p \rightarrow p_c$, in addition to the power law divergence of the global or the "lattice-like" contribution. It is also apparent then that it is enough that when $f(g)$ increases towards $g = 0$ (i. e., $f(g)$ diverges as $g \rightarrow 0$), a stronger (say, power law) divergence of $\langle \mathbf{r_c} \rangle$ with $p - p_c$ will be obtained. On the other hand, there will be no divergence of $\langle \mathbf{r_c} \rangle$ if $f(g)$ decreases (in this case $f(g)$ diminishes) towards $g = 0$. In the latter case we have that the approach of p to p_c is associated with the "disappearance" of the large $(1/g)$ resistors, while in the former, this approach is associated with the "appreciable" availability of these resistors ("at the expense" of the smaller $(1/g)$ resistors). From the above analysis we see that the crossover between the diverging and non diverging $\langle \mathbf{r_c} \rangle$, as $p \rightarrow p_c$, occurs when $f(g)$ "overcomes" the increase in the "summation" of the resistors, as reflected by the $(\mathrm{d}g/g)$ term in Eq. (22). This is easy to comprehend in the LNB picture as a competition between the possible introduction of large $(1/g)$ resistors in the links (the $f(g)$ effect), and the rate of increase in the "weighted" values of the "needed" individual resistors in the link (the $\mathrm{d}g/g$ effect) as the links become longer (i. e. as $p \rightarrow p_c$). This scenario of the "need to supply enough high-resistance resistors" in order to get the divergence of $\langle \mathbf{r_c} \rangle$ as $p \rightarrow p_c$, will be important as we turn to consider "real" continuum systems. Physically, we found here an important principle of the $\langle \mathbf{r_c} \rangle$ behavior, the appearance of non universal behavior will be determined by the competition between the $f(g)$ and the $\mathrm{d}g/g$ effects. One notes is passing that for other dynamical properties (that are described by functions of g other that $1/g$) the "$\mathrm{d}g/g$" effect will be changed accordingly [42].



Continuum Percolation, Figure 6
An illustration of a peaked distribution of the local conductances in a continuum system, as well as the variation of the participating g-values as $g \rightarrow 0$. The relation between the approach of g_c to the $g \rightarrow 0$ limit and the corresponding direction of the $p - p_c$ variations is indicated by the arrows in the figure. The principal equations that determine g_c and the approximated average resistor in the sample $\langle r \rangle$ are also given

In "real" systems, however, the $f(g)$ distributions may be more complicated than described above, yielding behaviors that were not accounted for in the literature until very recently [70,71]. To illustrate such a distribution and to summarize the above discussion, let us consider an $f(g)$ function that is peaked at some value g_m and diminishes towards both, $g = 1$ and $g = 0$. Such a distribution from $g_2 = 0.4$ to $g_1 = 0$ is illustrated in Fig. 6. As p is decreased towards p_c, the g_c values "move first" along the $f(g)$ part where $f(g)$ increases faster than $1/g$. Over this range, i. e. $0.4 \leq g_c \leq g_m$, we have (see above) the conditions that yield a diverging-like behavior of $\langle \mathbf{r_c} \rangle$. Hence, as the decrease of p is associated with the decrease of g_c, we will get an apparent non universal behavior. Continuing with the decrease of p to the p-range for which $f(g)$ decreases with $1/g$ (i. e. the $0 \leftarrow g \leq g_m$ range), the apparent behavior of $\langle \mathbf{r_c} \rangle$ will become of the type that does not have a divergence and will thus yield a universal-like behavior of the system in the corresponding p-range. We note that while, strictly speaking, the critical behavior (i. e., at the $p \rightarrow p_c$ limit) is universal for this $f(g)$, in "practice", i. e., when the p values are somewhat removed from p_c, the behavior observed (i. e. for $g > g_m$) may be characterized as non universal. In fact, considering the increase in the "pace" in which $f(g)$ varies with $1/g$ in the $g > g_m$ regime, one would expect for the large g (or large p) values, a behavior that is close to the universal behavior, and then that this behavior will turn to a more pronounced non universal-like behavior as g and p are decreasing (but still for $g > g_m$). This pronounced behavior will be weakened however, as g_m will be approached, until upon the decrease of $f(g)$ for $g < g_m$ a universal behavior is obtained. Hence, a peak in the "local" t value is expected for some p (i. e. at the inflection point of $f(g)$ in the $g_m < g < 0.4$ interval) and a transition from a non universal-like to a universal behavior will be obtained at some lower p. Examining the various experimental data in the literature [128] that suggested a non universal behavior for cases where $f(g)$ is expected to be of the form shown in Fig. 6, one must conclude that the p value that corresponds to the g_{m} peak lies below the p values-limit available in the acquisition of the experimental data. In fact, the range of the g values between g_2 and g_{m} usually consists of a few orders of magnitude in the variation of g [70] and thus the experimental data that is fitted to a range of a few orders of magnitude in $\mathbf{R}_L$, while appearing to represent a single "non universal" t value, does not disclose the variation in the value of t as $p \to p_{\mathrm{c}}$, due to the relatively limited range of the latter (see also Sect. "The Average Resistance in the Case of Tunneling-Percolation" below).

The Model of Kogut and Straley and Beyond

As we have shown above, the crucial factor in the determination of the critical behavior of an average dynamical property is the variation of its extreme value as $p - p_{\mathrm{c}} \to 0$. K&S have proposed a particular distribution of the local conductances in a system which can exhibit the three basic behaviors that we mentioned above. This distribution (which as we show below turned out to be very useful for the description of best known real systems) yielded then analytic expressions for the diverging or non diverging $\langle \mathbf{r}_{\mathrm{c}} \rangle$ values as $p \to p_{\mathrm{c}}$. The distribution that they suggested was

$$f(g) = 1/(1-\alpha) g^{-\alpha} \,, \tag{23}$$

which is normalizable for $\alpha < 1$, in the range $1 > g > 0$. For this distribution one gets from Eq. (21) that

$$g_{\mathrm{c}} = \left[(p - p_{\mathrm{c}})/p\right]^{1/(1-\alpha)} \tag{24}$$

and from Eq. (22), that

$$\langle \mathbf{r}_{\mathrm{c}} \rangle = \left[(1-\alpha)/\alpha\right]\left[g_{\mathrm{c}}^{-\alpha} - 1\right] . \tag{25}$$

We see here the three behaviors described above as follows. If $\alpha = 0$ we have that $f(g) = 1$ as discussed above. If $\alpha < 0$, $\langle \mathbf{r}_{\mathrm{c}} \rangle$ is not diverging as $p \to p_{\mathrm{c}}$ and it is the constant $(1-\alpha)/(-\alpha)$, i. e., the system behaves in the $g \to 0$ limit as if it is made of local resistors all having the latter value. If $\alpha > 0$, we get that in the interesting regime of $p \to p_{\mathrm{c}}$ (i. e., where $g_{\mathrm{c}} \to 0$),

$$\langle \mathbf{r}_{\mathrm{c}} \rangle \approx \left[(1-\alpha)/\alpha\right] g_{\mathrm{c}}^{-\alpha} \propto \left[(p - p_{\mathrm{c}})\right]^{-\alpha/(1-\alpha)} . \tag{26}$$

We recall that the resistance of the sample is approximated by $\mathbf{R}_L \propto \langle \mathbf{r}_{\mathrm{c}} \rangle (p - p_{\mathrm{c}}) -^{t_{\mathrm{un}}}$, where t_{un} describes the effect of the connectivity of the global network of the system (Eqs. (3)–(5)). Correspondingly, the critical behavior of the electrical conductivity of the system, which determines the critical exponent t (see Eq. (4)), will be given now by

$$t = t_{\mathrm{un}} + \alpha/(1-\alpha) \,. \tag{27}$$

In what follows we will show that the mapping of the most conspicuous systems in the continuum onto the K&S distribution is possible and that it is associated with the corresponding local geometrical parameters. We will also note the very important case where this mapping is not simple and other approaches must be taken in order to determine $\langle \mathbf{r} \rangle$ and thus the $\mathbf{R}_L$ dependence on $p - p_{\mathrm{c}}$, as the percolation threshold is approached. For example, the application of the Effective Medium Approximation (EMA) by Grimaldi and Balberg in 2006 [70] yielded the confirmation of the rather more complicated behavior that was discussed above in relation to Fig. 6, on the one hand, and which explained the moderate $t - t_{\mathrm{un}}$ (that are in the 1–10 range) values that were obtained in numerous composites [128], on the other hand.

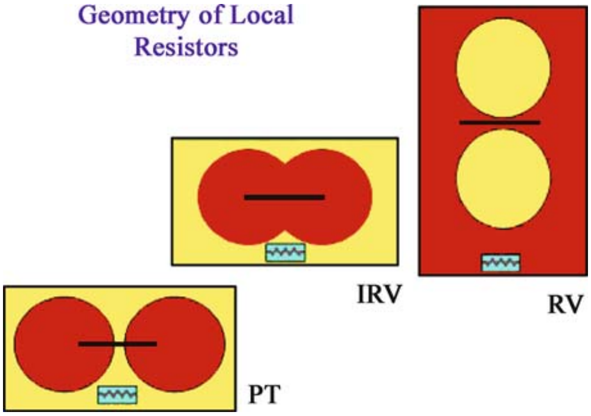
The Local Resistors and Their Distributions

As pointed out above, the conductors' distribution, as given by K&S was considered to be quite abstract until this resistors value distribution was suggested for some real systems by Halperin, Feng and Sen in 1985 [72]. Once this distribution is given, its mapping onto the K&S distribution is, as shown below [12], quite straight forward. Another way to derive the corresponding critical behavior was based on the use of a physically more transparent and more rigorous approach [59]. The corresponding findings were the very crucial steps that have finally explained the universal and non universal behaviors of "real" systems. In passing, we note that a posteriori, the critical behavior estimated by the above consideration of $\langle \mathbf{r}_{\mathrm{c}} \rangle$, is also confirmed by the latter approach, as will be shown below.

Since both approaches have been given in detail in the literature, we will only outline here the principal steps in their utilization for the determination of the corresponding t values. Starting then with the first approach, let us assume that the resistance of a local resistor in the system, $1/g$, is determined by a single local geometrical parameter, ε, and that the distribution of the ε values in the system, $h(\varepsilon)$, is known. The distribution function of the local conductor values, $f(g)$ is simply given then by [10]

$$f(g) = h(\varepsilon)(\mathrm{d}\varepsilon/\mathrm{d}g) \,. \tag{28}$$

"All one has to find" then, from the local configuration that occur in the system, are the $h(\varepsilon)$ function (which is usually not easy to do; see below) and the function $\varepsilon(g)$ (which is usually easy to do).



Continuum Percolation, Figure 7
The three major configurations of "real" systems in the continuum. The red color represents the conducting phase, and the yellow color represents the insulating phase in the systems. The bars and the resistors represent the "necks" and the corresponding dominant local resistors in the system. The configurations represent, from right to left, the random void, the inverted random void and the percolation tunneling models

There are three principal local configurations that were studied in detail. Let us discuss these configurations and consider the single geometrical parameter, ε, that describes them. These three configurations are illustrated in Fig. 7, where the "conducting" phase is colored in red and the "insulating" phase is colored in yellow. Starting from the random void (RV, sedimentary rock-like system, or a metallic bulk in which holes were punched out), that is also known as the Swiss cheese-like system, we can assume that the distance between the surfaces of the spherical pores is ε, and there is a "neck" of volume $(2\sqrt{b\varepsilon})C_D\varepsilon^{D-1}$ between them. Here, b is the radius of the pore and C_D is a constant (which is 1 in $D = 2$ and $\pi/4$ in $D = 3$). This neck determines the resistance of the neck that is associated with the two adjacent spheres [72] since beyond this "neck" the local resistance in the system is relatively small. The value of the local resistor in the system is well approximated then by the resistance of the "neck" which is given by $\rho_0(2\sqrt{b\varepsilon})/C_D\varepsilon^{D-1}$, where ρ_0 is the resistivity of the conducting medium. Hence, the resistance of the local resistor is varied with ε as $1/g \propto \varepsilon^{3/2-D}$ (i. e., the local conductance can be expressed by $g \propto \varepsilon^{D-3/2}$).

In the second configuration, known as the inverted random void (IRV) system, such as a granular metal above the onset of metallic conductivity [61], or a system of pores filled with a conducting liquid [32,59] the overlap between two spheres can be characterized by the parameter $\varepsilon = 2b - r$, where r is the distance between the centers of the adjacent spheres. Correspondingly, the neck for conduction here is approximated to be $2\sqrt{\varepsilon b}$ long, and it has a cross section of $C_D(b\varepsilon)^{(D-1)/2}$ ($C_D = 2$ in $D = 2$ and $C_D = \pi$ in $D = 3$), yielding that $g \propto \varepsilon^{D/2-1}$. For the above two cases we have then that $g \propto \varepsilon^m$, where $m = D - 3/2$ in the RV model and $m = D/2 - 1$ in the IRV model.

The third case is that of conduction between two spheres that do not overlap for which the distance between their centers is r and the distance between their surfaces is $r - 2b$. In that case, the conduction between the two spheres is by tunneling and the charge transfer probability between the surfaces is expected to decrease as $\exp[-(r - 2b)/2d]$, where $2d$ is the effective (say, the WKB) tunneling decay coefficient. This percolation-tunneling (PT) case will be considered in Sects. "The Tunneling Percolation Problem", "The Average Resistance in the Case of Tunneling-Percolation".

Let us turn now to the corresponding $h(\varepsilon)$ and $f(g)$ distributions in the first two cases. Noting that the derivation, theoretically or experimentally, of these distributions is the most difficult part (see above and below) in our problem, one can consider the first suggestion of a "real" $h(\varepsilon)$ dependence, and its use in the present context, to be the breakthrough that lead to the understanding of the non universal behavior of the dynamic properties in "real" systems. This was achieved, as pointed out already, in the 1985 [72] work of Halperin Feng and Sen, who showed that in porous media $h(\varepsilon)$ is a constant, h_0, for $\varepsilon \to 0$. The corresponding distribution is, however, normalizable since $h(\varepsilon)$ tapers off with increasing ε, above a certain value of ε (as is obvious from the finite size of the system, see below). Being interested in the smallest ε values, i. e., the largest resistors in the system, one can easily apply the "recipe" given by Eq. (28). Using then the general $g \propto \varepsilon^m$ relation, where m is the relevant exponent, we have that $f(g) \propto dg/d\varepsilon \propto h_0 g^{(1-m)/m}$ and thus, that the α value of K&S is $1 - 1/m$. In the above RV case, for which $m = D - 3/2$, we have then that $t - t_{\text{un}} = \alpha/(1 - \alpha) = m - 1 = D - 5/2$. Hence, from the consideration of K&S, it is easy to conclude that for $D = 2$ a universal behavior will be found, while for $D = 3$, a value of $t - t_{\text{un}} = 1/2$ is predicted, as confirmed indeed by the corresponding simulations of Sen, Roberts and Halperin in 1985 [111].

The above procedure was generalized by Balberg in 1998 [12] for any power law distribution of the $h(\varepsilon) \propto \varepsilon^{-\omega}$ type which, following the above, yields that $f(g) \propto g^{(1-m-\omega)/m}$ and correspondingly (see Eq. (27)) that $t - t_{\text{un}} = (m + \omega - 1)/(1 - \omega)$. Of course we note that we must have that $\omega < 1$, so that the distribution is normalizable. For the three dimensional RV case we will generally get then that $t - t_{\text{un}} = (D - 5/2 + \omega)/(1 - \omega)$. It is important to note that if ω varies from one ε range to another, the behavior obtained will be more complicated. In particular, transitions between universal-like and non universal-like behaviors may be obtained as $p \to p_c$. Such a variation, with the consequences of a none monotonic $h(\varepsilon)$, and thus a non monotonic $f(g)$, has not been considered in the literature for this case. However, following the above discussion of Fig. 6, one can account for now such transitions. In fact, a similar situation in principle, was evaluated very recently for the percolation tunneling problem, that will be described in Sect. "The Average Resistance in the Case of Tunneling-Percolation". A mapping similar to that of the RV model was obtained for the IRV model [12,23] yielding that $\alpha \leq 0$ for $D \leq 4$, and thus the divergence of the conductance is obtained (for the $\omega = 0$ case) only for $D \geq 5$.

Let us derive now these results in a physically more transparent manner [59,72]. We start by assuming the LNB model and try to find the relation between the parameter ε, that is relevant to the value of the resistors, and the proximity parameter $p - p_c$, so that a relation such as that of the K&S model can be found between the average local resistance $\langle \mathbf{r_c} \rangle$ and $p - p_c$. For the $\varepsilon \to 0$ limit, which corresponds to the larger $1/g$ resistors (that must be included in the network when $p \to p_c$), we assume, as above, that $h(\varepsilon) = h_0$ and that $g \propto \varepsilon^{-m}$. The corresponding $\langle \mathbf{r_c} \rangle$ is determined then by

$$\langle \mathbf{r_c} \rangle \propto \int_{\delta}^{\Delta} h_0 \varepsilon^{-m} \, \mathrm{d}\varepsilon \,, \tag{29}$$

where Δ is the ε value above which $h(\varepsilon)$ tapers off with increasing ε, and δ is the smallest ε value (which is associated with the smallest g value), in a typical link, that is made of L_1 singly connected bonds. The probability that ε will be larger than a given δ for a bond in this link is simply $1 - \int_0^\delta h_0 \, \mathrm{d}\varepsilon$, and thus for all L_1 bonds in the link it is $(1 - \int_0^\delta h_0 \, \mathrm{d}\varepsilon)^{L1} = \exp(-\delta L_1 h_0)$. This means that at $\delta \approx 1/L_1 h_0$ this probability starts to become significantly smaller than unity, and thus this relation provides a good estimate of δ, i. e. that $\delta \propto 1/L_1$. Having this value and considering that $\delta \ll \Delta$ and that $m > 1$ (see above), we get from Eq. (29) that

$$\langle \mathbf{r_c} \rangle \propto h_0 \delta^{-m+1} \approx h_0^m L_1^{m-1} \propto (p_- p_c)^{1-m} \,. \tag{30}$$

This result yields, as above, (see Eqs. (26) and (27)) that $t = t_{\mathrm{un}} + (m-1)$. As we saw for the RV model, $m = D - 3/2$, and thus for the corresponding three dimensional system we get $t = t_{\mathrm{un}} + 1/2$, in accordance with the above result that was obtained from the K&S model. Similarly, for the IRV model we got that $m = D/2 - 1$, and thus for $D = 3$, we have that $m < 1$ and $\alpha < 0$ yielding that $t = t_{\mathrm{un}}$. We can further attach the distribution term $\varepsilon^{-\omega}$ as above and get that instead of the relation of $\delta L_1 h_0 \approx 1$ we will have that $\delta^{(1-\omega)} L_1 h_0/(1-\omega) \approx 1$ and thus, that in the more general case, $\langle \mathbf{r_c} \rangle \propto h_0 \delta^{-m+1-\omega} \approx h_0^m L_1^{(\omega+m-1)/(1-\omega)} \propto (p_- p_c)^{-(\omega+m-1)/(1-\omega)}$ or $t = t_{\mathrm{un}} + (m+\omega-1)/(1-\omega)$, where ω is as above [12].

The Tunneling Percolation Problem

Before turning to the problem of the average value of the local resistor in the case of tunneling percolation, let us examine the unique relation between tunneling and percolation [13]. Tunneling is essentially the principal mechanism of electrical conduction in most classes of composite materials. This is to be distinguished from the cases that involve a coalescence of particles encounters one that in granular metals [2], porous media (the IRV model; see above) and microemulsions [69], where the onset of a continuous geometrical network is associated with the onset of a global "metallic" conductivity. In most other cases, such as in granular metals in the dielectric regime [1] and in various semiconductor composites [26,27], the conductivity between the particles (or objects) is generally controlled by tunneling.

As one considers the tunneling in the percolation-systems context, one notices that it challenges our above understanding of continuum percolation as follows. In lattices (S&Z-like), or in the above RV and IRV systems, the geometrical connectivity and the electrical connectivity are very clear and both are associated with the geometrical continuity of the conducting phase. This is also the case when we can define a soft shell that wraps the hard-core so that the overlaps of the soft shells provides a continuous conducting phase, as can be envisioned to be the case in microemulsions [35]. When the conductivity is provided only by tunneling, there is no geometrical connectivity (or continuity) but there is electrical connectivity. The latter connectivity, however, is problematic, since all the conducting objects are connected electrically to each other (albeit with a different "strength" that is determined by the exponential decay of the tunneling probability). Hence, the system has a "zero span" of the geometrical connectivity and an "infinite span" of the electrical connectivity. It appears then to be quite surprising [13] that systems corresponding to this "counter" percolation configuration, such as many types of composites, exhibit computational [24,78] or experimental [128] well-defined percolation thresholds with a universal or a non universal critical behavior. This problem was essentially clarified only recently as to be described below. We note then in passing that until quite recently the very many experimental or computational studies on numerous composite systems that were discussed in terms of a bona-fide percolation critical behavior have done so with no justification.

The first explanation for the above paradox was given only very recently by Toker et al. in 2003 [124]. They concluded, by examining the fractal dimension of the "electrical" percolation network (that was derived from local probe microscopy on real composites), that if the tunneling decay parameter is very small compared with the size of the objects (say, spheres), in practice, only the network of adjacent neighbors (see above) determines the global measurable conductivity of the system. In that case the percolation threshold is simply associated with by the concentration of "near neighbor" objects, such as in the case of a system of objects of a "large" hard-core and a thin soft shell. In this case the percolation tunneling problem can be mapped onto such a model. The dynamical properties are determined then by the local resistors distribution (see also Sect. "The Average Resistance in the Case of Tunneling-Percolation") as in the above given discussion.

When this is not the case, i. e. when the tunneling distance is of the order of the size of the conducting objects, such as the case in granular metals [1,26,124] or quantum dot semiconductors [27] in the dielectric regime, the observation of a percolation threshold and a critical behavior is less obvious. It turns out that the solution to this problem can be based on the same model as in the above case if we take into account the accumulated information that we have on the radial distribution (RDF), $G(r)$, of a collection of hard spheres. This function that is defined [138] as the probability $G(r)\mathrm{d}r$ to find a particle center within the interval r and $r + \mathrm{d}r$ from a given particle, shows that there is an evolution of the peak of "nearly touching" particles as we go from the dilute, or small b, limit to the (hard core) dense spheres limit, i. e. as the particles size and/or density increase. In particular, the common conclusion of all the corresponding models is that the larger the particles phase content (i. e. the particles density and/or their size) the narrower the "tail" of the near neighbors separations. This is equivalent in the present context to the statement that more particles will overlap their neighbors within the narrow shell, d, that is of the order of the tunneling distance. In other words, the way the hard spheres arrange themselves yields a preferred conduction network that can be looked upon as made of an IRV system with the soft shell, the thickness of which is of the order of the tunneling decay constant. Hence, the model is much like the one proposed above for the $b >> d$ case. It is of course appreciated that as the system becomes very dilute it also becomes reminiscent of the near neighbor hopping model [114]. Indeed, a close examination of the dependence of the conductivity on the conducting phase content in granular metals [26] reveals a density regime in which both, percolation and hopping, can simultaneously account for the experimental data, while only a percolation model can account for the data obtained on the "large spheres" system of Carbon black-Polymer composites [102].

The Average Resistance in the Case of Tunneling-Percolation

The determination of the average local resistance in the tunneling-percolation problem was solved thus far accurately only for the 1D (or quasi 1D; see below) systems [10,11]. At higher dimensions, the lack of analytic expressions for the $f(g)$ distribution does not enable a simple evaluation of the corresponding critical behavior. However, with the information already available [5,126] it is possible, as described below, to obtain a good physical understanding and derive semi-quantitative results for the critical behavior [78].

We start then with the simple case of a linear chain of geometrically, but not electrically, isolated "metallic" spheres [71], the hard core radius of which is b and the average distance between which is $2a$ ($= 1/N_1$, where N_1 is the average number of spheres per unit length). The random distribution of the distances of the centers of the nearest neighbor spheres from a given sphere center in the corresponding 1D system is the well-known 1D Hertz distribution [10,11,126] that can be written as [26]

$$h_1(r) = \big[1/(2a - 2b)\big]\exp\big[-(r - 2b)/(2a - 2b)\big] , \quad (31)$$

where r is the distance from the center of the reference sphere. On the other hand, the local tunneling conductance between two such spheres is simply given by

$$g = g_0 \exp\big[-(r - 2b)/2d\big] , \quad (32)$$

where g_0 is a corresponding geometrical-physical constant and $2d$ is the tunneling decay constant. For the calculation of the average value of the local resistance we will neglect the resistance between "non-adjacent" spheres, i. e., we assume that $b \gg d$. Applying Eqs. (28), (31) and (32), we get that for the nearest neighbor connections (or bonds; [10,11,26])

$$f_1(g) \propto \exp\big[(r - 2b)/2d\big]\big[1 - d/(a - b)\big] \propto g^{-\alpha} , \quad (33)$$

where,

$$\alpha = 1 - d/(a - b) . \quad (34)$$

Following the above K&S prediction (Eq. (27)) for the exponent t we get then that

$$t - t_{\text{un}} = (a - b)/d - 1 \, . \tag{35}$$

Considering the discussion in Sect. "The Basic Physics of the Non-universal Behavior of the Conductivity", regarding Eq. (22), we see that if d is larger than $(a - b)$, there is not "enough" supply of distant neighbors, or large resistors ("the $f(g)$ effect"), to overcome the needed $1/g$ resistors ("the $\mathrm{d}g/g$ effect") and thus to yield the diverging $\langle \mathbf{r_c} \rangle$. In contrast, if $d < a - b$, even though the "supply" (Eq. (31)) decreases with increasing r (i. e. with decreasing g values), this "supply" is enough for the "fastly" increasing $1/g$ values, to yield an increase in $f(g)$ as $g \to 0$. This is an important observation that results from the fact that we have two terms in Eq. (28) so that the behavior of $f(g)$ is not necessarily the behavior expected by just examining $h_1(r)$. Physically then, as above, in the $d < a - b$ case the decrease in the available small g values (or the large r values) is slower than the increase in the value of the large resistors, $1/g$, as $r \to \infty$ or (Eq. (32)). Correspondingly, one obtains the divergence of $f(g)$ (Eq. (33)) as $g \to 0$ i. e., when p approaches p_c. This description is manifested quantitatively of course by the above mapping of the tunneling resistor problem onto the K&S model, with an α value that determines the transition between a non-diverging to a diverging distribution, i. e. $\alpha = 0$, which in the present case corresponds to $d/(a - b) = 1$ (Eq. (34)).

For completeness, let us present the physically more transparent procedure for the determination of $t - t_{\text{un}}$, such as the one of Halperin, Feng and Sen in 1985 [72], which we outlined in Sect. "The Local Resistors and Their Distributions", but this time for the tunneling-percolation problem. This approach [102] is similar then to the one considered for the RV and IRV models, but for the present problem we assume that the largest distance between two adjacent spheres in the link is s (rather than the smallest distance, δ, in the RV and IRV problems). We are above interested then in the probability that the largest r in the link of L_1 singly connected resistors is smaller than some prechosen s. This probability (following Eq. (31) and the procedure given in Sect. "The Local Resistors and Their Distributions") is simply $\{1 - \exp[-(s - 2b)/(2a - 2b)]\}^{L1} = \exp\{-L_1 \exp[(s - 2b)/(2a - 2b)]\}$. Correspondingly, this probability starts to be small at a value of about $\exp(-1)$, and thus the condition for s being the largest "bond-length" in the link is that $L_1 \approx 1/\exp[(s - 2b)/(2a - 2b)]$. On the other hand, the average resistance in the link, $\langle \mathbf{r_c} \rangle$, is now proportional to $\int_{2b}^{s} \exp\{[(r - 2b)/(2a - 2b)] - [(r - 2b)/2d]\}\mathrm{d}r$. The dominant term in the integral is then $\exp\{[(s - 2b)/(2a - 2b)][1 - (a - b)/d]\}$, or in view of the above, the dominant term of $\langle \mathbf{r_c} \rangle$ is $L_1^{-[1-(a-b)/d]} \propto (p - p_c)^{1-[(a-b)/d]}$. Correspondingly, we get (as we already found above from our mapping of the problem onto the K&S distribution) that in this relatively simple 1D case, $t - t_{\text{un}} = (a - b)/d - 1$.

This finding has been adopted not only for the relativle simple 3D cases where the above $h_1(r)$ distribution is still appropriate [71] but also, very recently, to general 3D random systems. The simpler cases include cellular composites in which the system can be assumed to be made of insulating (say, cubic) blocks, and of conducting channels with a diameter just wider than $2b$ which include metallic spheres of diameter $2b$. The channels are further assumed to coincide with the edges of the insulating cubes [71,78]. In practice, however, the comparison of the prediction of Eq. (35) with experimental observations on "real" cellular composites [128] shows that the above predictions yield t values that are very exaggerated. If one tries to provide then a "more realistic" model of such a composite, one must also assume that the chains are not infinite and that thier number per unit length, N_1, (and thus a) may vary from one chain to the other, according to the make of the composite. This is even to the degrees that $(a - b)/d$ may be larger or smaller than unity in different chains. The corresponding distribution of N_1 in the chains may affect then the critical behavior by yielding an effective "global" α. For this case that was considered in detail by Grimaldi et al. in 2003 [71], it is obvious that the chains' $N_1(= 1/2a)$ values will contribute to a universal behavior (high N_1) or to a non universal (low N_1) behavior of the system. The global behavior will be determined then finally by the distribution of the various N_i's values in the system.

A priori one may think that this model of Grimaldi et al. should apply to systems of higher dimensions as one approaches the percolation threshold, since in the Links, Nodes, Blobs (LNB) model we consider a conducting network of such large chains that are, at first sight, one dimensional in nature. At second sight one notices two problems with that assumption. First, the distribution function of the nearest neighbor sites in the $D > 1$ case, $h_D(r)$, is different from the one given in Eq. (31) and, second, and more severe, the LNB link is made of singly connected bonds that are determined by the connectivity of the $D > 1$ network and not (necessarily) by the first nearest neighbors, as in the simple 1D case.

Before giving the recent "exact" solution to the problem [78] let us see the basic underlying physics of it. The global conductance as $p \to p_c$ can be concluded semi-quantitatively by noting that for the very large $(r - 2b)$ values (that are relevant to the critical behavior), the $h_3(r - 2b)$ distribution function is expected to decrease as $\exp[-(r/2a)^3]$ for all neighbors [5]. Here, a is the average (see below) distance between neighboring sphere centers in the above 3D system. Hence, "finally" (i. e., for the largest r's or smallest g's) the "supply" of the g's (i. e. the function $h_3(r-2b)$) will not meet the "demand" of Eq. (32). This situation is different then from the situation described above (Eq. (33)) for the 1D case of $d < a-b$, for any value of $d/(a-b)$. Correspondingly, as demonstrated below, for these large r's, $f(g)$ will be a decreasing function of $1/g$. Strictly speaking, for the general 3D system the "ultimate" (i. e. at the $p \to p_c$ limit) the critical behavior will be universal as discussed in Sect. "The Basic Physics of the Non-universal Behavior of the Conductivity". However, the latter limit is rather experimentally [70] and so far computationallay [78] a non accessible (see below) and thus there is a great interest in the critical behavior of the system on route to this limit. To consider the latter behavior let us use the 1990 [126] result of Torquato, Lu and Robinstein for the 3D distribution function of the inter-particle distances of the first nearest neighbors, $h_3(r)$. The latter result can guide us in predicting the behavior of the more general $D > 1$ systems.

The relevant dominant term that as implied by their $h_3(r)$ is

$$H_3(r-2b) \propto A(\rho)(r-2b)^2 \exp\left\{-\left[8A(\rho)/b^3\right]\left[r^3-(2b)^3\right]\right\} , \quad (36)$$

where, $A(\rho) = [\rho v(1+\rho v)/(1-\rho v)^3]$, v is the hard core spheres volume $(4\pi/3)b^3$ in a system of a unit volume and ρ is their concentration (i. e., per unit volume of the sample). This concentration is related to the above $2a(> 2b)$ parameter (i. e. the average distance between the sphere-centers of two nearest neighbors) that, we simply define it here by $\int_{2b}^{\infty} h_3(r) r \mathrm{d}r$ [70]. A very rough, but geometrically more transparent, estimate of a (which is enough for the present discussion of the basic physics of the problem) can be given by assuming that the spheres are implanted randomly (as in the $a \gg b$ case). In this case a is associated with the average "territory" of a sphere, and is thus given by the relation

$$(4\pi/3)a^3\rho = 1 . \quad (37)$$

Returning to Eq. (36), we can follow the behavior of $f(g)$ by applying Eqs. (28), (32) and (36) in a way that yielded Eq. (33). The first feature that we encounter in $H_3(r-2b)$, and which has no counter part in $h_1(r-2b)$, is a peak at some r_m that is of the order of $2(a-b)$. For $r < r_m$ both $H_3(r-2b)$ and the resistance term $1/g \propto \exp[(r-2b)/2d]$ increase with increasing r and thus $f(g)$ increases with $1/g$. The rate of this increase may be very large for the small $r-2b$ values due to the dominance of the $(r-2b)^2$ term in Eq. (36) that can be translated to a $g^{-\alpha}$ behavior with $\alpha > 1$ for a limited g range, in some cases [24]. Of course, this strong dependence will be weakened for larger r's (smaller g's) as follows. Since $H_3(r-2b)$ stops increasing at r_m it is obvious that the maximum rate of the increase of $f(g)$ with $1/g$ will be obtained at some intermediate value r_p that is smaller than r_m. This yields then a corresponding g_p value (Eq. 32) such that $g_p > g_m$. However, the increase of $f(g)$ with $1/g$ does not necessarily stop at g_m. To see that let us now consider the $r > r_m$ regime. The dominant term of $H_3(r-2b)$ there is the exponential decrease of this function and thus, using the procedure suggested by Eqs. (33) and (36), we can express $f(g)$ now as $f(g) \propto g^{-\alpha(r)}$ where now we have a "local α" that is given by:

$$\alpha(r) = 1 - \left\{2d\left[8A(\rho)/b^3\right]\left[r^2+br+(2b)^2\right]\right\} . \quad (38)$$

Noting that r is a measure of $1/g$ we have here then, unlike the constant α of the 1D case, a "locally" varying $\alpha(r)$. Hence, for the $r > r_m$ regime we will have that for small (in comparison with $2a-2b$), $r-2b$ values, $\alpha(r) > 0$, i. e. $f(g)$ will still be increasing with increasing $1/g$. For larger r's however, $\alpha(r)$ will "eventually" become negative and thus $f(g)$ will decrease with $1/g$. Denoting the r value, at which $\alpha(r) = 0$, by r_t, we can summarize the behavior of $f(g)$ as follows. For small $1/g$ values $f(g)$ increases with g, obtaining a maximum rate of this increase at some g_p which is larger than g_m. Then, at some g_t (such that $g_t < g_m$) $f(g)$ will decrease as a function of $1/g$. Following now the premise of our model (as described in Sect. "The Basic Physics of the Non-universal Behavior of the Conductivity") i. e. the increase of $1/g_c$ with the decrease of $p-p_c$) we can now predict the critical behavior of the global conductance of the system and in particular the corresponding critical exponent as p approaches p_c. For p far above p_c, $\langle \mathbf{r_c} \rangle$ will have a diverging-like behavior with $p-p_c$ and thus a non universal value will be obtained for t. The value of this non universal t will reach its maximum at a p value, p_p, and then, for still smaller p values, it will decrease until at a low enough p value, p_t, it will become the universal value t_{un}. Hence, as we saw above, strictly speaking, the critical (i. e. in the $p \to p_c$ limit) behavior of the general $D > 1$ systems will always be the universal behavior.

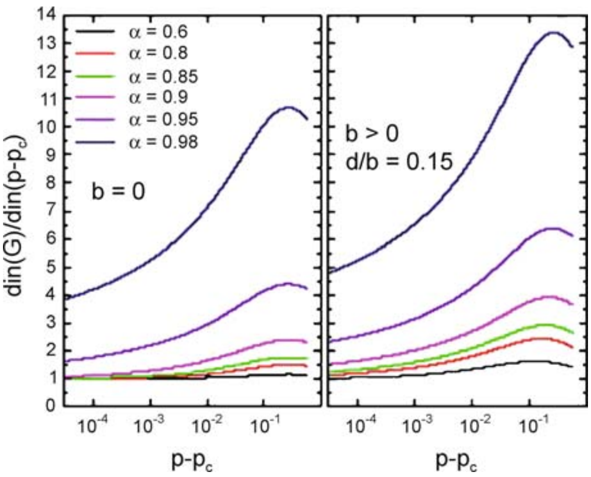
Since the value of $1/g_c$ and thus the values of $\langle \mathbf{r_c} \rangle$ and $\langle \mathbf{r} \rangle$ diverge as $p-p_c \to 0$, minute variations in the value of p cause very strong (orders of magnitude) variations in the former quantities. This implies that the variations that we described above for $1/g$, and thus for $f(g)$, actually take place over a narrow range of p when it is close to p_c. The first implication of that is that the use of a constant a (say, the one that applies to p_c) in the above equations (e. g. Eq. (37)) is a very good approximation. The second implication is that the "locally" (r range, g range or p range, see also Eq. (38)) determined value of t should be defined differentially as

$$t = \mathrm{d}[\ln G]/\,\mathrm{d}[\ln(p - p_c)] \,, \tag{39}$$

where G is the global conductance of the system. Using this definition and the g_c dependence on $p - p_c$, the above behavior of $f(g)$ can be translated to a p (or a $p - p_c$) dependence of t. We concluded above that far from p_c, t will increase with decreasing $p - p_c$ until it obtains its maximum, and from there on (i. e. for smaller p values) it will decrease until it reaches the corresponding universal value. The questions that arise are, of course, at what p, and for what value of $t - t_{\text{un}}$ will the peak t-value and the t-value transition happen, and how does this behavior depend on the parameters of the system [78]. The mathematical expression for the above described behaior has been developed very recently by Johner et al. [78] using the application of the Effective Medium Approximation (EMA). In that case, one essentially usses a lattice model, but with a distribution of bond or g-values taken by using the $f_3(g)$ function that is obtained by applying Eqs. (28), (32) and (36) to the basic EMA equation. They were able to show then that:

$$t - t_{\text{un}} = \left[(a-b)/d\right] \left\{\ln\left[yp/\left(p - p_c\right)\right]\right\}^{-2/3} \,, \tag{40}$$

where, $y = \exp[2b/(2a - 2b)]^3$. It is easy to appreciate that this expression follows the expected behavior as otlined above.



Continuum Percolation, Figure 8

The dependence of the (differential) conductivity exponent as calculated using the Effective Medium Approximation for a system of spheres, where 2*b* is their diameter and 2*d* is the tunneling decay constant. The results are presented for various values of the 1D parameter *α*, such that the higher the values of *α*, the smaller the local conductance values that are involved in the global conductance. (From [70])

Following our above discussion and the result given in Eq. (40), it is apparent that we cannot have a simple-complete mapping of the $D > 1$ systems (Eq. (38)) onto the K&S model as we had for the 1D model (Eq. (33)). On the other hand, the constant parameter α, of Eqs. (33) and (34), can be applied for the characterization of the above general $D > 1$ systems, since the three relevant length scales for all ($D \geq 1$) tunneling models are still the same i. e., a (see above), b and d. Using then $\alpha = 1 - d/(a - b)$ as the system characterization parameter, we show in Fig. 8 the above differential $t(p - p_c)$ dependence, as obtained by the application of the Effective Medium Approximation (EMA) to the 3D percolation tunneling system [70]. In this approximation, one notes that $t_{\text{un}} = 1$, but this (EMA value) is not important in the present context, since we are interested here in $t - t_{\text{un}}$. The results shown in Fig. 8, confirm then the above mentioned expectations and in particular that (as in the 1D case) the larger the value of α (in the $0 \leq \alpha \leq 1$ interval), i. e. the smaller the tunneling decay range in comparison with the inter-particle distances r (or, the smaller the corresponding g-values that are involved in the conduction), the more pronounced the non universal-like behavior (the larger the t values) that accompanies the conduction process. However, in contrast with the $D = 1$ case, the $t > t_{\text{un}}$ values are maintained only over a limited $p - p_c$ range, and "eventually" (as $p \to p_c$) a decrease of t towards the value of t_{un} takes place. In addition to this main result, i. e., the confirmation of the peak in the t dependence on p and its decrease to t_{un}, as p approaches p_c, we can derive further conclusions regarding the α dependence of the curves. First, for very small values of α, a non universal behavior will hardly be observed, and second, even though the shift of the $t(p)$ peak with increasing α is very small, it is obvious from our comparison of Eqs. (32) and (36) that this increase shifts the peak towards lower $p - p_c$ values. As to the meaning of this shift in practice, let us note that for a given type of a system (i. e., b and d fixed), the larger the value of a (the smaller the value of p), the more dilute the system. Correspondingly, as apparent from the smaller g values involved in the conduction process then, the decay of $H_3(r)$ in Eq. (36) will be "slower", and the non universal behavior will be also emphasized by the extension of the non universal behavior values into the lower $p - p_c$ regime. A similar argument can be derived from Eq. (38)

since the smaller the value of d/b the larger the value of $\alpha(r)$ i. e. the value of t for a given r. In passing, we also note that the parameters that determine the critical behavior can be controlled externally. For example, a voltage application effect on d, or a pressure application and temperature change on a, will change the value of α and thus the value of t. In that case, we expect to get a new class of phenomena that may be called "physically controlled percolation".

Considering the corresponding numerous data in the literature [128], one finds that there is no convincing direct experimental or Monte Carlo evidence, on a genuine continuum system, for the variation of $t(p)$ as predicted in Fig. 8, but there is some indirect evidence for that behavior (see below). The main reason for not observing the peaked $t(p)$ dependence in the experimental and computational works seems to be the accuracy of these data, or more importantly, the limited $p - p_c$ range that is actually accessible in the corresponding studies. For the experimental data, the latter effect can be appreciated as follows [70]. The proximity to the percolation threshold is usually given, as in Fig. 8, by the parameter $p - p_c$. However, this is not a good parameter for that purpose of the description of the proximity to the threshold in continuum systems. If both p and p_c are small, a small $p - p_c$ does not mean that the system is in close proximity to the critical regime. Rather, a quantity such as $(p - p_c)/p_c$ is a better indication of the proximity to the threshold. The problem is not too severe in lattices when one considers the resistors distribution imposed on lattices [88,111] since the p_c values are larger than 0.1 for two and three dimensions. This problem may be very severe, however, when one interprets experimental data in the continuum, where the proximity parameter is given in terms of the fractional occupied volume τ, and the corresponding critical fractional volume τ_c, i. e. by $(\tau - \tau_c)$. Here, while both, τ and τ_c can be very small, $(\tau - \tau_c)/\tau_c$ can be quite large, indicating that the system is far from the critical region. In this case, single, high, non universal-like, t values, are not too surprising in view of the above discussion, and the limited $(\tau - \tau_c)/\tau_c$ range hardly enables one to detect deviations from a single t value [70]. Computer simulations, on the other hand, "suffer" from finite size effects, and while a variable $t(p)$ behavior has been found [24,78], it has not yet been established to be beyond the influence of such effects.

However, more convincing, though indirect, confirmation of the t value variation with $p - p_c$ seems to be indicated by the experimental data on other electrical properties such as the resistance noise. For this property, the variation of the corresponding exponent, as a function of $\tau - \tau_c$, was found to indicate the decrease of t as τ approaches τ_c [40,42,85]. Hence, in view of the above, the transitions from a non universal behavior to a universal behavior seem to be well understood and well accounted for semiquantitatively, but further work is needed in order to examine this effect more quantitatively.

Future Directions

The relevance of the connectivity to the many areas of science mentioned in Sect. "Definition of the Subject", makes the achievements in the theory of continuum percolation of great importance to numerous natural and artificial systems. In particular, the understanding and conclusions derived from this theory in the last forty years serve as guides to the analysis and understanding of the various properties of these systems. Naturally then, future studies of continuum percolation are expected to proceed along two directions. The first is the further development of the theory itself and issues of principle and depth of understanding, such as the ones emphasized in the present review. The other direction is the application of the general conclusions and principles derived from the first direction for the understanding of the systems and properties such as those mentioned in Sect. "Definition of the Subject". The latter direction deserves quite a few reviews that will center on particular types of systems (say, solid composites and porous media) or particular properties (such as electrical noise and heterogeneous-fluid viscosity). Considering the wide scope of these many issues, we will mention here, as examples, only a couple of them that are relevant to systems and properties that became of wide interest only very recently.

On the very basic-scientific level, the establishment of continuum percolation as a phase transition that can be mapped onto the lattice percolation phase transition is desired, in order to determine rigorously and conclusively that the critical behavior of the geometrical-statistical properties is the same. Another fundamental theoretical problem is the derivation of the percolation thresholds and critical exponents rigorously. The first question that arises in that context is whether this can be done by another approach than the common one used thus far, i. e., the application of the theory of liquids, and whether such an approach can yield better results than that of the latter application. The second question is whether within the rather developed application of the theory of liquids, one can find a way to facilitate its implementation to the degree that will enable unbiased determination of percolation thresholds and critical exponents, as well as its application for "less trivial" objects than those considered thus far. On the more fundamental level of the latter problem, it is very interesting to know why this application is so successful for the description of the percolation behavior, even though the theory of liquids was not designed for the study of phase transitions.

Considering the empirical approach to the theory we note that the usefulness and the generalization of the concept of "pointedness" have not been studied extensively, in spite of its impressive predications concerning trends in the behavior of percolation thresholds in the continuum. The transparent physical-geometrical meaning of this concept calls then for its further utilization on the one hand, and for trying to understand the reasons for its success on the other hand. In particular we note that at present, the latter approach can be used much more readily and for many more systems than the above mentioned rigorous approach that is based on the theory of liquids. A specific problem that needs attention in order to derive a more complete understanding of the percolation threshold in continuum percolation is the value of B_c at the hardcore limit

Turning to the behavior of the dynamical properties, it appears that most of the basic principles of the corresponding theory are well understood by now. However, the problem of variable "bond strengths" is not well accounted for, in general, and beyond the "strong-interaction" limit (e. g., tunneling only to nearest neighbors) in particular. The meaning of the percolation threshold in corresponding systems and the predictions of the corresponding critical behaviors, are still open. In particular, the evaluation of the "residual" conductivity in comparison with the dominant percolation-like subnetwork has not been given thus far. For this, there is the need for the other corresponding parameters that characterize the charge, the mass, or energy, transfer between the objects in the system. It further appears however, that the most important information that one needs, when these principles are applied for the understanding of particular continuum systems, is the distribution function of the sizes and shapes of the objects in the system, as well as their separations. As emphasized in this review, this information is usually scarce, thus limiting the possible applications of the general understanding and principles outlined in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum" to specific systems. Here, the utilization of very modern experimental characterization techniques, such as local probe microscopies, should provide the link needed for the above application. On the theoretical-computational end, the distribution functions associated with the structure, and thus with the dynamical physical properties, need to be derived, in order to appreciate trends in the critical behavior (i. e., the values of the critical exponents) in various systems. These trends will become amenable for examination, however, only with the ability to produce or find systems where the percolation threshold can be approached much more closely than at present, in the experiments, and to create much larger systems (in order to avoid finite size effects) in the computations. Such developments are expected to yield the (generally missing so far) information on the dependence of the critical exponents on the proximity to the percolation threshold. It appears then that the meaning of the percolation threshold, when farther than closest neighbors are involved (such as in the case of tunneling-percolation systems where the "interaction" range is not much smaller than the size of the objects), as well as the critical behavior of such systems, beyond the intuitive picture suggested in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum", is very much called for.

Let us now turn to outline a few typical (of the numerous) questions associated with specific systems, in order to illustrate the issues and systems whose understanding rests on the concepts mentioned above. Such are the effects of charging and quantum confinement within corresponding particles that are embedded in a percolating system. The understanding of the interplay between these effects and the effects of the neighboring network is still in its infancy. This interplay has been shown recently to have pronounced effects not only on the transport, but also on the phototransport in such systems. Also, the effects of special features of the network, such as its fractal dimension, on the dynamical behavior are still a matter of controversy. Again, local probe microscopy, on the experimental end, and corresponding theoretical-computational work, on the other end, may resolve the corresponding problems. Another issue of great interest is the consideration of objects that are more complicated in their shape than the convex objects dealt with in the present review. Such are the systems that are of great present interest, i. e., carbon nanotube composites, where the objects can be described as having a "wavy" shape. One question that arises there, and that was dealt with only very recently, is the effect of this "wavyness" on the percolation threshold. A broader issue that is relevant to many systems, but for which no systematic discussion has been given thus far, is the relation between the trends in the critical behavior of the electrical, mechanical and rheological properties, in corresponding solids or molten composites. These are of great importance in the fields of chemical-material and electrical-material engineering.

Finally, a subject that has not been discussed in the context of continuum percolation, and, as we saw above, is mainly concerned with global properties of systems, is the statistics of finite clusters where there is, or there is no, interaction between the objects. This is of great importance for systems where the finite clusters as such determine the properties of interest, such as in cases where there is a charge transfer by a delocalization process. This is of importance, for example, in the cases of illumination-generated charge carriers, since there, the local connectivity of the system around the particle, in which the carriers were optically generated, determines the probability of their separation and thus the "deconfinement" and corresponding elimination of the radiative recombination.

Acknowledgments

The present review could not have been written without the stimulation and the intensive collaboration that I had with the many colleagues and students, whose papers that were co-authored with me, are cited in this review. In particular, I would like to thank Dr. C. Grimaldi and Dr. N. Wagner for the on going collaboration and for the critical reading of the present manuscript. This work was supported by the Israel Science Foundation (ISF). The author holds the Enrique Berman chair in Solar Energy Research at the Hebrew University.

Bibliography

Primary Literature

1. Abeles B (1976) Granular metal films. Appl Solid State Sci 6:1–117
2. Abeles B, Pinch HL, Gittleman JI (1975) Percolation conductivity in W-Al_2O_3 granular metal films. Phys Rev Lett 36:257–260
3. Alon U, Drory A, Balberg I (1990) Systematic derivation of percolation thresholds in continuum systems. Phys Rev A 42:4634–4638
4. Alon U, Balberg I, Drory A (1991) New, heuristic, percolation criterion for continuum systems. Phys Rev Lett 66:2879–2882
5. Ambegaokar V, Halperin BI, Langer JS (1971) Hopping conductivity in disordered systems. Phys Rev B 4:2612–2620
6. Andrade JS et al (2000) Flow between two sites on a percolation cluster. Phys Rev E 62:8270–8281
7. Archie GE (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. Trans Am Inst Min Metall Pet Eng 146:54–62
8. Azulay D et al (2003) Electrical-themal switching in carbon black-polymer composites as a local effect. Phys Rev Lett 90:236601
9. Balberg I (1986) Excluded-volume explanation of Archie's law. Phys Rev B 33:3618–3620
10. Balberg I (1987) Tunneling and nonuniversal conductivity in composite materials. Phys Rev Lett 59:1305–1308
11. Balberg I (1987) Recent developments in continuum percolation. Philos Mag B 56:991–1002
12. Balberg I (1998) New limits on the continuum-percolation transport exponents. Phys Rev B 57:13351–13354
13. Balberg I (2002) A comprehensive picture of the electrical transport phenomena in carbon black-polymer composites. Carbon 40:139–143
14. Balberg I, Binenbaum N (1983) A Computer study of the percolation threshold in a two-dimensional anisotropic system of conducting sticks. Phys Rev B 28:3799–3812
15. Balberg I, Binenbaum N (1985) Cluster structure and conductivity of three-dimensional continuum Systems. Phys Rev A 31:1222–1225
16. Balberg I, Binenbaum N (1987) Invariant properties of the percolation thresholds in the soft core-hard core transition. Phys Rev A 35:5174–5177
17. Balberg I, Binenbaum N (1987) Scher and Zallen criterion: applicability to composite systems. Phys Rev B 35:8749–8752
18. Balberg I, Blanc J (1985) Capacitive noise spectra of a disordered material. Phys Rev B 31:8295–8297
19. Balberg I, Bozowski S (1982) Percolation in composites of random stick-like conducting particles. Solid State Commun 44:551–554
20. Balberg I, Binenbaum N, Anderson CH (1983) Critical behavior of the two-dimensional sticks system. Phys Rev Lett 51:1605–1609
21. Balberg I, Binenbaum N, Wagner N (1984) Percolation thresholds in the three-dimensional sticks system. Phys Rev Lett 52:1465–1609
22. Balberg I et al (1984) Excluded volume and its relation to the onset of percolation. Phys Rev B 30:3933–3943
23. Balberg I et al (1988) Critical behavior of the electrical resistance and its noise in inverted random-void systems. Phys Rev Lett 60:1887–1890
24. Balberg I et al (1990) Tunneling and percolation behavior in granular metals. Mater Res Soc Symp Proc 195:233–238
25. Balberg I, Berkowitz B, Drachsler GE (1991) Application of a percolation model to flow in fractured hard rocks. J Geophys Rev 96:10015–10021
26. Balberg I et al (2004) Percolation and tunneling in composite materials. Int J Mod Phys B 18:2091–2121
27. Balberg I et al (2007) Fundamental transport processes in ensembles of silicon quantum dots. Phys Rev B 75:225–329
28. Bergman DJ (2003) Exact relation between critical exponents for elastic stiffness and electrical conductivity of percolation systems. Phys B Condens Matter 338:240–246
29. Bergman DJ et al (1983) Critical behavior of the low-field Hall conductivity at a percolation threshold. Phys Rev Lett 50:1512–1515
30. Berhan L, Sastry AM (2007) Modeling percolation in high-aspect-ration fiber systems. Phys Rev E 75:041–121
31. Berkowitz B, Balberg I (1992) Percolation approach to the problem of hydraulic conductivity in porous media. Transp Porous Media 9:275–286
32. Berkowitz B, Balberg I (1993) Percolation theory and its application to groundwater hydrology. J Water Resour Res 29:775–794
33. Broadbent SR, Hammersley JM (1957) Percolation processes, crystals and mazes. Proc Camb Philos Soc 53:629–641
34. Bug ALR, Safran SA, Webman I (1985) Continuum percolation of rods. Phys Rev Lett 54:1412–1415
35. Bug ALR et al (1985) Do interactions raise or lower a percolation threshold? Phys Rev Lett 55:1896–1899

36. Cametti C et al (1990) Theory and experiment of electrical conductivity and percolation locus in water. Phys Rev Lett 64:1461–1464
37. Charlaix E, Guyon E, Rivier N (1984) A criterion for percolation threshold in a random array of plates. Solid State Commun 50:999–1002
38. Charlaix E, Guyon E, Roux S (1987) Permeability of a random array of fractures of widely varying apertures. Transp Porous Media 2:31–43
39. Chatterjee AP (2000) Continuum percolation in macromolecular fluids. J Chem Phys 113:9310–9317
40. Chen CC, Chou YC (1985) Electrical-conductivity fluctuations near the percolation threshold. Phys Rev Lett 54:2529–2532
41. Chiew YC, Glandt ED (1983) Percolation behavior of permeable and of adhesive spheres. J Phys A 16:2599–2608
42. Chiteme C, McLachlan DS, Balberg I (2003) 1/f or flicker noise in cellular percolation systems. Phys Rev B 67:024207
43. Clerc JP et al (1990) The ac electrical conductivity of binary-disordered systems, percolation clusters, fractals and related models. Adv Phys 39:191–308
44. Coniglio A, DeAngelis U, Forlani A (1977) Pair connectedness and cluster size. J Phys A Math Gen 10:1123–1139
45. Dalmas F et al (2006) Carbon nanotube-filled polymer composites. Numerical simulations of electrical conductivity in three-dimensional entangled fibrous networks. Acta Mater 54:2923–2931
46. Dalton NW, Domb C, Sykes MF (1964) Dependence of the critical concentration of dilute ferromagnet on the range of interaction. Proc Phys Soc 83:496–498
47. Day AR et al (2003) Spectral representation of the electrical properties of layered materials. Phys B Condens Matter 338:24–30
48. de Arcangelis L, Redner S, Coniglio A (1985) Anomalous voltage distribution of random resistor networks and a new model for the backbone at the percolation threshold. Phys Rev B 31:4725–4727
49. DeSimone T, Stratt RM, Demoulini S (1986) Continuum percolation in an interacting system: Exact solution of the Percus-Yevick equation for connectivity in liquids. Phys Rev Lett 56:1140–1143
50. Drory A (1996) Theory of continuum percolation. I. General formalism. Phys Rev E 54:5992–6002
51. Drory A (1996) Theory of continuum percolation. II. Mean field theory. Phys Rev E 54:6003–6013
52. Drory A (1997) Exact solution of a one-dimensional continuum percolation model. Phys Rev E 55:3878–3885
53. Drory A et al (1991) Analytic derivation of percolation thresholds in anisotropic systems of permeable objects. Phys Rev A 43:6604–6612
54. Drory A, Balberg I, Berkowitz B (1994) Random-adding determination of percolation thresholds in interacting systems. Phys Rev E 49:R949–952
55. Drory A, Balberg I, Berkowitz B (1995) Application of the central-particle potential approximation for percolation in interacting systems. Phys Rev E 52:4482–4494
56. Drory A et al (1997) Theory of continuum percolation. III. Low-density expansion. Phys Rev E 56:1379–1395
57. Du F et al (2004) Nanotube networks in polymer nanocomposites: Rheology and electrical conductivity. Macromolecules 37:9048–9055
58. Englman R, Gur Y, Jaeger Z (1983) Fluid flow through a crack network in rocks. J Appl Mech 50:707–711
59. Feng S, Halperin BI, Sen PN (1987) Transport properties of continuum systems near the percolation threshold. Phys Rev B 35:197–214
60. Flory PM (1941) Molecular size distribution in three dimensional polymers. J Am Chem Soc 63:3083–3100
61. Fonseca LF, Balberg I (1993) Resistivity and electrical noise in granular metal composites. Phys Rev B 48:14915–14924
62. Fortuin CM, Kasteleyn PW (1972) Random-cluster model. 1. Introduction and relation to other models. Phys 57:536–564
63. Fortunato S, Stauffer D, Coniglio A (2004) Percolation in high dimension is not understood. Phys A 334:307–311
64. Foygel M et al (2004) Theoretical computational studies of carbon nanotube composites and suspensions: electrical and thermal conductivity. Phys Rev B 71:164201
65. Gawlinski ET, Redner S (1983) Monte Carlo renormalization group for continuum percolation with excluded-volume interactions. J Phys A Math Gen 16:1063–1071
66. Gawlinski ET, Stanley HE (1981) Continuum percolation in two dimensions: Monte carlo tests of scaling and universality for non-interacting discs. J Phys A Math Gen 14:L291–L299
67. Grannan DM, Garland JC, Tanner DB (1981) Critical behavior of the dielectric constant of a random composite near the percolation threshold. Phys Rev Lett 46:375–378
68. Grassberger P (2003) Critical percolation in high dimensions. Phys Rev E 67:036101
69. Grest GS et al (1986) Dynamic percolation in microemulsions. Phys Rev A 33:2842–2845
70. Grimaldi C, Balberg I (2006) Tunneling and non-universality in continuum percolation systems. Phys Rev Lett 96:066602
71. Grimaldi C et al (2003) Segregated tunneling-percolation model for transport nonuniversality. Phys Rev B 68:024207
72. Halperin BI, Feng S, Sen PN (1985) Differences between lattice and continuum percolation transport exponents. Phys Rev Lett 54:2391–2394
73. Haan SW, Zwanzig R (1977) Series expansions in a continuum percolation problem. J Phys A Math Gen 10:1547–1555
74. Hansen PH, McDonald RM (1986) Theory of Simple Liquids. Academic, London
75. Heyes DM, Melrose JR (1998) Percolation thresholds of simple liquids. J Phys A Math Gen 21:4075–4081
76. Holcomb DF, Rehr JJ (1969) Percolation in heavily doped semiconductors. Phys Rev 183:773–776
77. Isichenko MB (1992) Percolation, statistical topography, and transport in random media. Rev Mod Phys 64:961–1043
78. Johner N et al (2008) Transport exponent in a three-dimensional continuum tunneling-percolation model. Phys Rev B 77:174–204
79. Kapitulnik A, Deutscher G (1982) Percolation characteristics in discontinuous thin films of Pb. Phys Rev Lett 43:1444–1448
80. Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
81. Knite M et al (2002) Electric and elastic properties of conductive poylmer nanocomposites on macro- and nanoscales. Mater Sci Eng C 19:15–19
82. Kogut PM, Straley J (1979) Distribution–induced non-universality of the percolation conductivity exponents. J Phys C Solid State Phys 12:2151–2159

83. Laria D, Vericat F (1989) Percolation behavior of long permeable objects: A reference interaction-site-model study. Phys Rev B 40:353–360
84. Lin C-R, Chen W-J (1999) The links–nodes-blobs model for shear thinning-yield stress fluids. Colloid Polym Sci 277:1019–1025
85. Mandal P et al (1997) Temperature and magnetic field dependence of the resistivity of carbon-black composites. Phys Rev B 55:452–456
86. McCarthy JF (1987) Continuum percolation of disks and the random lattice. Phys Rev Lett 58:2242–2244
87. Miller A, Abrahams E (1960) Impurity conduction in low concentrations. Phys Rev 120:745–755
88. Murat M, Mariner S, Bergman DJ (1986) A transfer matrix study of the conductivity and permeability exponents in continuum percolation. J Phys A Math Gen 19:L275–L279
89. Neda Z, Florian R, Brechet Y (1999) Reconsideration of continuum percolation of isotropically oriented sticks in three dimensions. Phys Rev E 59:3717–3719
90. Netemeyer SC, Glandt ED (1986) Percolation behavior of the square-well fluid. J Chem Phys 85:6054–6059
91. Octavio M et al (1988) Nonuniversal critical behavior in the critical current of superconducting composites. Phys Rev B 37:9292–9297
92. Pagnotta SE et al (2005) Glassy behavior of a percolative water-protein system. Phys Rev E 71:031506
93. Park S et al (2004) Percolation conduction in the half-metallic ferromagnetic and ferroelectric mixture of (La,Lu,Sr)MnO_3. Phys Rev Lett 92:167206
94. Pike GE (1978) Conductivity of thick films (cermet) resistors as a function of metallic particle volume fraction. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media (AIP Conf Proc 40). AIP, New York, pp 366–371
95. Pike GE, Seager CH (1974) Percolation and conductivity: a computer study I. Phys Rev B 10:1421–1434
96. Planes J et al (1998) Transport properties of polyanilline-cellulose-acetate blends. Phys Rev B 58:7774–7785
97. Powell MJ (1979) Site percolation in randomly packed spheres. Phys Rev B 20:4194–4198
98. Rapp O, Shinivas V, Poon SJ (2005) Critical exponents at the metal-insulator transition in AlPdRe quasicrystals. Phys Rev B 71:012202
99. Re A et al (2006) Correlated fragile site expression allows the identification of candidate fragile genes involved in immunity and associated with carcinogenesis. BMC Bioinformatics 7:413
100. Rintoul MD, Torquato S (1997) Precise determination of the critical threshold and exponents in three-dimensional continuum percolation model. J Phys A Math Gen 30:L585–L592
101. Robinson PC (1983) Numerical calculations of critical densities for lines and planes. J Phys A Math Gen 17:2823–2830
102. Rubin Z et al (1999) Critical behavior of the electrical transport properties in a tunneling-percolation system. Phys Rev B 49:12196–12199
103. Saar MO, Manga M (2002) Continuum percolation of randomly oriented soft-core prisms. Phys Rev E 65:056131
104. Satz H, Fortunato S (2001) Percolation and confinement in SU(2) gauge theory. Nucl Phys A 681:466C–471C
105. Sahimi M (1994) Application of Percolation Theory. Taylor, London
106. Sahimi M (1998) Non-linear and non-local transport in heterogeneous media: from long-range correlated percolation to fracture and materials breakdown. Phys Rep Rev Sect Phys Lett 306:214–295
107. Sahimi M, Imdakm AO (1988) The effect of morphological disorder on hydrodynamic dispersion in flow through porous media. Phys A Math Gen 21:3833–3870
108. Scher H, Zallen R (1970) Critical density in percolation processes. J Chem Phys 53:3759–3761
109. Schrijver CJ et al (1992) Patterns in the photosphereic magnetic-field and percolation theory. Astron Astrophys 253: L1–L4
110. Sen PN, Scala C, Cohen MH (1981) A self-similar model for sedimentary rocks with application to the dielectric constant of fused glass beads. Geophysics 46:781–795
111. Sen PN, Roberts JN, Halperin BI (1985) Nonuniversal critical exponents for transport in percolating systems with a distribution of bond strengths. Phys Rev B 32:3306–3308
112. Shante VKS, Kirkpatrick S (1971) An introduction to percolation theory. Adv Phys 20:325–357
113. Shimoni N et al (2002) tomographic-like reconstruction of the percolation cluster as a phase transition. Phys Rev B 66:020102R
114. Shklovskii BI, Efros AL (1984) Electronic properties of doped semiconductors. Springer, New York
115. Skal AS, Shklovskii BI (1974) Influence of impurity concentration on the hopping conduction in semiconductors. Sov Phys Semicond 7:1058–1059
116. Smart JS (1968) Effective field theories of Magnetism. Sanders, Philadelphia
117. Sokolowska D, Krol-Otwinowska A, Moscicki JK (2004) Water network percolation transition in hydrated yeast. Phys Rev E 70:052901
118. Song Y, Lee S-I, Gaines JR (1992) Ac conductivity and $1/f$ noise in a Cr-film lattice-percolation system. Phys Rev B 46:14–20
119. Stanley HE (1971) Introduction to phase transitions and critical phenomena. Clarendon, Oxford
120. Stanley HE (1977) Cluster shapes at the percolation threshold: an effective cluster dimensionality and its connection with critical-point exponents. J Phys A Math Gen 10:L211–L220
121. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor, London
122. Stinchcombe RB (1976) Conductivity and spin-wave stiffness in disordered systems-an exactly soluble model. J Phys C Solid State Phys 7:179–203
123. Stockmayer WH (1943) Theory of molecular size distribution and gel formation in branched-chin polymers. J Chem Phys 11:45–55
124. Toker D et al (2003) Tunneling and percolation in metal-insulator composite materials. Phys Rev B 68:041403
125. Thompson AH, Katz AJ, Krohn CE (1987) The microgeometry and transport properties of sedimentary rocks. Adv Phys 365:625–694
126. Torquato S, Lu B, Rubinstein J (1990) Nearest neighbor distribution function in many-body systems. Phys Rev A 41:2059–2075
127. Tremblay A-MS, Fourcade B, Breton P (1989) Multifractals and noise in metal-insulator composites. Phys A 157:89–100
128. Vionnet-Menot S et al (2005) Tunneling-percolation origin of nonuniversality: theory and experiments. Phys Rev B 76:064201

129. Vyssotsky VA et al (1961) Critical percolation probabilities (bond problem). Phys Rev 123:1566 1567
130. Wagner N, Balberg I (1987) Anomalous diffusion and continuum percolation. J Stat Phys 59:369–382
131. Wagner N, Balberg I, Klein D (2006) Monte carlo results for continuum percolation in low and high dimensions. Phys Rev E 74:021127
132. Williams JC, Snyder SA (2005) Restoring habitat corridors in fragmented landscapes using optimization and percolation models. Environ Model Assess 10:239–250
133. Wilkinson D (1986) Percolation effects in immiscible displacement. Phys Rev A 34:1380–1391
134. Wille JJ, Elson EL, Okamoto RJ (2006) Cellular and matrix mechanics of bioartificial tissues during continuous cyclic stretch. Ann Biomed Eng 34:1678–1690
135. Wu ZH et al (2006) Optimal paths in complex networks with correlated weights: the worldwide airport network. Phys Rev E 74:056104
136. Zallen R (1983) The Physics of Amorphous Solids. Wiley, New York
137. Zallen R, Scher H (1971) Percolation on a continuum and the localization delocalization transition in amorphous semiconductors. Phys Rev B 4:4471–4479
138. Ziman JM (1979) Models of Disorder. Cambridge University Press, New York

Books and Reviews

Balberg I (1987) Recent developments in continuum percolation. Philos Mag B 56:991–1002
Balberg I et al (2004) Percolation and tunneling in composite materials. Int J Mod Phys B 18:2091–2121
Berkowitz B, Balberg I (1993) Percolation theory and its application to groundwater hydrology. J Water Resour Res 29:775–794
Isichenko MB (1992) Percolation, statistical topography, and transport in random media. Rev Mod Phys 64:961–1043
Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
Sahimi M (1994) Applications of Percolation Theory. Taylor, London
Sahimi M (1998) Non-linear and non-local transport in heterogeneous media: from long-range correlated percolation to fracture and materials breakdown. Phys Rep Rev Sect Phys Lett 306:214–295
Shklovskii BI, Efros AL (1984) Electronic properties of doped semiconductors. Springer, New York
Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor, London
Zallen R (1983) The Physics of Amorphous Solids. Wiley, New York

Continuum Robots

Ian D. Walker[1], Keith E. Green[2]
[1] Department of Electrical and Computer Engineering, Clemson University, Clemson, USA
[2] School of Architecture, Clemson University, Clemson, USA

Article Outline

Glossary

Degree of freedom (Sub)set of a physically reconfigurable system whose configuration can always be specified by a single variable.

Kinematics Geometric (and differential geometric) models of mechanical systems, not including effects of dynamics (masses, forces, inertias, etc.).

Robot A physically reconfigurable system that has multiple degrees of freedom and is (at least partially) computer-controlled.

Actuators Devices used to transfer mass (herein portions of a robot system) from one place to another within the system's environment.

Sensors Devices used to infer information, usually either about the internal state of a system, or about external environment surrounding that system.

Control Use of sensors and actuators to reconfigure a system (herein robots) according to a desired plan.

Continuum Continuous in nature; herein usually the backbone/core of a robot.

OctArm Continuum robot manipulator inspired by octopus arms.

AWE ("Animated Work Environment") Reconfigurable environment featuring continuum robot components, focused on work environments featuring computing.

Definition of the Subject

In this article, we discuss some of the key issues involved in the design and implementation of the emerging class of "invertebrate-like" continuum robots. Using two case studies of continuum robots developed recently at Clemson University, we overview the issues involved in realizing continuum robots and their deployment. The potential of these types of robots for enhanced productivity in novel applications is discussed.

In the first case study, we describe the design of the "OctArm" continuum manipulator robot hardware, and discuss the results of field testing of these novel "trunk-like" robots. OctArm robots are able to adapt their shape to their environment, to access difficult-to-reach areas, and to perform adaptive grasping using their entire arms. Lessons learned and implications for future robot manipulators in the field are discussed.

In the second case study, we describe the new "Animated Work Environment" (AWE) concept. AWE is an articulated, programmable, interior environment and chassis, embedded with integrated digital technologies. Featuring a continuum, morphing robot surface controlled by a user-friendly interface, AWE is characterized as adaptive and robust when applied to a range of work activities and working populations currently not accommodated by robotics.

The OctArm and AWE were selected as case studies to represent the potential broad range of application of continuum robots: the former suggests how continuum robots might operate in environments too harsh and hazardous for routine human activity, and the latter offers the promise of continuum robots in everyday interior environments supporting human collaboration and productivity.

Introduction

Industrial robot manipulators typically feature a small finite number (typically 4–6) of serially-connected rigid links. This is well-matched to factory "pick-and-place" applications, but such manipulators have not been generally successful in tackling more complex tasks (such as interacting with humans) in less structured environments. A frequently considered question in robot manipulator research is: "what happens as the number of joints becomes much larger, and ultimately increases towards infinity?" Examples of serial rigid-link systems with many joints exist in nature (snakes, mammalian and fish backbones, for example), and these biological "existence proofs" have provided motivation and insight for robot analysts and hardware designers for many years. This interest has resulted in the development of several special sub-classes of manipulators, collectively known as *redundant manipulators* [86].

A manipulator is said to be (kinematically) *redundant* if its configuration (joint) space degrees of freedom exceed its task space degrees of freedom. Hence, since spatial (end effector) task motions are generally six-dimensional, seven or eight degree of freedom spatial rigid-link manipulators are usually considered to be redundant. A manipulator is generally considered to be *hyperredundant* if its configuration (joint) space degrees of freedom greatly exceed its task space degrees of freedom. Hyperredundant manipulators have enhanced potential to use their extra joints for whole arm grasping/manipulation and maneuvering within tight obstacle fields. Their anticipated applications therefore include operations in congested environments (disaster relief, medical applications, etc.).

The field of hyperredundant manipulators has evolved in two major directions: "vertebrate-like" rigid-link designs, and "invertebrate-like" continuum manipulators [86]. Here we describe recent developments in continuum robots. Continuum robots are most typically defined by their "continuous backbone structures" – structures without the skeletal design of traditional rigid-link, robot manipulators [66,78]. Defined by this unusual structure, continuum robots may be described as "invertebrate-like" as opposed to the "vertebrate-like" nature of traditional robots [33]. Much research in the area takes inspiration from biology, and has often been strongly influenced by the design and functionality of invertebrate structures such as octopus arms [46,50,79,81] and elephant trunks [19,30,46,76,85]. As continuum designs can be scaled down to very small sizes [69,72,73], potential applications include medicine [4,12,68,75,82].

The main feature of continuum structures is their inherent smoothness. Instead of bending at discrete points (joints, or "elbows") along the "backbone", they can smoothly bend anywhere along their structure. Several designs, including the OctArm robots discussed in this article, additionally feature the ability to extend (elongate) along the length of the backbone. Almost all continuum robot designs exhibit significant compliance, i. e. they inherently present a "soft" rather than "stiff" interface to the environment.

Two functions in particular define the potential advantages of continuum robots: operation in congested environments and whole-arm grasping. In congested environments, the lack of "elbows" and the ability to bend at many locations allows continuum structures to "snake" in and through very tight obstacle fields. (Snakes are vertebrates, but their "links" are small and many, and their general movements match that of continuum structures.) There are numerous important arenas in which key applications of traditional robotics are currently impractical (many surgical applications, search and rescue operations in collapsed buildings, etc.) and where the effective deployment of continuum structures exhibiting snake-like behaviors would make a significant contribution to Society.

In whole-arm grasping, continuum robots again prove their potential advantage: their smooth and compliant nature allows them to gently interact with the world by adapting their shape to that of environmental objects (i. e.

contact along a continuum of the robot) [67]. Note that this is the form of adaptive manipulation used by biological continuum structures such as elephant trunks and octopus arms.

In the following case studies, we outline the design and development of two very different types of continuum robot design, the "OctArm" series of manipulator, and the "AWE" continuum surface. In both cases, the ability of the continuum robot structures to bend smoothly is key to enabling the innovative application scenarios.

Case Study 1: Trunk Continuum Robots – OctArm

Introduction to Octarm Robots

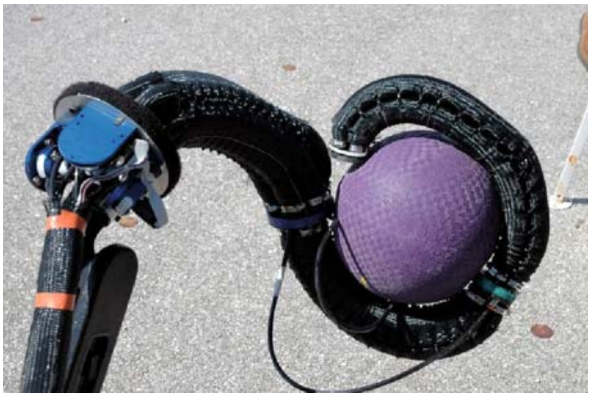
Continuum robot hardware designs can be sub-divided (via their means of actuation) into the categories of "extrinsically actuated" and "intrinsically actuated". Extrinsically actuated continuum robots have their actuators located outside the continuum structure itself, and typically transmit power to the continuum structure though tendons. Intrinsically actuated continuum robots have the actuators embedded in the continuum structure.

At Clemson, we have developed continuum robot hardware actuated both extrinsically and intrinsically. In this section, we discuss the OctArm series of intrinsically actuated continuum arms developed over the past several years under funding from DARPA. The OctArm design was significantly influenced by the underlying design of octopus arms [46,50,70]. Key OctArm design goals were to keep the arm as "soft" and compliant as possible, while incorporating the actuators in the body of the arm.

The resulting robots feature a serially arranged set of "sections". Each section corresponds roughly to a link in a conventional rigid link robot [71]. The variables determining changes in shape of each section correspond conceptually to "joint angles" in conventional robot arms. However, the shape of each section in the OctArm design is determined by two or more variables in a coupled nonlinear relationship. A series of kinematic models have been developed in order to command the OctArms into desired shapes. These models are easily generalized to a wide variety of previously proposed and constructed continuum robots.

Overview of Octarm Robots

The OctArm series of continuum robots feature three (or four) sections, each section having three degrees of freedom, for a total of nine (or twelve) degrees of freedom per arm (see Fig. 1, showing a three-section version) [64]. Each section has two degrees of bending freedom, and one degree of extension. The bending and extension in a section are achieved via variation of pressure in three "air muscles" (McKibben actuators) built into the system (in fact, the air muscles – tubes in the figure – form the majority of the arm structure).



Continuum Robots, Figure 1
Three-section Octarm V continuum robot

The arms are approximately one meter long, with a roughly circular cross-sectional profile, and are tapered from base to tip. The air pressure in the actuators is controlled via commercial pressure regulators, which form the inner loop of the feedback control system. At a higher level, commands to the pressure regulators are provided via a real-time kernel on a PC-104 system, with wireless interface to an operator control PC/joystick. The control code is written in C+. Algorithm development and user interface code utilize Matlab and Simulink modules [20].

The shape of the arms is measured via cables integrated into the arm. The cables change length as the arm moves; changes in cable length are then measured by encoders at the base; and the shape of the arm is inferred by changes in the cable length, following a geometric model. This approach gives a good approximate shape of the arm, which is displayed to the operator via a graphical interface, driven by the sensor information in real-time.

Related Work

Continuum robot hardware designs (usually arms/trunks or fingers/whiskers, though some robot snake/worm [33] designs also fit the classification) have appeared in the literature since the early 1960's [2]. Although there have been many proposed continuum robot designs [51], and numerous hardware realizations [3,10,15,29,35,36,37,47,53,61,62,84], until recently they have largely been "laboratory curiosities", used for a few "proof-of-concept" demonstrations and then abandoned. The main reason for this has been a lack of underlying theory, corresponding to traditional robot kinematics and dynamics, to enable the coordination of their degrees of freedom. This deficiency meant that even "simple" tasks such as moving the tip of the robots in a straight line were difficult if not impossible to achieve. This in turn has made it very difficult to apply continuum robots to real tasks.

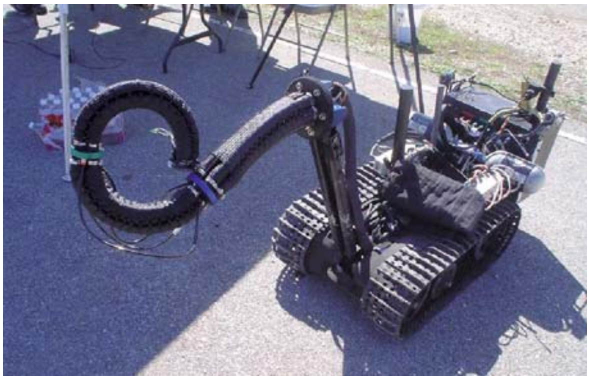
However, in the last ten years, a complete and general body of theory (at least for kinematics) has been established [38,39,40,58,59]. Our group at Clemson University has been a major contributor to this theory [24,25,26,30, 31,32,41,43,44,45,74,80] building on existing pioneer efforts [13,14,16,17,83]. This emerging body of understanding now enables the consistent coordination of the degrees of freedom within general continuum structures. Programming of continuum robots can now theoretically be done in terms of "task space" and "configuration space" in an analogous way to that for conventional robots.

At Clemson, the first author's group has demonstrated the ability for the "OctArm" continuum robot hardware to successfully maneuver through complex obstacle fields and grasp and manipulate objects of many different shapes over a wide range (orders of magnitude) of size and weight, and with widely different physical characteristics (rigid, soft, flexible ...) [8,9,18]. The following section describes the deployment of the OctArm hardware in realistic field trials.

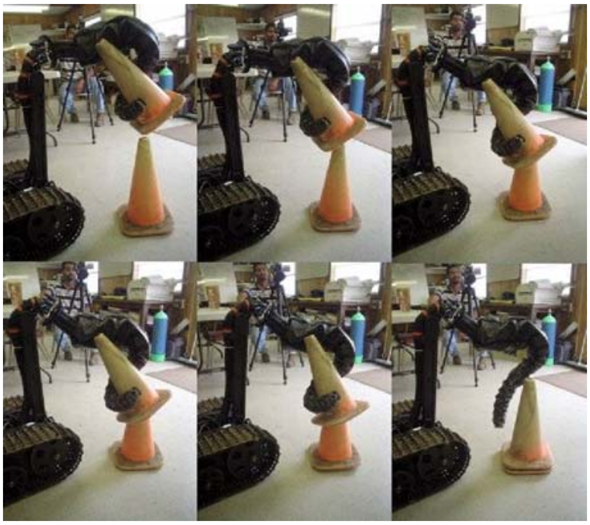
Example Octarm Field Trials

The OctArm system underwent extensive field trials in the spring of 2005 and 2006 at the Southwest Research Institute (SwRI) in San Antonio, Texas, USA [42,52]. The SwRI testbed is specially designed for the systematic evaluation of robotic systems in the field. The test environment includes rubble piles and a dry riverbed which can be flooded to create turbulent conditions. The main goal for the trials was to evaluate the ability of the OctArm system to stably and adaptively grasp a wide range of objects under a variety of conditions. For the field tests, OctArm V was mounted on a Foster–Miller TALON platform (shown in Fig. 2).

The OctArm base was attached to the second link of the TALON robot arm. The control valves and two air tanks provided nine channels of controlled pneumatic pressure. The control computer was mounted on the back of the TALON. The TALON and arm were controlled via wireless connection. The system was operated under joystick control via the wireless link and in view of the operator. More details of the user interface used are described in [20].



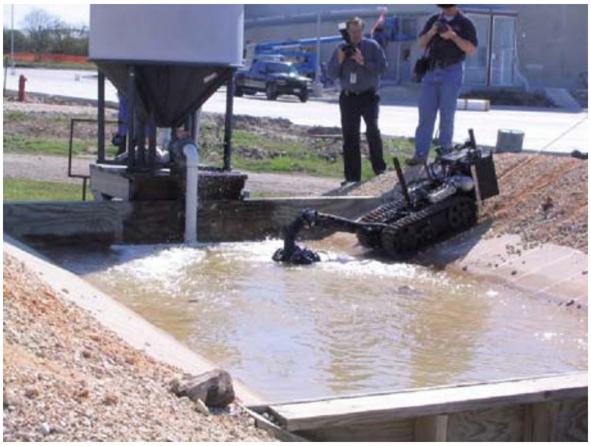
Continuum Robots, Figure 2
OctArm V mounted on Foster–Miller TALON system



Continuum Robots, Figure 3
Cone stacking task

Tasks included stacking and unstacking traffic cones (see Fig. 3). The ability of the system to grasp objects, such as spheres and cylinders, over a wide range of scales was recorded. These tasks are inherently problematic for traditional parallel jaw-end effectors. The operations were timed, videotaped, and recorded in detail to provide a baseline performance measure for continuum robots under the above conditions.

The system was also operated in water. Submerged in water (shown in Fig. 4), the OctArm attempted to grasp various payloads and to maintain grasps under turbulent flow. This tested the potential of the system for robust grasping under disturbances. The system was operated as well alongside rubble piles. This provided an initial benchmark for remote operation in congested environments.



Continuum Robots, Figure 4
Grasp in turbulent water

During the trials, the OctArm grasped and manipulated various sized cylinders and spheres. The system was able to successfully conform to the varying shapes of these payloads, demonstrating an ability to adapt to environmental conditions not found in traditional manipulators. The continuum system was able to successfully grasp both spherical (balls) and elongated objects (pieces of wood) within the underwater environment, and to maintain the grasps despite significant flow.

The results of the trials clearly demonstrated that continuum robot arms exhibit more adaptive and versatile operation than do conventional robot manipulators. Note that the continuum arm in these trials featured no shape sensing; the shapes displayed to the user was inferred only from the pressures at the actuator input valves. This proved a highly inaccurate estimate, as external effects such as sag due to gravity or external loading were not accounted for in any way. Shortly after the trials described in this section were performed, we integrated shape sensors into the design in order to reduce this inaccuracy.

Case Study 2: Continuum Robot Surfaces – AWE

Introduction to AWE

Our second case study features a novel robotic system where the continuum "trunk" is expanded into a continuum "surface". The surface forms a key part of our "Animated Work Environment" (AWE), aimed at work activities and working populations currently not accommodated by robotics [27,28].

A recent trend in the nature, place and organization of working life is the growing complexity of work, the emergence of new working populations (older workers, under-skilled workers, telecommuters and flexible shift workers), and the increasing likelihood that workers are working, at least part-time, at home. As the home becomes more an office, the office is becoming more a home, where "hot desks," lounges, sofas, WI-FI and internal networks have replaced office cubicles and hard-wired, isolated workstations. These dramatic transformations in the nature of work, combined with unprecedented new technologies associated with working life, suggest a re-evaluation of the relations between workers, their technologies and their work environments, and a redesign of the work environment itself as a socially and technologically responsive system occupying both home and office.

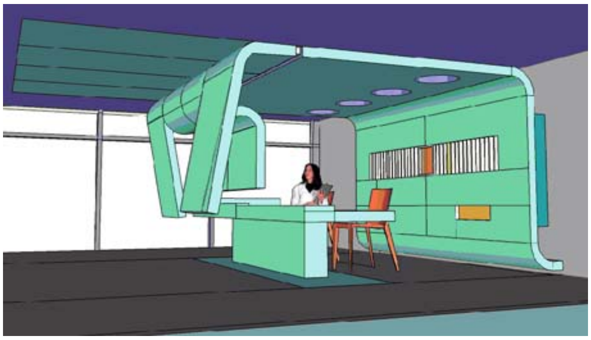
While investigators continue to realize promising components of the "intelligent" work environment – projectors, screens, tablets, sensors, actuators and other digital devices – our collaborative team is focused upon a novel but wholly compatible aspect of the "intelligent" work environment: the physical work space itself. The AWE concept challenges knowledge and understanding in both Architecture and Engineering by defining the "robot as a room" and the "room as a robot." This is a redefinition of what constitutes Architecture and Robotics and is not only a conceptual leap in the respective disciplines, but a fully appropriate, even necessary response to a condition in working life that is both social and technological.

The need to "program the room" both stimulates and is enabled by existing and ongoing efforts in Information Technology and "intelligent environments." More broadly, AWE challenges present understanding in both Architecture and Computer and Information Science and Engineering by recognizing computer software, networks and devices not as isolated aspects of a digital society but as constituents of an integral environmental system, far more productive and more accessible than any of its parts. In this way, the AWE concept aims to expand the vision of researchers, developers and manufacturers of information technologies to recognize the physical environment as an integral and necessary part of the dynamic interaction between people and the digital realm.

Overview of AWE

The ongoing dramatic transformation in working life, including the introduction of ever new digital technologies, presents problems and opportunities to all workers, particularly to segments of the working population that are emerging and neglected: telecommuters, flexible shift workers, single parents, elders, recent immigrants, the obese, the handicapped and, other individuals requiring special accommodations.

Continuum Robots, Figure 5
AWE concept – SLEEPING



Continuum Robots, Figure 6
AWE concept – COMPOSING

This dramatic shift in the nature, place and organization of working life motivates our research which, in the simplest of terms, involves the designing, prototyping, demonstrating and evaluating of a prototypical "robot-room" with embedded Information Technologies that we call an "Animated Work Environment" (AWE) (Figs. 5 and 6).

The strength of AWE is made clearer by recognizing what it isn't: it isn't a building, or a room, or a "stand-alone" device, or a software application, or a piece of furniture. Instead, AWE is a user-friendly, programmable environment, both digital and analog, high-tech and low-tech, fitted to home and office, that users adjust along a continuum, providing the sense of being more "at home" or more "at work," more leisurely or more productive, more efficient or more innovative, while facilitating multiple activities.

In concept, AWE is envisioned as an information-rich environment featuring the ability to continuously "morph" its continuum surfaces to accommodate a wide range of user needs. At the core of this environment (though not exclusively comprising it) are smooth, continuously deformable "smart" continuum surfaces whose configuration, and hence functionality, are user-controllable. In addition to this novel aspect, AWE embodies a range of "off-the-shelf" Information Technology (IT) components: embedded commercially-available sensors that, when suitably exploited, make AWE user-friendly and intelligent; radio-frequency identification (RFID) tags that allow AWE to associate printed and digital materials; and integrated display screens, scanners, projectors, keyboards and audio speakers that make AWE useful as a total work environment programmable to suit a range of work needs and situations.

Related Work

Few precedents for AWE are found in the work environments being researched and developed by IT and the architectural design industries. The most promising of these efforts are perhaps IBM's "Blue Space" [34] and IDEO's "Q" [65], both important steps towards integrating IT and Design. Compared to our vision of AWE, however, "Blue Space" contains fewer and more timid "smart" components and a narrower range of embedded IT peripherals; and "Q," accommodates only one user who must be seated, supports a far more limited range of work activities, and cannot be reconfigured in the way AWE promises. A more enticing integration of design and IT is found in numerous efforts offered by Phillips as "Ambient Intelligence" 0; but as single products rather than complex environments, these too fall short of AWE's promise.

Additionally, while numerous research efforts in ubiquitous computing [6,7,22,23,60] are viewed by the AWE team as significant, compatible, and parallel efforts to our own research, the difference between the AWE project and these efforts is made clear by a compelling research project from University of Massachusetts titled "NeTS: Animated Spaces for the Digital Society" [60]. While this project's title and broad objective to "bridge the gap between the physical and digital worlds" sounds strikingly similar to our own, "NeTS" focuses on employing Radio Frequency Identification to link objects with their placement in a room, rather than creating, as envisioned for AWE, the physical room itself, rendered intelligent and embedded with a range of IT technologies. Our AWE research effort focuses less on developing and interfacing individual IT elements that might be embedded in AWE, and more on cultivating the dynamic, reinforcing and reconfigurable relationships between the digital and physical (i. e. environmental) realms responsive to workers and working life in a digital society.

Precedents that most significantly impact our development of AWE, a robot-room, are those few where Architectural and CISE researchers collaborate to realize intelligent physical environments. In broad theoretical terms, the AWE team is inspired by two such convergences drawn from William Mitchell's trilogy of books on IT and the built environment [55,56,57]. The first is Mitchell's vision of the building as a computer: "The building of the near future will function more and more like large computers" [56]. The second is Mitchell's vision of the building as a robot: "Our buildings will become … robots for living in" [56].

It is worth noting that Mitchell moves easily between MIT's School of Architecture and its Media Lab, and that his vision of the building-as-robot is not altogether radical when we recognize that the first known treatise of architecture, by Vitruvius, laid claims to machine design as the proper domain of architecture [49,77]. Indeed, the "smart house" projects by Mitchell's group [48] and others, and, more significantly, the programmable, robotic wall by dECOi [21] and robotic pavilions by Kas Oosterhuis [63], hint at the promise of converging architectural design, interface design and robotic design. AWE draws from these precedents the open source, chassis/plug-in strategy of MIT's *House* [48], the seamless integration of IT and architecture suggested by the MIT/FPC *Media House* [54], and the real-time configurability of the programmable *WEB* pavilion of Oosterhuis [63].

However as "houses," "walls," and "pavilions," these significant precedents for our research, by definition, are not explicitly designed for specified human activities and so, suffer from being too diffuse; whereas AWE, operating at the smaller scale of the room, employing novel morphing surfaces, and devoted to facilitating working life, makes more of its association of new technologies, the physical environment and very particular social conditions.

Example AWE Scenario

The following scenario begins to illustrate the operation and promise of AWE.

Situation Laura, a single-mother/biologist, presents her research proposal before a committee in two weeks. Laura's proposal draws significantly from documents in digital format, notes from a recent conference, a video she made, and several books. She is working on her proposal mostly at home but also at her faculty office. Meanwhile Laura's nephew Roberto, a college student, is visiting from Latin America.

Before AWE Laura's documents are spread over the desk, computer, floor, and filing cabinet of her home office. Books are opened, face-down, to critical pages. One of these books and two of the digital papers are most important to her proposal. She's periodically disrupted by her four-year old child, Eric, who pulls materials off Laura's desk when he wants her attention. Laura senses that her nephew Roberto has a real interest in science, but she doesn't know how to sustain this interest over his visit, and, anyhow, he's on vacation – easily distracted and easily bored.

Working with AWE for the first time AWE is installed and ready to operate in Laura's living room (Fig. 5). It looks to her more like a dining room table than an office desk, a cubicle, or a work station. Laura quickly recognizes six prominent icons on AWE's work surface that read: COMPOSING, PRESENTING, COLLABORATING, MEETING, VIEWING, PLAYING.

As Laura urgently needs to work on her research proposal, she imagines COMPOSING best defines her needs and engages that icon. Quickly and steadily, four ribbons descend from the ceiling, each embedded with a computer screen arrayed at eye level when seated at the work surface (Fig. 6). The four screens light-up and a keyboard and mouse pop up from the work surface. In the short time that AWE sets itself up for COMPOSING, Laura discovers a toggle option marked SMART / PROGRAM and decides on toggling SMART mode. Nothing happens; that is, until Laura has organized her printed materials and coffee cup on AWE's work surface and sits before its array of four lit screens.

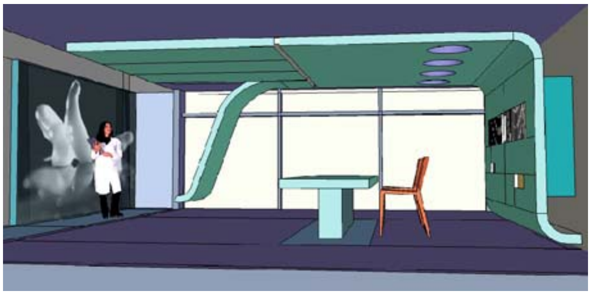
Laura takes notice of a slowly moving curtain encircling one end of AWE's work surface to provide privacy and sound dampening for COMPOSING. Laura's scientific mind correctly infers that AWE, in its simple intelligence, has recognized that she is presently seated, COMPOSING, and so is likely to require this kind of gentle enclosure to facilitate her work. She also senses that AWE has adjusted its overhead lighting to facilitate her focused work. But the desk is too low for Laura's height. Laura moves her hand towards the work surface where she sees the outline of a hand, marked RISE, at which point the surface rises to a height at which Laura is comfortable. The surface stops rising when Laura removes her hand. Laura now presses and holds the COMPOSING button to save this particular setting and then types "Laura – Research Presentation" to define this personal setting. At a later occasion, Laura can simply press COMPOSING to return AWE to this configuration.

Laura now assigns four digital documents to the four digital screens: two screens display the two most important articles, a third displays the video, and the fourth screen displays the titles of some non-digital documents Laura is using to compose the presentation – her handwritten conference notes, nine printed articles, and four books. Laura clicks on the title of a particular printed article she's soon to need, and the screen reports to her that this article (with attached RFID tag), is located in slot-24 of AWE's physical filing system. The lit door over file slot-24 helps Laura quickly locate and retrieve the article.

Laura now touches, one at a time, the two screens displaying the two most important articles for her proposal and these, in AWE's simple intelligence, follow Laura's hand, extending and slightly twisting to a more accessible position for her. In a like manner, Laura sends the two screens displaying the video and the non-digital documents a short distance away and above her so that she is aware of, but not distracted by them.

Forty minutes into her work, Laura becomes aware of her son Eric, growing restless. But Eric, seeing his mother intently working through AWE's translucent privacy curtain, entertains himself for another five minutes. Then, the inevitable: Eric, intent on getting his mother's attention, lunges for the printed article Laura has just retrieved. At that moment, Laura toggles the icon on AWE's work surface labeled COVER/UNCOVER to COVER. A thin but sturdy surface originates from below, covering Laura's materials that had been distributed methodically by her across the work surface. The new surface provides space for Laura and Eric to color some pictures for ten minutes – their pact.

After ten minutes of coloring, Eric seems content to entertain himself for a little while longer, and Laura toggles UNCOVER to reveal the organization of her work surface again. Laura decides it's time to rehearse the first segment of her presentation. Simply engaging the PRESENTING icon establishes the new configuration (Fig. 7). (Alternatively, Laura could select INTENT INFERENCING in AWE's set-up menu, so that AWE begins configuring the PRESENTING mode when it recognizes Laura standing in a location where one would give a presentation.) In PRESENTING mode, three of AWE's display screens retract into the ceiling and the fourth screen, displaying Laura's developing *Powerpoint* presentation, spins 180 degrees to face her. Three projectors also are turned on: one projects the *Powerpoint* presentation on the wall surface next to Laura while the other two projectors display, on two adjacent walls, the visual sensation of a small lecture hall where one might make a presentation. AWE's lighting and acoustics are likewise adjusted to simulate the atmosphere of the hall.



Continuum Robots, Figure 7
AWE concept – Presenting

After 30 minutes of rehearsing, Laura remembers its time she drop Eric at the day care and then, to return to her faculty office for a scheduled appointment. Before leaving the house, Laura withdraws, from a USB port in AWE, a portable storage device containing the digital information held in her home-based AWE system.

Stopped at a red traffic light on her way to the University, Laura has an epiphany about her developing proposal. Upon entering her faculty lab where AWE is also installed, Laura inserts the portable storage device into AWE to quickly match the AWE configuration she left at home. Without having to reach for a notepad and without losing the time of transcription, Laura is easily integrating her "epiphany" into the developing presentation. Laura also adjusts AWE's on-screen ATMOSPHERE settings to better approximate, for her next rehearsal, the sense of the conference room where she will finally present her proposal.

At home, Roberto employs AWE to view materials that Laura has given him access to. Among these is a recording of Laura's earlier conversation with Roberto about AWE and her developing work, presented by AWE as a video for Roberto to review and examine. Roberto, a college student, prefers lounging in an upholstered chair to sitting at a desk. Using AWE's simple intelligence, Roberto adjusts AWE's work surface from the default VIEWING configuration to a slightly tilted and lower configuration that better suits his lounging. As well, Roberto elects to bring the single screen displaying Laura's video more proximate to him, guiding the screen with his hand. Roberto saves this VIEWING figuration as, simply, "Roberto" for future retrieval.

When Laura returns from the University, Roberto excitedly announces to his aunt that using AWE, he's discovered a curious link between something she's considering in her research and research activities occurring in his native country, which he discovered on the internet.

When Roberto grows tired from all his explorations facilitated by AWE, Laura, with a simple activation of the COMPOSING option, recovers the physical organization of her materials as before, and continues developing her presentation.

In sum, AWE affords responsiveness to needs – particular and dynamic, cognitive and kinesthetic – of individual users in fluid social organizations and work patterns.

Summary

The two case studies discussed in this article demonstrate the potential for continuum structures to transform the nature of robot applications. The OctArm continuum robots described herein have successfully grasped and manipulated objects over a wide range of sizes and scales in the laboratory, demonstrating the ability to adapt their shape to that of a wide variety of payloads. Additionally, the field tests demonstrated the ability of the continuum robots to operate successfully in realistic application environments, both in air and in water, and to maintain grasp-stability under dynamic disturbances.

AWE, by recognizing the physical environment as an integral aspect of the dynamic interaction between people and the digital realm, constitutes an early, socially significant initiative featuring continuum robots by a team of "collaborative environment designers," defined by Mark Burry as "architects" working "along with their new collaborating experts … in computer science … and engineering" [11].

Future Directions

The AWE environment, featuring continuum robot surfaces, is a "look at the future" for continuum robotics. AWE is a direct extension of the activities within Clemson University's *Animated Architecture Lab*, a research and teaching body founded by the second author in which engineering, architecture and the social sciences converge. The AWE project is currently being realized, as a result of these efforts, as a working prototype at full-scale.

Our ongoing efforts with the OctArm robots focus on enhanced designs incorporating significantly higher arm strength, and on the integration of sensors to guide remote operation in congested obstacle fields. We anticipate "multi-arm" versions of the hardware being developed in the near future.

Acknowledgments

The authors acknowledge support from the US National Science Foundation under grant number IIS-0534423, and from the Defense Advanced Research Projects Agency (DARPA) Defense Sciences Office through the Space and Naval Warfare Systems Center, San Diego, Contract Number N66001-03-C-8043.

Bibliography

1. Aarts E, Marzano S (2003) The New Everyday: Views on Ambient Intelligence. 010 Publishers, Rotterdam
2. Anderson VC, Horn RC (1967) Tensor arm manipulator design. Trans ASME, vol. 67-DE-57. American Society of Mechanical Engineers, New York, pp 1–12
3. Aoki T, Ochiai A, Hirose S (2004) Study on slime robot: development of the mobile robot prototype model using bridle bellows. In: Proc. IEEE International Conference on Robotics and Automation, New Orleans, Louisiana, pp 2808–2813
4. Bailly Y, Amirat Y (2005) Modeling and Control of a Hybrid Continuum Active Catheter for Aortic Aneurysm Treatment. In: Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 936–941
5. Beyer H, Holtzblatt K (1999) Contextual Design: Defining Customer-Centered Systems. Morgan Kaufman Publishers, San Francisco
6. Binder T, De Michelis G, Gervautz M, Jacucci G, Matkovic K, Psik T, Wagner I (2004) Supporting configurability in a mixed-media environment for design students. In: Thomas P (ed) Personal Ubiquitous Computing, vol. 8, Issue 5. Springer, London, pp 310–325
7. Bondarenko O, Janssen R (2005) Documents at Hand: Learning from Paper to Improve Digital Technologies. In: Proceedings of the Computer-Human-Interaction (CHI) 2005 Conference, Association for Computing Machinery. ACM, New York
8. Braganza D, Nath N, Walker ID, Dawson D (2006) Neural Network Grasping Controller for Continuum Robots. Proc. IEEE Conference on Decision and Control, San Diego, CA, pp 6445–6449
9. Braganza D, McIntyre M, Walker ID, Dawson D (2006) Whole Arm Grasping Control for Redundant Robot Manipulators. In: Proc. American Control Conference, Minneapolis, MN, pp 3194–3199
10. Buckingham R (2002) Snake arm robots, Industrial Robot. Int J 29(3):242–245
11. Burry M (2004) Convergent Design. In: Redmond J, Durling D, de Bono A (eds) Futureground Design Research Society International Conference vol 2 Proc. Monash University Press, Victoria
12. Chen G, Tu PM, Herve TR, Prelle C (2005) Design and modeling of a micro-robotic manipulator for colonoscopy. In: Proc. 5th Int Workshop on Research and Education in Mechatronics, Annecy, France, pp 109–114
13. Chrikjian GS (1992) Theory and Applications of Hyper-redundant Robotic Mechanisms. Ph D Thesis, Department of Applied Mechanics, California Institute of Technology. Pasadena
14. Chirikjian GS, Burdick JW (1991) Selected Applications and Intrinsic Kinematics of Hyper-Redundant Robots. In: Proc. 4th American Nuclear Society Topical Meeting on Robotics and Remote Systems, pp 61–70
15. Chirikjian GS, Burdick JW (1993) Design and Experiments with a 30 DOF Robot. In: Proc. IEEE International Conference on Robotics and Automation, Atlanta, pp 113–119
16. Chirikjian GS (1995) Hyper-redundant manipulator dynamics: a continuum approximation. Adv Robot 9(3):217–243
17. Chirikjian GS, Burdick JW (1994) A modal approach to hyper-redundant manipulator kinematics. IEEE Trans Robot Autom 10(3):343–354

18. Chitrakaran V, Behal A, Dawson D, Walker ID (2007) Setpoint Regulation of Continuum Robots Using a Fixed Camera. Robotica 25(5):581–586
19. Cieslak R, Morecki A (1999) Elephant trunk type elastic manipulator – a tool for bulk and liquid type materials transportation. Robotica 17:11–16
20. Csencsits M, Jones BA, McMahan W, Walker ID (2005) User interfaces for continuum robot arms. In: Proc. IEEE/RSJ Int Conf on Intelligent Robots and Systems, Edmonton, Canada, pp 3011–3018
21. deCOI (2004) Aegis Hypersurface. In: Mark Galthorpe/dECOi, "Precise Indeterminacy", Praxis: J Writ Build, Issue 6:28–45
22. Ganz A (2006) Principal Investigator, NeTS: Animated Spaces for the Digital Society: The ASPEN Architecture, NSF Award Abstract – #0434985
23. Garlan D, Siewiorek D, Smailagic A, Steenkiste P (2002) Project Aura: Toward Distraction-Free Pervasive Computing. In: Satyanarayanan M (ed) IEEE Pervasive Computing. IEEE, New York, pp 22–30
24. Gravagne IA, Walker ID (2000) Kinematic Transformations for Remotely Actuated Planar Continuum Robots. In: Proc. IEEE International Conference on Robotics and Automation, San Francisco, CA. IEEE, New York, pp 19–26
25. Gravagne IA, Walker ID (2002) Manipulability, Force and Compliance Analysis for Planar Continuum Manipulators. IEEE Trans Robot Autom 18(3):263–273
26. Gravagne IA, Rahn CD, Walker ID (2003) Large deflection dynamics and control for planar continuum robots. IEEE/ASME Trans Mechatron 8(2):299–307
27. Green KE, Gugerty LJ, Walker ID, Witte JC (2005) AWE (Animated Work Environment): Ambient Intelligence in Working Life. In: Proc. Conference on Intelligent Ambience and Well–Being (Ambience 05), Tampere, Finland, September 2005, pp 1–7 (CD-ROM proceedings)
28. Green KE, Walker ID, Gugerty LJ, Witte JC (2006) Three Robot–Rooms/The AWE Project. In: Proc. CHI 2006, Montreal, Canada, April 2006, pp 809–814
29. Greypilgrim Company http://www.greypilgrim.com. Accessed 2 Aug 1999
30. Hannan MW, Walker ID (2001) Analysis and Experiments with an Elephant's Trunk Robot. Adv Robot 15(8):847–858
31. Hannan MW, Walker ID (2003) Kinematics and the Implementation of an elephant's trunk manipulator and other continuum style robots. J Robot Syst 20(2):45–63
32. Hannan MW, Walker ID (2005) Real-Time Shape Estimation for Continuum Robots Using Vision. Robotica 23(5):645–651
33. Hirose S (1993) Biologically inspired robots. Oxford University Press, New York
34. IBM Blue Space Research (2005) http://www.research.ibm.com/bluespace
35. Ikuta K, Ichikawa H, Suzuki K, Yajima D, Multi-degree of Freedom Hydraulic Pressure Driven Safety Active Catheter. In: Proc. IEEE International Conference on Robotics and Automation, pp 4161–4166
36. Immega G (1992) Tentacle-like manipulators with adjustable tension lines. US Patent #5,317,952
37. Immega G, Antonelli K (1995) The KSI tentacle manipulator. In: Proc. IEEE Intl. Conf. Robotics and Automation, Nagoya, Japan, pp 3149–3154
38. Ivanescu M, Stoian V (1995) A variable structure controller for a tentacle manipulator. In: Proc. IEEE Intl. Conf. Robotics and Automation, Nagoya, Japan, pp 3155–3160
39. Ivanescu M, Bizdoaca N, Pana D (2003) Dynamic control for a tentacle manipulator with SMA actuators. In: Proc. IEEE Intl. Conf. Robotics and Automation, Taipei, Taiwan, pp 2079–2084
40. Ivanescu M, Popescu N, Popescu D (2005) A Variable Length Tentacle Manipulator Control System. In: Proc. IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 3274–3279
41. Jones BA, McMahan W, Walker ID (2004) Design and analysis of a novel pneumatic manipulator. In: Proc. 3rd IFAC Symposium on Mechatronic Systems, Sydney, Australia, pp 745–750
42. Jones BA, Csencsits M, McMahan W, Chitrakaran V, Grissom M, Pritts M, Rahn CD, Walker ID (2006) Grasping, manipulation, and exploration tasks with the OctArm continuum manipulator, video in Proceedings of the International Conference on Robotics and Automation, Orlando, FL
43. Jones BA, Walker ID (2006a) Three-Dimensional Modeling and Display of Continuum Robots. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, China, pp 5872–5877
44. Jones BA, Walker ID (2006b) Kinematics for Multi-Section Continuum Robots. IEEE Trans Robot 22(1):43–55
45. Jones BA, Walker ID (2006c) Practical Kinematics for Real-Time Implementation of Continuum Robots. IEEE Trans Robot 22(6):1087–1099
46. Kier WM, Smith KK (1985) Tongues, tentacles and trunks: The biomechanics of movement in muscular-hydrostats. Zool. J Linn Soc 83:307–324
47. Lane DM, Davies JBC, Robinson G, O'Brien DJ, Sneddon J, Seaton E, Elfstrom E (1999) The AMADEUS dextrous subsea hand: design, modeling, and sensor processing. IEEE J Ocean Eng 24(1):96–111
48. Larson K (2004) House_n Current Projects http://architecture.mit.edu/~kll/Project%20List%20Sept-2004.pdf. Accessed 2 Jun 2008
49. McCarter R (ed) (1987) Building; Machines. Pamphlet Architecture/Princeton Architectural Press, New York
50. McMahan W, Jones BA, Walker ID, Chitrakaran V, Seshadri A, Dawson D (2004) Robotic manipulators inspired by cephalopod limbs. In: Proc. of the CDEN Design Conf., Montreal, Canada, pp 1–10
51. McMahan W, Jones BA, Walker ID (2005) Design and implementation of a multi-section continuum robot: Air-Octor. In: Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, Edmonton, Canada, pp 3345–3352
52. McMahan W, Jones BA, Chitrakaran V, Csencsits M, Grissom M, Pritts M, Rahn CD, Walker ID (2006), Field trials and testing of the OctArm continuum manipulator. In: Proceedings of the International Conference on Robotics and Automation, Orlando, FL, USA, pp 2336–2341
53. Mehling JS, Diftler MA, Chu M, Valvo M (2006) A Minimally Invasive Tendril Robot for In-Space Inspection. In: Proc. BioRob 2006 Conference, pp 690–695
54. MIT Media Lab/FPC (2002), Media House Project: BCN, 26.09–06.10, Barcelona: FPC
55. Mitchell WJ (2000) City of Bits. MIT, Cambridge
56. Mitchell WJ (2000) e-topia. MIT, Cambridge
57. Mitchell WJ (2003) ME++. MIT, Cambridge
58. Mochiyama H, Shimemura E, Kobayashi H (1998) Shape Correspondence between a Spatial Curve and a Manipulator with

Hyper Degrees of Freedom. Proc. IEEE International Conference on Robotics and Automation, pp 161–166

59. Mochiyama H, Suzuki T (2003) Kinematics and dynamics of a cable-like hyper-flexible manipulator. ICRA, Taipei, Taiwan, pp 3672–3677
60. Niranjan RK, Ganz A (2004) Animated Space Architecture for Multimedia Experience – ASkME. In: Proceedings, Broadnets 2004, San Jose, pp 1–6
61. O. C. Robotics http://www.ocrobotics.com/mediagallery/images.htm. Accessed 30 May 2008
62. Ohno H, Hirose S (2001) Design of slim slime robot and its gait of locomotion. In: Proc. IEEE/RSJ Int Conf on Intelligent Robots and Systems, Maui, Hawaii, pp 707–715
63. Oosterhuis K (2003) Hyperbodies: Towards an E-motive architecture. Birkäuser, Basel
64. Pritts MB, Rahn CD (2004) Design of an artificial muscle continuum robot. In: Proc. IEEE Intl. Conf. Robotics and Automation, New Orleans, Louisiana, pp 4742–4746
65. 'Q' Mobile workstation concept for Steelcase (2005) http://www.ideo.com/portfolio/re.asp?x=12378. Accessed 2 Jun 2008
66. Robinson G, Davies JBC (1999) Continuum robots – a state of the art. In: Proc. IEEE Intl. Conf. Robotics and Automation, Detroit, Michigan, pp 2849–2854
67. Salisbury K, Townsend W, Ebrman B, DiPietro D (1988) Preliminary design of a whole-arm manipulation system (WAMS). In: Proc. IEEE Intl. Conf. Robotics and Automation, Philadelphia, Pennsylvania, pp 254–260
68. Sears P, Dupont P (2006) A Steerable Needle Technology Using Curved Concentric Tubes. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, pp 2850–2856
69. Simaan N (2005) Snake-Like Units Using Flexible Backbones and Actuation Redundancy for Enhanced Miniaturization. In: Proc. IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 3023–3028
70. Smith KK, Kier WM (1989) Trunks, tongues, and tentacles: Moving with skeletons of muscle. Am Sci 77:28–35
71. Spong MW, Hutchinson S, Vidyasagar M (2006) Robot modeling and control. John Wiley & Sons, New York
72. Suzumori K, Iikura S, Tanaka H (1991) Development of flexible microactuator and its applications to robotic mechanisms. In: Proc. IEEE Intl. Conf. Robotics and Automation, Sacramento, California, pp 1622–1627
73. Suzumori K, Iikura S, Tanaka H (1992), Applying a flexible microactuator to robotic mechanisms. Control Syst Mag IEEE 12:21–27
74. Tatlicioglu E, Walker ID, Dawson D (2007) Dynamic Modelling for Planar Extensible Continuum Robot Manipulators. In: Proc. IEEE International Conference on Robotics and Automation, Rome, Italy, pp 1357–1362
75. Thomann F, Betemps M, Redarce T (2003) The development of a bendable colonoscopic tip. In: International Conference on Robotics and Automation, Taipei, Taiwan, pp 658–663
76. Tsukagoshi H, Kitagawa A, Segawa M (2001) Active Hose: an artificial elephant's nose with maneuverability for rescue operation. In: Proc. IEEE Intl. Conf. Robotics and Automation, Seoul, Korea, pp 2454–2459
77. Vitruvius (1985) De Architectura [On Architecture]. trans. F. Granger. Harvard University Press, Cambridge
78. Walker ID (2000) Issues in Creating 'Invertebrate' Robots. In: Proc. Conference on Adaptive Manipulation with Animals and Machines, Montreal, Canada, pp 1–6
79. Walker ID, Dawson D, Flash T, Grasso F, Hanlon R, Hochner B, Kier WM, Pagano C, Rahn CD, Zhang Q (2005) Continuum Robot Arms Inspired by Cephalopods. In: Proc. SPIE Conference on Unmanned Ground Vehicle Technology VII, Orlando, FL, pp 303–314
80. Walker ID, Carreras C, McDonnell R, Grimes G (2006) Extension Versus Bending for Continuum Robots. Int J Adv Robot Syst 3(2):171–178
81. Wilson JF, Li D, Chen Z, George RT (1993) Flexible Robot Manipulators and Grippers: Relatives of elephant Trunks and Squid Tentacles. In: Dario P (ed) Robots and Biological Systems: Towards a New Bionics? Nato Asi Series. Springer, New York, pp 474–494
82. Xu K, Simaan N (2006) Actuation Compensation for Flexible Surgical Snake-like Robots with Redundant Remote Actuation. In: Proc. IEEE International Conference on Robotics and Automation, pp 4148–4154
83. Yamada H, Hirose S (2006) Study on the 3D Shape of Active Cord Mechanism. In: Proc. IEEE International Conference on Robotics and Automation, pp 2890–2895
84. Yang J, Potratz J, Abdel-Malek K (2006) A Hyper-Redundant Continuous Robot. In: Proc. IEEE International Conference on Robotics and Automation. pp 1854–1859
85. Yang J, Pena Pitarch E, Potratz J, Beck S, Abdel-Malek K (2006) Synthesis and analysis of a flexible elephant trunk robot. Adv Robot 20(6): 631–659
86. Siciliano B, Khatib O (2008) Springer Handbook of Robotics. Chapter 11, Kinematically Redundant Manipulators. in press

Control of Non-linear Partial Differential Equations

Fatiha Alabau-Boussouira[1],
Piermarco Cannarsa[2]
[1] L.M.A.M., Université de Metz, Metz, France
[2] Dipartimento di Matematica, Università di Roma "Tor Vergata", Rome, Italy

Article Outline

Glossary

$\mathbb{R}$ denotes the **real line**, $\mathbb{R}^n$ the n-dimensional Euclidean space, $x \cdot y$ stands for the Euclidean scalar product of $x, y \in \mathbb{R}^n$, and $|x|$ for the norm of x.

State variables quantities describing the state of a system; in this note they will be denoted by u; in the present setting, u will be either a function defined on a subset of $\mathbb{R} \times \mathbb{R}^n$, or a function of time taking its values in an Hilbert space H.

Space domain the subset of $\mathbb{R}^n$ on which state variables are defined.

Partial differential equation a differential equation containing the unknown function as well as its partial derivatives.

State equation a differential equation describing the evolution of the system of interest.

Control function an external action on the state equation aimed at achieving a specific purpose; in this note, control functions they will be denoted by f; f will be used to denote either a function defined on a subset of $\mathbb{R} \times \mathbb{R}^n$, or a function of time taking its values in an Hilbert space F. If the state equation is a partial differential equation of evolution, then a control function can be:

1. *distributed* if it acts on the whole space domain;
2. *locally distributed* if it acts on a subset of the space domain;
3. *boundary* if it acts on the boundary of the space domain;
4. *optimal* if it minimizes (together with the corresponding trajectory) a given cost;
5. *feedback* if it depends, in turn, on the state of the system.

Trajectory the solution of the state equation u_f that corresponds to a given control function f.

Distributed parameter system a system modeled by an evolution equation on an infinite dimensional space, such as a partial differential equation or a partial integro-differential equation, or a delay equation; unlike systems described by finitely many state variables, such as the ones modeled by ordinary differential equations, the information concerning these systems is "distributed" among infinitely many parameters.

$\mathbb{1}_A$ denotes the **characteristic function** of a set $A \subset \mathbb{R}^n$, that is,

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in \mathbb{R}^n \setminus A \end{cases}$$

∂_t , ∂_{x_i} denote **partial derivatives** with respect to t and x_i, respectively.

$L^2(\Omega)$ denotes the **Lebesgue space** of all real-valued square integrable functions, where functions that differ on sets of zero Lebesgue measure are identified.

$H^1_0(\Omega)$ denotes the **Sobolev space** of all real-valued functions which are square integrable together with their *first order* partial derivatives in the sense of distributions in Ω, and vanish on the boundary of Ω; similarly $H^2(\Omega)$ denotes the space of all functions which are square integrable together with their *second order* partial derivatives.

$H^{-1}(\Omega)$ denotes the dual of $H^1_0(\Omega)$.

$\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional **Hausdorff measure**.

H denotes a **normed spaces** over $\mathbb{R}$ with norm $\|\cdot\|$, as well as an **Hilbert space** with the scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$.

$L^2(0, T; H)$ is the space of all square integrable functions $f\colon [0, T] \to H$; $C([0, T]; H)$ (continuous functions) and $H^1(0, T; H)$ (Sobolev functions) are similarly defined .

Given Hilbert spaces F and H, $\mathcal{L}(F, H)$ denotes the (Banach) space of all bounded linear operators $\Lambda\colon F \to H$ with norm $\|\Lambda\| = \sup_{\|x\|=1} \|\Lambda x\|$ (when $F = H$, we use the abbreviated notation $\mathcal{L}(H)$); $\Lambda^*\colon H \to F$ denotes the adjoint of Λ given by $\langle \Lambda^* u, \phi\rangle = \langle u, \Lambda\phi\rangle$ for all $u \in H, \phi \in F$.

Definition of the Subject

Control theory (abbreviated, CT) is concerned with several ways of influencing the evolution of a given system by an external action. As such, it originated in the nineteenth century, when people started to use mathematics to analyze the perfomance of mechanical systems, even though its roots can be traced back to the calculus of variation, a discipline that is certainly much older. Since the second half of the twentieth century its study was pursued intensively to address problems in aerospace engineering, and then economics and life sciences. At the beginning, CT was applied to systems modeled by ordinary differential equations (abbreviated, ODE). It was a couple of decades after the birth of CT—in the late sixties, early seventies—that the first attempts to control models described by a partial differential equation (abbreviated, PDE) were made. The need for such a passage was unquestionable: too many interesting applications, from diffusion phenomena to elasticity models, from fluid dynamics to traffic flows on networks and systems biology, can be modeled by a PDE.

Because of its peculiar nature, control of PDE's is a rather deep and technical subject: it requires a good knowledge of PDE theory, a field of enormous interest in its own right, as well as familiarity with the basic aspects of CT for ODE's. On the other hand, the effort put into this research direction has been really intensive. Mathematicians and engineers have worked together in the construction of this theory: the results—from the stabilization of flexible structures to the control of turbulent flows—have been absolutely spectacular.

Among those who developed this subject are A. V. Balakrishnan, H. Fattorini, J. L. Lions, and D. L. Russell, but many more have given fundamental contributions.

Introduction

The basic examples of controlled partial differential equations are essentially two: the heat equation and the and the wave equation. In a bounded open domain $\Omega \subset \mathbb{R}^n$ with sufficiently smooth boundary Γ the *heat equation*

$$\partial_t u = \Delta u + f \quad \text{in } Q_T \doteq (0,T)\times\Omega \tag{1}$$

describes the evolution in time of the temperature $u(t,x)$ at any point x of the body Ω. The term $\Delta u = \partial^2_{x_1} u + \cdots + \partial^2_{x_n} u$, called the Laplacian of u, accounts for heat diffusion in Ω, whereas the additive term f represents a heat source. In order to solve the above equation uniquely one needs to add further data, such as the initial distribution u_0 and the temperature of the boundary surface Γ of Ω. The fact that, for any given data $u_0 \in L^2(\Omega)$ and $f \in L^2(Q_T)$ Eq. (1) admits a unique weak solution u_f satisfying the boundary condition

$$u = 0 \quad \text{on} \quad \Sigma_T \doteq (0,T)\times\Gamma \tag{2}$$

and the initial condition

$$u(0,x) = u_0(x) \quad \forall x = (x_1,\dots,x_n) \in \Omega \tag{3}$$

is well-known. So is the maximal regularity result ensuring that

$$u_f \in H^1\left(0,T;L^2(\Omega)\right) \cap C\left([0,T];H^1_0(\Omega)\right) \cap L^2\left(0,T;H^2(\Omega)\right) \tag{4}$$

whenever $u_0 \in H^1_0(\Omega)$. If problem (1)–(3) possesses a unique solution which depends continuously on data, then we say that the problem is *well-posed*.

Similarly, the *wave equation*

$$\partial_t^2 u = \Delta u + f \quad \text{in} \quad Q_T \tag{5}$$

describes the vibration of an elastic membrane (when $n=2$) subject to a force f. Here, $u(t,x)$ denotes the displacement of the membrane at time t in x. The initial condition now concerns both initial displacement and velocity:

$$\forall x \in \Omega \qquad \begin{cases} u(0,x) = u_0(x) \\ \partial_t u(0,x) = u_1(x)\,. \end{cases} \tag{6}$$

It is useful to treat the above problems as a first order *evolution equation* in a Hilbert space H

$$u'(t) = Au(t) + Bf(t) \quad t \in (0,T)\,, \tag{7}$$

where $f(t)$ takes its valued in another Hilbert space F, and $B \in \mathcal{L}(F,H)$. In this abstract set-up, the fact that (7) is related to a PDE translates into that the closed linear operator A is not defined on the whole space but only on a (dense) subspace $D(A) \subset H$, called the *domain* of A; such a property is often referred to as the *unboundedness* of A.

For instance, in the case of the heat equation (1), $H = L^2(\Omega) = F$, and A is defined as

$$\begin{cases} D(A) = H^2(\Omega) \cap H^1_0(\Omega) \\ Au = \Delta u\,, \quad \forall u \in D(A)\,, \end{cases} \tag{8}$$

whereas $B = I$.

As for the wave equation, since it is a second order differential equation with respect to t, the Hilbert space H should be given by the product $H^1_0(\Omega)\times L^2(\Omega)$. Then, problem (5) is turned into the first order equation

$$U'(t) = \mathcal{A}U(t) + Bf(t) \quad t \in (0,T)\,,$$

where

$$U = \begin{pmatrix} u \\ v \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ I \end{pmatrix}, \quad F = L^2(\Omega)\,.$$

Accordingly, $\mathcal{A}\colon D(\mathcal{A}) \subset H \to H$ is given by

$$\begin{cases} D(\mathcal{A}) = \left(H^2(\Omega)\cap H^1_0(\Omega)\right)\times H^1_0(\Omega) \\ \mathcal{A}U = \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix} U = \begin{pmatrix} v \\ Au \end{pmatrix} \quad \forall U \in D(\mathcal{A})\,, \end{cases}$$

where A is taken as in (8).

Another advantage of the abstract formulation (7) is the possibility of considering locally distributed or boundary source terms. For instance, one can reduce to the same set-up the equation

$$\partial_t u = \Delta u + \mathbb{1}_\omega f \quad \text{in} \quad Q_T\,, \tag{9}$$

where $\mathbb{1}_\omega$ denotes the characteristic function of an open set $\omega \subset \Omega$, or the nonhomogeneus boundary condition of Dirichlet type

$$u = f \quad \text{on} \quad \Sigma_T\,, \tag{10}$$

or Neumann type

$$\frac{\partial u}{\partial \nu} = f \quad \text{on} \quad \Sigma_T\,, \tag{11}$$

where ν is the outward unit normal to Γ. For Eq. (9), B reduces to multiplication by $\mathbb{1}_\omega$—a bounded operator on $L^2(\Omega)$; conditions (10) and (11) can also be associated to suitable linear operators B—which, in this case, turn out to be unbounded. Similar considerations can be adapted to the wave equation (5) and to more general problems.

Having an efficient way to represent a source term is essential in control theory, where such a term is regarded as an external action, the *control function*, exercised on the *state variable* u for a purpose, of which there are two main kinds:

- *positional*: $u(t)$ is to approach a given target in X, or attain it exactly at a given time $t > 0$;
- *optimal*: the pair (u, f) is to minimize a given functional.

The first criterion leads to *approximate* or *exact controllability* problems in time t, as well as to *stabilization* problems as $t \to \infty$. Here, the main tools will be provided by certain estimates for partial differential operators that allow to study the states that can be attained by the solution of a given controlled equation. These issues will be addressed in Sects. "Controllability" and "Stabilization" for linear evolution equations. Applications to the heat and wave equations will be discussed in the same sections.

On the other hand, *optimal control problems* require analyzing the typical issues of optimizations: existence results, necessary conditions for optimality, sufficient conditions, robustness. Here, the typical problem that has been successfully studied is the Linear Quadratic Regulator that will be discussed in Sect. "Linear Quadratic Optimal Control".

Control problems for nonlinear partial differential equations are extremely interesting but harder to deal with, so the literature is less rich in results and techniques. Nevertheless, among the problems that received great attention are those of fluid dynamics, specifically the *Euler equations*

$$\partial_t u + (u \cdot \nabla)u + \nabla p = 0$$

and the *Navier–Stokes equations*

$$\partial_t u - \mu\Delta u + (u \cdot \nabla)u + \nabla p = 0$$

subject to a *boundary control* and to the incompressibility condition $\operatorname{div} u = 0$.

Controllability

We now proceed to introduce the main notions of controllability for the evolution equation (7). Later on in this section we will give interpretations for the heat and wave equations.

In a given Hilbert space H, with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, let

$$A \colon D(A) \subset H \to H$$

be the *infinitesimal generator* of a *strongly continuous semigroup* e^{tA}, $t \geq 0$, of bounded linear operators on X. Intuitively, this amounts to saying that $u(t) \doteq e^{tA} u_0$ is the unique solution of the Cauchy problem

$$\begin{cases} u'(t) = Au(t) & t \geq 0 \\ u(0) = u_0 \,, \end{cases}$$

in the classical sense for $u_0 \in D(A)$, and in a suitable generalized sense for all $u_0 \in H$. Necessary and sufficient conditions in order for an unbounded operator A to be the infinitesimal generator of a strongly continuous semigroup are given by the celebrated Hille–Yosida Theorem, see, e. g. [99] and [55].

Abstract Evolution Equations

Let F be another Hilbert space (with scalar product and norm denoted by the same symbols as for H), the so-called *control space*, and let $B \colon F \to H$ be a linear operator, that we will assume to be bounded for the time being. Then, given $T > 0$ and $u_0 \in H$, for all $f \in L^2(0, T; F)$ the Cauchy problem

$$\begin{cases} u'(t) = Au(t) + Bf(t) & t \geq 0 \\ u(0) = u_0 \end{cases} \tag{12}$$

has a unique *mild solution* $u_f \in C([0, T]; H)$ given by

$$u_f(t) = e^{tA} u_0 + \int_0^t e^{(t-s)A} Bf(s) \quad \forall t \geq 0 \tag{13}$$

Note 1 Boundary control problems can be reduced to the same abstract form as above. In this case, however, B in (12) turns out to be an unbounded operator related to suitable fractional powers of $-A$, see, e. g., [22].

For any $t \geq 0$ let us denote by $\Lambda_t \colon L^2(0, t; F) \to H$ the bounded linear operator

$$\Lambda_t f = \int_0^t e^{(t-s)A} Bf(s)\, ds \quad \forall f \in L^2(0, t; F)\,. \tag{14}$$

The *attainable* (or *reachable*) set from u_0 at time t, $\mathcal{A}(u_0, t)$ is the set of all points in H of the form $u_f(t)$ for some control function f, that is

$$\mathcal{A}(u_0, t) \doteq e^{tA} u_0 + \Lambda_t L^2(0, t; F)\,.$$

We introduce below the main notions of controllability for (7). Let $T > 0$.

Definition 1 System (7) is said to be:

- **exactly controllable** in time T if $\mathcal{A}(u_0, T) = H$ for all $u_0 \in H$, that is, if for all $u_0, u_1 \in H$ there is a control function $f \in L^2(0, T; F)$ such that $u_f(T) = u_1$;
- **null controllable** in time T if $0 \in \mathcal{A}(u_0, T)$ for all $u_0 \in H$, that is, if for all $u_0 \in H$ there is a control function $f \in L^2(0, T; F)$ such that $u_f(T) = 0$;
- **approximately controllable** in time T if $\mathcal{A}(u_0, T)$ is dense in H for all $u_0 \in H$, that is, if for all $u_0, u_1 \in H$ and for any $\varepsilon > 0$ there is a control function $f \in L^2(0, T; F)$ such that $\|u_f(T) - u_1\| < \varepsilon$.

Clearly, if a system is exactly controllable in time T, then it is also null and approximately controllable in time T. Although these last two notions of controllability are strictly weaker than strong controllability, for specific problems—like when A generates a strongly continuous group—some of them may coincide.

Since controllability properties concern, ultimately, the range of the linear operator Λ_T defined in (14), it is not surprising that they can be characterized in terms of the adjoint operator $\Lambda_T^*\colon H \to L^2(0, T; F)$, which is defined by

$$\int_0^T \langle \Lambda_T^* u(s), f(s)\rangle \mathrm{d}s = \langle u_0, \Lambda_T f\rangle \qquad \forall u \in H\,,\ \forall f \in L^2(0, T; F)\,.$$

Such a characterization is the object of the following theorem. Notice that the above identity and (14) yield

$$\Lambda_T^* u(s) = B^* \mathrm{e}^{(T-s)A^*} u \quad \forall s \in [0, T]\,.$$

Theorem 1 *System (7) is:*

- ***exactly controllable*** *in time T if and only if there is a constant $C > 0$ such that*
$$\int_0^T \left\|B^* e^{tA^*} u\right\|^2 dt \geq C\|u\|^2 \quad \forall u \in H\,; \tag{15}$$
- ***null controllable*** *in time T if and only if there is a constant $C > 0$ such that*
$$\int_0^T \left\|B^* e^{tA^*} u\right\|^2 dt \geq C\left\|e^{TA^*} u\right\|^2 \quad \forall u \in H\,; \tag{16}$$
- ***approximately controllable*** *in time T if and only if, for every $u \in H$,*
$$B^* e^{tA^*} u = 0 \quad t \in [0, T] \ a.e. \implies u = 0\,. \tag{17}$$

To benefit the reader who is more familiar with optimization theory than abstract functional analysis, let us explain, by a variational argument, why estimate (16) implies null controllability. Consider, for every $\varepsilon > 0$, the penalized problem

$$\min\left\{J_\varepsilon(f) \,:\, f \in L^2(0, T; H)\right\},$$

where

$$J_\varepsilon(f) = \frac{1}{2}\int_0^T \|f(t)\|^2 \,\mathrm{d}t + \frac{1}{2\varepsilon}\|u_f(T)\|^2 \qquad \forall f \in L^2(0, T; H)\,.$$

Since J_ε is strictly convex, it admits a unique minimum point f_ε. Set $u_\varepsilon = u_{f_\varepsilon}$. Recalling (13) we have, By Fermat's rule,

$$0 = J'_\varepsilon(f_\varepsilon)g = \int_0^T \langle f_\varepsilon(t), g(t)\rangle \,\mathrm{d}t + \frac{1}{\varepsilon}\langle u_\varepsilon(T), \Lambda_T g\rangle \quad \forall g \in L^2(0, T; H)\,. \tag{18}$$

Therefore, passing to the adjoint of Λ_T,

$$\int_0^T \left\langle f_\varepsilon(t) + \frac{1}{\varepsilon}\left(\Lambda_T^* u_\varepsilon(T)\right)(t), g(t)\right\rangle \mathrm{d}t = 0 \qquad \forall g \in L^2(0, T; H)\,,$$

whence, owing to (14),

$$f_\varepsilon(t) = -\frac{1}{\varepsilon}\left(\Lambda_T^* u_\varepsilon(T)\right)(t) = -B^* v_\varepsilon(t) \qquad \forall t \in [0, T]\,, \tag{19}$$

where $v_\varepsilon(t) \doteq \frac{1}{\varepsilon}\mathrm{e}^{(T-t)A^*} u_\varepsilon(T)$ is the solution of the *dual problem*

$$\begin{cases} v' + A^* v = 0 & t \in [0, T] \\ v(T) = \frac{1}{\varepsilon} u_\varepsilon(T)\,. \end{cases}$$

It turns out that

$$\frac{1}{2}\int_0^T \|f_\varepsilon(t)\|^2 \,\mathrm{d}t + \frac{1}{\varepsilon}\|u_\varepsilon(T)\|^2 \leq C\|u_0\|^2 \qquad \forall \varepsilon > 0 \tag{20}$$

for some positive constant C. Indeed, observe that, in view of (19),

$$\begin{cases} \left\langle u'_\varepsilon - Au_\varepsilon + BB^* v_\varepsilon\,,\, v_\varepsilon\right\rangle = 0\,, & u_\varepsilon(0) = u_0 \\ \left\langle v'_\varepsilon + A^* v_\varepsilon\,,\, u_\varepsilon\right\rangle = 0\,, & v_\varepsilon(T) = \frac{1}{\varepsilon} u_\varepsilon(T)\,. \end{cases}$$

So,

$$\int_0^T \left[\frac{\mathrm{d}}{\mathrm{d}t} \langle u_\varepsilon \, , \, v_\varepsilon \rangle + \left\| B^* v_\varepsilon \right\|^2 \right] dt = 0 \, ,$$

hence

$$\frac{1}{\varepsilon} \| u_\varepsilon(T) \|^2 + \int_0^T \left\| B^* v_\varepsilon \right\|^2 \, \mathrm{d}t \; = \; \langle u_0 \, , \, v_\varepsilon(0) \rangle \, . \qquad (21)$$

Now, apply estimate (16) with $u = \frac{u_\varepsilon(T)}{\varepsilon}$ and note that $v_\varepsilon(T - t) = \mathrm{e}^{tA^*} \frac{u_\varepsilon(T)}{\varepsilon}$ to obtain

$$\int_0^T \left\| B^* v_\varepsilon(t) \right\|^2 \, \mathrm{d}t \geq C \, \| v_\varepsilon(0) \|^2$$

for some positive constant C. Hence, (20) follows from (21) and (19).

Finally, from (20) one deduces the existence of a weakly convergent subsequence f_{ε_j} in $L^2(0, T; F)$. Then, called f_0 the weak limit of f_{ε_j}, $u_{\varepsilon_j}(t) \to u_{f_0}(t)$ for all $t \in [0, T]$. So, owing to (20), $u_{f_0}(T) = 0$.

Heat Equation

It is not hard to see that the heat equation (9) with Dirichlet boundary conditions (2) fails to be exactly controllable. On the other hand, one can show that it is null controllable in any time $T > 0$, hence approximately controllable. Let ω be an open subset of Ω such that $\overline{\omega} \subset \Omega$.

Taking

$$H = L^2(\Omega) = F \, , \quad Bf = \mathbb{1}_\omega f \quad \forall f \in L^2(\Omega)$$

and A as in (8), one obtains that, for any $u_0 \in L^2(\Omega)$ and $f \in L^2(Q_T)$, the initial-boundary value problem

$$\begin{cases} \partial_t u = \Delta u + \mathbb{1}_\omega f & \text{in } \; Q_T \\ u = 0 & \text{on } \; \Sigma_T \\ u(0, x) = u_0(x) & x \in \Omega \end{cases} \qquad (22)$$

has a unique mild solution $u_f \in C([0, T]; L^2(\Omega))$. Moreover, multiplying both sides of equation (9) by u and integrating by parts, it is easy to see that

$$\partial_{x_i} u \in L^2(Q_T) \quad \forall i = 1, \ldots, n \, . \qquad (23)$$

Notice that the above property already suffices to explain why the heat equation cannot be exactly controllable: it is impossible to attain a state $u_1 \in L^2(\Omega)$ which is not compatible with (23).

On the other hand, null controllability holds true in any positive time.

Theorem 2 *Let $T > 0$ and let ω be an open subset of Ω such that $\overline{\omega} \subset \Omega$. Then the heat equation (9) with homogeneous Dirichlet boundary conditions is null controllable in time T, i. e., for every initial condition $u_0 \in L^2(\Omega)$ there is a control function $f \in L^2(Q_T)$ such that the solution u_f of (22) satisfies $u_f(T, \cdot) \equiv 0$. Moreover,*

$$\iint_{Q_T} |f|^2 dx dt \leq C_T \int_\Omega |u_0|^2 dx$$

for some positive constant C_T.

The above property is a consequence of the abstract result in Theorem 1 and of concrete estimates for solutions of parabolic equations. Indeed, in order to apply Theorem 1 one has to translate (16) into an estimate for the heat operator. Now, observing that both A and B are self-adjoint, one promptly realizes that (16) reduces to

$$\int_0^T \int_\omega |v(t, x)|^2 \mathrm{d}x\mathrm{d}t \geq C \int_\Omega |v(T, x)|^2 \mathrm{d}x \qquad (24)$$

for every solution v of the problem

$$\begin{cases} \partial_t v = \Delta v & \text{in } \; Q_T \\ v = 0 & \text{on } \; \Sigma_T \, . \end{cases} \qquad (25)$$

Estimate (24) is called an *observability inequality* for the heat operator for obvious reasons: problem (25) is not well-posed since the initial condition is missing. Nevertheless, if, "observing" a solution v of such a problem on the "small" cylinder $(0, T) \times \omega$, you find that it vanishes, then you can conclude that $v(T, \cdot) \equiv 0$ in the whole domain Ω. Thus, $v(0, \cdot) \equiv 0$ by backward uniqueness.

In conclusion, as elegant as the abstract approach to null controllability may be, one is confronted by the difficult task of proving observability estimates. In fact, for the heat operator there are several ways to prove inequality (24). One of the most powerful, basically due to Fursikov and Imanuvilov [65], relies on global *Carleman estimates*. Compared to other methods that can be used to derive observability, such a technique has the advantage of applying to second order parabolic operators with variable coefficients, as well as to more general operators.

Global Carleman estimates are a priori estimates in weighted norms for solutions of the problem

$$\begin{cases} \partial_t v = \Delta v + f & \text{in } \; Q_T \\ v = 0 & \text{on } \; \Sigma_T \, . \end{cases} \qquad (26)$$

regardless of initial conditions. The weight function is usually of the form

$$\psi_r(t, x) \doteq \theta(t) \big(\mathrm{e}^{2r\|\phi\|_{\infty,\Omega}} - \mathrm{e}^{r\phi(x)} \big) \quad (t, x) \in Q_T \, , \; (27)$$

where r is a positive constant, ϕ is a given function in $C^2(\overline{\Omega})$ such that

$$\nabla\phi(x) \neq 0 \quad \forall x \in \overline{\Omega}\,, \tag{28}$$

and

$$\theta(t) \doteq \frac{1}{t(T-t)} \quad 0 < t < T\,.$$

Note that

$$\theta > 0\,, \quad \theta(t) \to \infty \quad t \to 0\,, T$$

$$\psi_r > 0\,, \quad \psi_r(t,x) \to \infty \quad t \downarrow 0\,, t \uparrow T\,.$$

Using the above notations, a typical global Carleman estimate for the heat operator is the following result obtained in [65]. Let us denote by $\nu(x)$ the outword unit normal to Γ at a point $x \in \Gamma$, and by

$$\frac{\partial\phi}{\partial\nu}(x) = \nabla\phi(x)\cdot\nu(x)$$

the normal derivative of ϕ at x.

Theorem 3 *Let Ω be a bounded domain of $\mathbb{R}^n$ with boundary of class C^2, let $f \in L^2(Q_T)$, and let ϕ be a function satisfying (28). Let v be a solution of (26). Then there are positive constants r, s_0 and C such that, for any $s > s_0$,*

$$\begin{aligned} s^3 \iint_{Q_T} & \theta^3(t)|v(t,x)|^2 e^{-2s\psi_r} dx dt \\ & \leq C \iint_{Q_T} |f(t,x)|^2 e^{-2s\psi_r} dx dt \\ & + Cs \int_0^T \theta(t) dt \\ & \quad \times \int_\Gamma \frac{\partial\phi}{\partial\nu}(x) \left|\frac{\partial v}{\partial\nu}(t,x)\right|^2 e^{-2s\psi_r} \mathrm{d}\mathcal{H}^{n-1}(x) \end{aligned} \tag{29}$$

It is worth underlying that, thanks to the singular behavior of θ near 0 and T, the above result is independent of the initial value of v. Therefore, it can be applied, indifferently, to any solution of (26) as well as to any solution of the *backward problem*

$$\begin{cases} \partial_t v + \Delta v = f & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T\,. \end{cases}$$

Moreover, inequality (29) can be completed adding first and second order terms to its left-hand side, each with its own adapted power of s and θ.

Instead of trying to sketch the proof of Theorem 3, which would go beyond the scopes of this note, it is interesting to explain how it can be used to recover the observability inequality (24), which is what is needed to show that the heat equation is null controllable. The reasoning—not completely straightforward—is based on the following topological lemma, proved in [65].

Lemma 1 *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with boundary Γ of class C^k, for some $k \geq 2$, and let $\omega \subset \Omega$ be an open set such that $\overline{\omega} \subset \Omega$.*

Then there is function $\phi \in C^k(\overline{\Omega})$ such that

$$\begin{cases} (i) \quad \phi(x) = 0 \quad \text{and} \quad \frac{\partial\phi}{\partial\nu}(x) < 0 \quad \forall x \in \Gamma \\ (ii) \quad \{x \in \Omega | \nabla\phi(x) = 0\} \subset \omega\,. \end{cases} \tag{30}$$

Now, given a solution v of (25) and an open set ω such that $\overline{\omega} \subset \Omega$, let $\omega' \subset\subset \omega'' \subset\subset \omega$ be subdomains with smooth boundary. Then the above lemma ensures the existence of a function ϕ such that

$$\{x \in \Omega | \nabla\phi(x) = 0\} \subset \omega'\,.$$

"Localizing" problem (25) onto $\Omega' \doteq \Omega \setminus \omega'$ by a cutoff function $\eta \in C^\infty(\mathbb{R}^n)$ such that

$$0 \leq \eta \leq 1\,, \quad \eta \equiv 1 \quad \text{on } \mathbb{R}^n \setminus \omega''\,, \quad \eta \equiv 0 \quad \text{on } \omega'\,,$$

that is, taking $w = \eta v$, gives

$$\begin{cases} \partial_t w = \Delta w + h & \text{in } Q'_T \doteq (0,T)\times\Omega' \\ w(t,\cdot) = 0 & \text{on } \partial\Omega' = \partial\Omega \cup \partial\omega'\,, \end{cases} \tag{31}$$

with $h := -v\Delta\eta + 2\nabla\eta\cdot\nabla u$. Since $\nabla\phi \neq 0$ on Ω', Theorem 3 can be applied to w on Q'_T to obtain

$$\begin{aligned} s^3 \iint_{Q'_T} & \theta^3 |w|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \\ & \leq C \iint_{Q'_T} |h|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \\ & + Cs \int_0^T \theta \mathrm{d}t \int_\Gamma \frac{\partial\phi}{\partial\nu} \left|\frac{\partial w}{\partial\nu}\right|^2 \mathrm{e}^{-2s\psi_r} d\mathcal{H}^{n-1} \\ & + Cs \int_0^T \theta \mathrm{d}t \int_{\partial\omega'} \frac{\partial\phi}{\partial\nu} \left|\frac{\partial w}{\partial\nu}\right|^2 \mathrm{e}^{-2s\psi_r} d\mathcal{H}^{n-1} \\ & \leq C \iint_{Q'_T} |h|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \end{aligned}$$

for s sufficiently large. On the other hand, for any $0 < T_0 < T_1 < T$,

$$s^3 \iint_{Q'_T} \theta^3 |w|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \geq s^3 \int_{T_0}^{T_1} \mathrm{d}t \int_{\Omega\backslash\omega} \theta^3 |w|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \geq \int_{T_0}^{T_1} \mathrm{d}t \int_{\Omega\backslash\omega} |v|^2 \mathrm{d}x$$

Therefore, recalling the definition of h,

$$\begin{aligned} \int_{T_0}^{T_1} \mathrm{d}t \int_{\Omega\backslash\omega} |v|^2 \mathrm{d}x &\leq C \iint_{Q'_T} |h|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \\ &\leq C \int_0^T \mathrm{d}t \int_{\omega''\backslash\omega'} \left[|\nabla^2 \eta|^2 v^2 + |\nabla \eta|^2 |\nabla v|^2 \right] \mathrm{e}^{-2s\psi_r} \mathrm{d}x \\ &\leq C \int_0^T \mathrm{d}t \int_\omega |v|^2 \mathrm{d}x + C \int_0^T \mathrm{d}t \\ &\quad \times \int_{\omega''\backslash\omega'} |\nabla v|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x \,. \end{aligned}$$

Now, fix $T_0 = T/3\,,\ T_1 = 2T/3$ and use Caccioppoli's inequality (a well-known estimate for solution of elliptic and parabolic PDE's)

$$\int_0^T \mathrm{d}t \int_{\omega''\backslash\omega'} |\nabla v|^2 \mathrm{e}^{-2s\psi_r}\, \mathrm{d}x \leq C \int_0^T \mathrm{d}t \int_\omega |v|^2 \mathrm{e}^{-2s\psi_r}\, \mathrm{d}x \,,$$

to conclude that

$$\int_{T/3}^{2T/3} \mathrm{d}t \int_{\Omega\backslash\omega} |v|^2\, \mathrm{d}x \leq C \int_0^T \mathrm{d}t \int_\omega |v|^2\, \mathrm{d}x$$

or

$$\int_{T/3}^{2T/3} \mathrm{d}t \int_\Omega |v|^2\, \mathrm{d}x \leq (1 + C) \int_0^T \mathrm{d}t \int_\omega |v|^2\, \mathrm{d}x$$

for some constant C. Then, the dissipativity of the heat operator (that is, the fact that $\int_\Omega |v(t,x)|^2 \mathrm{d}x$ is decreasing with respect to t) implies that

$$\begin{aligned} \int_\Omega v^2(T,x)\, \mathrm{d}x &\leq \frac{3}{T} \int_{T/3}^{2T/3} \mathrm{d}t \int_\Omega v^2(t,x)\, \mathrm{d}x \\ &\leq (1 + C) \frac{3}{T} \int_0^T \mathrm{d}t \int_\omega v^2(t,x)\, \mathrm{d}x \,, \end{aligned}$$

which is exactly (24).

Wave Equation

Compared to the heat equation, the wave equation (5) exhibits a quite different behavior from the point of view of exact controllability. Indeed, on the one hand, there is no obstruction to exact controllability since no regularizing effect is connected with wave propagation. On the other hand, due to the finite speed of propagation, exact controllability cannot be expected to hold true in arbitrary time, as null controllability does for the heat equation.

In fact, a typical result that holds true for the wave equation is the following, where a boundary control of Dirichlet type acts on a part $\Gamma_1 \subset \Gamma$, while homogeneous boundary conditions are imposed on $\Gamma_0 = \Gamma \setminus \Gamma_1$:

$$\begin{cases} \partial_t^2 u = \Delta u & \text{in } Q_T \\ u = f \mathbb{1}_{\Gamma_1} & \text{on } \Sigma_T \\ u(0,x) = u_0(x)\,, \partial_t u(0,x) = u_1(x) & x \in \Omega \end{cases} \qquad (32)$$

Observe that problem (32) is well-posed taking

$$\begin{aligned} u_0 &\in L^2(\Omega)\,, \quad u_1 \in H^{-1}(\Omega) \\ f &\in L^2(0,T; L^2(\Gamma)) \\ u &\in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))\,. \end{aligned}$$

Theorem 4 *Let Ω be a bounded domain of $\mathbb{R}^n$ with boundary of class C^2 and suppose that, for some point $x_0 \in \mathbb{R}^n$,*

$$\begin{cases} (x - x_0) \cdot \nu(x) > 0 & \forall x \in \Gamma_1 \\ (x - x_0) \cdot \nu(x) \leq 0 & \forall x \in \Gamma_0\,. \end{cases}$$

Let

$$R = \sup_{x \in \Omega} |x - x_0|\,.$$

If $T > 2R$, then, for all $(u_0, u_1), (v_0, v_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ there is a control function $f \in L^2(0,T; L^2(\Gamma))$ such that the solution u_f of (32) satisfies

$$u_f(T,x) = v_0(x)\,, \quad \partial_t u_f(T,x) = v_1(x)\,.$$

As we saw for abstract evolution equations, the above exact controllability property is proved to be equivalent to an observability estimate for the dual homogeneous problem using, for instance, the Hilbert Uniqueness Method (HUM) by J.-L. Lions [86].

Bibliographical Comments

The literature on controllability of parabolic equations and related topics is so huge, that no attempt to provide a comprehensive account of it would fit within the scopes of this note. So, the following comments have to be taken as a first hint for the interested reader to pursue further bibliographical research.

The theory of exact controllability for parabolic equations was initiated by the seminal paper [58] by Fattorini and Russell. Since then, it has experienced an enormous development. Similarly, the multiplier method to obtain observability inequalities for the wave equation was developed in [17,73,74,77,86]. Some fundamental early contributions were surveyed by Russell [108]. The next essential progress was made in the work by Lebeau and Robbiano [83] and then by Fursikov and Imanuvilov in a series of papers. In [65] one can find an introduction to global Carleman estimates, as well as applications to the controllability of several ODE's. In particular, the presentation of this paper as for observability inequalities and Carleman estimates for the heat operator is inspired by the last monograph. General perspectives for the understanding of global Carleman estimates and their applications to unique continuation and control problems for PDE's can be found in the works by Tataru [113,114,115,116]. Usually, the above approach requires coefficients to be sufficiently smooth. Recently, however, interesting adaptations of Carleman estimates to parabolic operators with discontinuous coefficients have been obtained in [21,82].

More recently, interest has focussed on control problems for nonlinear parabolic equations. Different approaches to controllability problems have been proposed in [57] and [44]. Then, null and approximate controllability results have been improved by Fernandez–Cara and Zuazua [61,62]. Techniques to produce insensitizing controls have been developed in [117]. These techniques have been successfully applied to the study of Navier–Stokes equations by several authors, see e. g. [63].

Fortunately, several excellent monographs are now available to help introduce the reader to this subject. For instance, the monograph by Zabczyk [121] could serve as a clean introduction to control and stabilization for finite- and infinite-dimensional systems. Moreover, [22,50,51], as well as [80,81] develop all the basic concepts of control and system theory for distributed parameter systems with special emphasis on abstract formulation. Specific references for the controllability of the wave equation by HUM can be found in [86] and [74]. More recent results related to series expansion and Ingham type methods can be found in [75]. For the control of Navier–Stokes equations the reader is referred to [64], as well as to the book by Coron [43], which contains an extremely rich collection of classical results and modern developments.

Stabilization

Stabilization of flexible structures such as beams, plates, up to antennas of satellites, or of fluids as, for instance, in aeronautics, is an important part of CT. In this approach, one wants either to derive feedback laws that will allow the system to autoregulate once they are implemented, or study the asymptotic behavior of the stabilized system i. e. determine whether convergence toward equilibrium states as times goes to infinity holds, determine its speed of convergence if necessary or study how many feedback controls are required in case of coupled systems.

Different mathematical tools have been introduced to handle such questions in the context of ODE's and then of PDE's. Stabilization of ODE's goes back to the work of Lyapunov and Lasalle. The important property is that trajectories decay along Lyapunov functions. If trajectories are relatively compact in appropriate spaces and the system is autonomous, then one can prove that trajectories converge to equilibria asymptotically. However, the construction of Lyapunov functions is not easy, in general.

This section will be concerned with some aspects of the stabilization of second order hyperbolic equations, our model problem being represented by the wave equation with distributed damping

$$\begin{cases} \partial_{tt}u - \Delta u + a(x)u_t = 0 & \text{in } \Omega \times \mathbb{R}\,, \\ u = 0 & \text{on } \Sigma = (0,\infty) \times \Gamma \\ (u, \partial_t u)(0) = (u^0, u^1) & \text{on } \Omega\,, \end{cases} \tag{33}$$

in a bounded domain $\Omega \subset \mathbb{R}^n$ with a smooth boundary Γ. For $n = 2$, $u(t,x)$ represents the displacement of point x of the membrane at time t. Therefore, equation (33) describes an elastic system. The energy of such a system is given by

$$E(t) = \frac{1}{2} \int_\Omega \left[|u_t(t,x)|^2 + |\nabla u(t,x)|^2 \right] \mathrm{d}x\,.$$

When $a \geq 0$, the *feedback* term $a(x)u_t$ models friction: it produces a loss of energy through a dissipation phenomenon. More precisely, multiplying the equation in (33) by u_t and integrating by parts on Ω, it follows that

$$E'(t) = -\int_\Omega a(x)|u_t|^2 \, \mathrm{d}x \leq 0\,, \quad \forall t \geq 0\,. \tag{34}$$

On the other hand, if $a \equiv 0$, then the system is *conservative*, i. e., $E(t) = E(0)$ for all $t \geq 0$.

Another well-investigated stabilization problem for the wave equation is when the feedback is localized on a part Γ_0 of the boundary Γ, that is,

$$\begin{cases} \partial_{tt}u - \Delta u = 0 & \text{in } \Omega \times \mathbb{R} \\ \frac{\partial u}{\partial \nu} + u_t = 0 & \text{on } \Sigma_0 = (0,\infty) \times \Gamma_0 \\ u = 0 & \text{on } \Sigma_1 = (0,\infty) \times (\Gamma \setminus \Gamma_0) \\ (u, \partial_t u)(0) = (u^0, u^1) \end{cases} \tag{35}$$

In this case, the dissipation relation (34) takes the form

$$E'(t) = -\int_{\Gamma_0} |u_t|^2 \, \mathrm{d}\mathcal{H}^{n-1} \leq 0\,, \quad \forall t \geq 0\,.$$

In many a situation—such as to improve the quality of an acoustic hall—one seeks to reduce vibrations to a minimum: this is why stabilization is an important issue in CT. We note that the above system has a unique stationary solution—or, *equilibrium*—given by $u \equiv 0$. Stabilization theory studies all questions related to the convergence of solutions to such an equilibrium: existence of the limit, rate of convergence, different effects of nonlinearities in both displacement and velocity, effects of geometry, coupled systems, damping effects due to memory in viscoelastic materials, and so on.

System (33) is said to be:

- **strongly stable** if $E(t) \to 0$ as $t \to \infty$;
- **(uniformly) exponentially stable** if $E(t) \leq Ce^{-\alpha t}E(0)$ for all $t \geq 0$ and some constants $\alpha > 0$ and $C \geq 0$, independent of u^0, u^1.

This note will focus on some of the above issues, such as geometrical aspects, nonlinear damping, indirect damping for coupled systems and memory damping.

Geometrical Aspects

A well-known property of the wave equation is the so-called *finite speed of propagation*, which means that, if the initial conditions u^0, u^1 have compact support, then the support of $u(t,\cdot)$ evolves in time at a finite speed. This explains why, for the wave equation, the geometry of Ω plays an essential role in all the issues related to control and stabilization.

The size and localization of the region in which the feedback is active is of great importance. In this paper such a region, denoted by ω, is taken as a subset of Ω of positive Lebesgue measure. More precisely, a is assumed to be continuous on $\overline{\Omega}$ and such that

$$a \geq 0 \quad \text{on} \quad \Omega \quad \text{and} \quad a \geq a_0 \quad \text{on} \quad \omega\,, \tag{36}$$

for some constant $a_0 > 0$. In this case, the feedback is said to be *distributed*. Moreover, it is said to be *globally* distributed if $\omega = \Omega$ and *locally* distributed if $\Omega \setminus \omega$ has positive Lebesgue measure.

Two main methods have been used or developed to study stabilization, namely the *multiplier method* and *microlocal analysis*. The one that gives the sharpest results is based on microlocal analysis. It goes back to the work of Bardos, Lebeau and Rauch [17], giving geodesics sufficient conditions on the region of active control for exact controllability to hold. These conditions say that each ray of geometric optics should meet the control region. Burq and Gérard [25] showed that these results hold under weaker regularity assumptions on the domain and coefficients of the operators (see also [26,27]). These geodesics conditions are not explicit, in general, but they allow to get decay estimates of the energy under very general hypotheses.

The multiplier method is an explicit method, based on energy estimates, to derive decay rates (as well as observability and exact controllability results). For boundary control and stabilization problems it was developed in the works of several authors, such as Ho [38,73], J.-L. Lions [86], Lasiecka–Triggiani, Komornik–Zuazua [76], and many others. Zuazua [123] gave an explicit geometric condition on ω for a semilinear wave equation subject to a locally distributed damping. Such a condition was then relaxed K. Liu [87] (see also [93]) who introduced the so-called piecewise multiplier method. Lasiecka and Triggiani [80,81] introduced a sharp trace regularity method which allows to estimate boundary terms in energy estimates. There also exist intermediate results between the geodesics conditions of Bardos–Lebeau–Rauch and the multiplier method, obtained by Miller [95] using differentiable escape functions.

Zuazua's multiplier geometric condition can be described as follows. If a subset O of $\overline{\Omega}$ is given, one can define an ε-neighborhood of O in $\overline{\Omega}$ as the subset of points of Ω which are at distance at most ε of O. Zuazua proved that if the set ω is such that there exists a point $x_0 \in \mathbb{R}^n$—an *observation* point—for which ω contains an ε-neighborhood of $\Gamma(x^0) = \{x \in \partial\Omega\,,\ (x - x^0) \cdot \nu(x) \geq 0\}$, then the energy decays exponentially. In this note, we refer to this condition as (MGC).

If a vanishes for instance in a neighborhood of the two poles of a ball Ω in $\mathbb{R}^n$, one cannot find an observation point x_0 such that (MGC) holds. K. Liu [87] (see also [93]) introduced a piecewise multiplier method which allows to choose several observation points, and therefore to handle the above case.

Introduce disjoint lipschitzian domains Ω_j of Ω, $j = 1, \dots, J$, and observation points $x^j \in \mathbb{R}^N$, $j = 1, \dots, J$ and define

$$\gamma_j(x^j) = \{x \in \partial\Omega_j \, , \, (x - x^j) \cdot \nu_j(x) \geq 0\}$$

Here ν_j stands for the unit outward normal vector to the boundary of Ω_j. Then the piecewise multiplier geometrical condition for ω is:

$$\omega \supset \mathcal{N}_\varepsilon \left(\cup_{j=1}^J \gamma_j(x^j) \cup \left(\Omega \setminus \cup_{j=1}^J \, \Omega_j \right) \right) \qquad \text{(PWMGC)}$$

It will be denoted by (PWMGC) condition in the sequel.

Assume now that a vanishes in a neighborhood of the two poles of a ball in $\mathbb{R}^n$. Then, one can choose two subsets Ω_1 and Ω_2 containing, respectively, the two regions where a vanishes and apply the piecewise multiplier method with $J = 2$ and with the appropriate choices of two observation points and ε. The multiplier method consists of integrating by parts expressions of the form

$$\int_t^T \int_\Omega \left(\partial_t^2 u - \Delta u + a(x) u_t \right) Mu \, \mathrm{d}x \, \mathrm{d}t = 0 \qquad \forall 0 \leq t \leq T \, ,$$

where u stands for a (strong) solution of (33), with an appropriate choice of Mu. Multipliers have generally the form

$$Mu = (m(x) \cdot \nabla u + c\, u)\, \psi(x) \, ,$$

where m depends on the observation points and ψ is a cut-off function. Other multipliers of the form $Mu = \Delta^{-1}(\beta u)$, where β is a cut-off function and Δ^{-1} is the inverse of the Laplacian operator with homogeneous Dirichlet boundary conditions, have also used.

The geometric conditions (MGC) or (PWMGC) serve to bound above by zero terms which cannot be controlled otherwise. One can then prove that the energy satisfies an estimate of the form

$$\int_t^T E(s) \, \mathrm{d}s \leq cE(t) + \int_t^T \left(\int_\Omega a(x)|u_t|^2 + \int_\omega |u_t|^2 \right) \mathrm{d}s \qquad \forall t \geq 0 \, . \quad (37)$$

Once this estimate is proved, one can use the dissipation relation to prove that the energy satisfies integral inequalities of Gronwall type. This is the subject of the next section.

Decay Rates, Integral Inequalities and Lyapunov Techniques

The Linear Feedback Case Using the dissipation relation (34), one has

$$\int_t^T \int_\Omega a|u_t|^2 \, \mathrm{d}x \, \mathrm{d}s \leq \int_t^T -E'(s) \, \mathrm{d}s \leq E(t) \qquad \forall \; 0 \leq t \leq T \, .$$

On the other hand, thanks to assumption (36) on a

$$\int_t^T \int_\omega u_t^2 \, \mathrm{d}x \, \mathrm{d}s \leq \frac{1}{a_0} \int_t^T \int_\Omega a|u_t|^2 \, \mathrm{d}x \, \mathrm{d}s \leq \frac{1}{a_0} E(t) \quad \forall 0 \leq t \leq T \, .$$

By the above inequalities and (37), E satisfies

$$\int_t^T E(s) \, \mathrm{d}s \leq cE(t) \, , \quad \forall 0 \leq t \leq T \, . \qquad (38)$$

Since E is a nonincreasing function and thanks to this integral inequality, Haraux [71] (see also Komornik [74]) proved that E decays exponentially at infinity, that is

$$E(t) \leq E(0) \exp\left(1 - t/c\right) , \quad \forall t \geq c \, . \qquad (39)$$

This proof is as follows. Define

$$\phi(t) = \exp(t/c) \int_t^\infty E(s) \, \mathrm{d}s \quad \forall t \geq 0 \, .$$

Thanks to (38) ϕ is nonincreasing on $[0, \infty)$, so that

$$\phi(t) \leq \phi(0) = \int_0^\infty E(s) \, \mathrm{d}s \, .$$

Using once again (38) with $t = 0$ in this last inequality and the definition of ϕ, one has

$$\int_t^\infty E(s) \, \mathrm{d}s \leq cE(0) \exp(-t/c) \quad \forall t \geq 0 \, .$$

Since E is a nonnegative and nonincreasing function

$$cE(t) \leq \int_{t-c}^t E(s) \, \mathrm{d}s \leq \int_{t-c}^\infty E(s) \, \mathrm{d}s \leq cE(0) \exp(-(t-c)/c) \, ,$$

so that (39) is proved.

An alternative method is to introduce a modified (or perturbed) energy E_ε which is equivalent to the natural one for small values of the parameter ε as in Komornik and

Zuazua [76]. Then one shows that this modified energy satisfies a differential Gronwall inequality so that it decays exponentially at infinity. The exponential decay of the natural energy follows then at once. In this case, the modified energy is indeed a Lyapunov function for the PDE. The natural energy cannot be in general such a Lyapunov function due to the finite speed of propagation (consider initial data which have compact support compactly embedded in $\Omega\backslash\omega$).

There are also very interesting approaches using the frequency domain approach, or spectral analysis such as developed by K. Liu [87] Z. Liu and S. Zheng [88]. In the sequel, we concentrate on the integral inequality method. This method has been generalized in several directions and we present in this note some results concerning extensions to

- nonlinear feedback
- indirect or single feedback for coupled system
- memory type feedbacks

Generalizations to Nonlinear Feedbacks Assume now that the feedback term $a(x)u_t$ in (33) is replaced by a nonlinear feedback $a(x)\rho(u_t)$ where ρ is a smooth, increasing function satisfying $v\rho(v) \geq 0$ for $v \in \mathbb{R}$, linear at ∞ and with polynomial growth close to zero, that is: $\rho(v) = |v|^p$ for $|v| \leq 1$ where $p \in (1, \infty)$.

Assume moreover that ω satisfies Zuazua's multiplier geometric condition (MGC) or Liu's piecewise multiplier method (PWMGC). Then using multipliers of the space and time variables defined as $E(s)^{(p-1)/2}\, Mu(x)$ where $Mu(x)$ are multipliers of the form described in section 5.1 and integrating by parts expressions of the form

$$\int_t^T E(s)^{(p-1)/2} \times \int_\Omega \left(\partial_t^2 u - \Delta u + a(x)\rho(u_t)\right) Mu(x)\, \mathrm{d}x\, \mathrm{d}s = 0\,,$$

one can prove that the energy E of solutions satisfies the following inequality for all $0 \leq t \leq T$

$$\int_t^T E^{(p+1)/2}(s)\, \mathrm{d}t \leq cE^{(p+1)/2}(t) + c\int_t^T E^{(p-1)/2}(s) \times \left(\int_\Omega \rho(u_t)^2 + \int_\omega |u_t|^2\right) .$$

One can remark than an additional multiplicative weight in time depending on the energy has to be taken. This weight is $E^{(p-1)/2}$. Then as in the linear case, but in a more involved way, thanks to the dissipation relation

$$E'(t) = -\int_\Omega a(x) u_t \rho(u_t)\,, \tag{40}$$

one can prove that E satisfies the following nonlinear integral inequality

$$\int_t^T E^{(p+1)/2}(s)\, \mathrm{d}s \leq cE(t)\,, \quad \forall 0 \leq t \leq T\,.$$

Thanks to the fact that E is nonincreasing, a well-known result by Komornik [74] shows that E is polynomially decaying, as $t^{-2/(p-1)}$ at infinity. The above type results have been obtained by many authors under weaker form (see also [40,41,71,98,122]).

Extensions to nonlinear feedbacks without growth conditions close to zero have been studied by Lasiecka and Tataru [78], Martinez [93], W. Liu and Zuazua [89], Eller Lagnese and Nicaise [56] and Alabau–Boussouira [5]. We present the results obtained in this last reference since they provide *optimal* decay rates.

The method is as follows. Define respectively the linear and nonlinear kinetic energies

$$\begin{cases} \int_\omega |u_t|^2\, \mathrm{d}x\,, \\ \int_\Omega |a(x)\rho(u_t)|^2\, \mathrm{d}x\,, \end{cases}$$

and use a weight function in time $f(E(s))$ which is to be determined later on in an optimal way. Integrating by parts expressions of the form

$$\int_t^T f(E(s)) \int_\Omega \left(\partial_t^2 u - \Delta u + a(x)\rho(u_t)\right) Mu(x)\, \mathrm{d}x\, \mathrm{d}s = 0\,,$$

one can prove that the energy E of solutions satisfies the following inequality for all $0 \leq t \leq T$

$$\int_t^T E(s) f(E(s))\, \mathrm{d}s \leq c f(E(t)) + c\int_t^T f(E(s)) \times \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2\right) . \tag{41}$$

The difficulty is to determine the optimal weight under general growth conditions on the feedback close to 0, in particular for cases for which the feedback decays to 0 faster than polynomials.

Assume now that the feedback satisfies

$$g(|v|) \leq |\rho(v)| \leq Cg^{-1}(|v|)\,, \quad \forall |v| \leq 1\,, \tag{42}$$

where g is continuously differentiable on $\mathbb{R}$ strictly increasing with $g(0) = 0$ and

$$\begin{cases} g \in C^2([0, r_0])\,,\, r_0 \text{ sufficiently small }, \\ H(.) = \sqrt{.}g(\sqrt{.}) \text{ is strictly convex on } \left[0, r_0^2\right], \\ g \text{ is odd}\,. \end{cases}$$

Moreover, ρ is assumed to have a linear growth at infinity. We define the *optimal* weight function f as follows.

We first extend H to a function $\hat{H}$ define on all $\mathbb{R}$

$$\hat{H}(x) = \begin{cases} H(x) \text{ if } x \in \left[0, r_0^2\right], \\ +\infty \text{ otherwise}, \end{cases}$$

then, define a function F as follows:

$$F(y) = \begin{cases} \dfrac{\hat{H}^*(y)}{y} & \text{if } \; y \in (0, +\infty)\,, \\ 0 & \text{if } \; y = 0\,, \end{cases}$$

where $\hat{H}^*$ stands for the convex conjugate of $\hat{H}$, that is

$$\hat{H}^*(y) = \sup_{x \in \mathbb{R}}\{x\,y - \hat{H}(x)\}\,.$$

Then the *optimal* weight function f is determined in the following way

$$f(s) = F^{-1}(s/2\beta) \quad s \in \left[0, 2\beta r_0^2\right)\,,$$

where β is of the form $\max(\eta_1, \eta_2 E(0))$, η_1 and η_2 being explicit positive constants.

One can prove that the above formulas make sense, and in particular that F is invertible and smooth. More precisely, F is twice continuously differentiable strictly increasing, one-to-one function from $[0, +\infty)$ onto $[0, r_0^2)$. Note that since the feedback is supposed to be linear at infinity, if one wants to obtain results for general growth types of the feedback, one can assume convexity of H only in a neighborhood of 0.

One can prove from (41) that there exists an (explicit) $T_0 > 0$ such that for all initial data, E satisfies the following nonlinear integral inequality

$$\int_t^T E(s) f(E(s))\,\mathrm{d}s \leq T_0 E(t) \quad \forall 0 \leq t \leq T\,. \tag{43}$$

This inequality is proved thanks to convexity arguments as follows. Thanks to the convexity of $\hat{H}$, one can use Jensen's inequality and (42), so that

$$\int_{\Omega_t} |a(x)\rho(u_t)|^2\,\mathrm{d}x \leq \gamma_1(t)\hat{H}^{-1} \times \left(\frac{1}{\gamma_1(t)} \int_\Omega a(x) u_t \rho(u_t)\,\mathrm{d}x\right)$$

In a similar way, one proves that

$$\int_{\omega_t} |u_t|^2\,\mathrm{d}x \leq \gamma_2(t)\hat{H}^{-1}\left(\frac{1}{\gamma_2(t)} \int_\Omega a(x) u_t \rho(u_t)\,\mathrm{d}x\right)$$

where Ω_t and ω_t are time-dependent sets of respective Lebesgue measures $\gamma_1(t)$ and $\gamma_2(t)$ on which the velocity $u_t(t, x)$ is sufficiently small. Using the above two estimates, together with the linear growth of ρ at infinity, one proves

$$\int_t^T f(E(s))\left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2\right) \leq \int_t^T f(E(s))\hat{H}^{-1}\left(\frac{1}{c}\int_\Omega a(x) u_t \rho(u_t)\,\mathrm{d}x\right)$$

Using then Young's inequality, together with the dissipation relation (40) in the above inequality, one obtains

$$\int_t^T f(E(s))\left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2\right) \leq C_1 \int_t^T \hat{H}^\star\left(f(E(s)\right)\,\mathrm{d}s + C_2 E(t)\,, \tag{44}$$

where $C_i > 0$ $i = 1, 2$ is a constant independent of the initial data. Using the dissipation relation (40) in the above inequality, this gives for all $0 \leq t \leq T$

Combining this last inequality with (41) gives

$$\int_t^T E(s) f(E(s))\,\mathrm{d}s \leq \beta \int_t^T (\hat{H})^\star\left(f(E(s)\right)\,\mathrm{d}s + C_2 E(t)$$

where β is chosen of the form $\max(\eta_1, \eta_2 E(0))$, η_1 and η_2 being explicit positive constants to guarantee that the argument E of f stays in the domain of definition of f. Thus (43) is proved, thanks to the fact that the weight function has been chosen so that

$$\beta\hat{H}^\star(f(E(s)) = \frac{1}{2}E(s) f(E(s)) \quad \forall\, 0 \leq s\,.$$

Therefore E satisfies a nonlinear integral inequality with a weight function $f(E)$ which is defined in a semi-explicit way in general cases of feedback growth.

The last step is to prove that a nonincreasing and nonnegative absolutely continuous function E satisfying a nonlinear integral inequality of the form (43) is decaying at infinity, and to establish at which rate this holds. For this, one proceeds as in [5].

Let $\eta > 0$ and $T_0 > 0$ be fixed given real numbers and F be a strictly increasing function from $[0, +\infty)$ on $[0, \eta)$, with $F(0) = 0$ and $\lim_{y \to +\infty} F(y) = \eta$.

For any $r \in (0, \eta)$, we define a function K_r from $(0, r]$ on $[0, +\infty)$ by

$$K_r(\tau) = \int_\tau^r \frac{\mathrm{d}y}{yF^{-1}(y)}\,, \tag{45}$$

and a function ψ_r which is a strictly increasing function from $[\frac{1}{F^{-1}(r)}, +\infty)$ onto $[\frac{1}{F^{-1}(r)}, +\infty)$ by

$$\psi_r(z) = z + K_r(F(\frac{1}{z})) \geq z, \quad \forall z \geq \frac{1}{F^{-1}(r)}\,, \tag{46}$$

One can prove that if E is a nonincreasing, absolutely continuous function from $[0, +\infty)$ on $[0, +\infty)$, satisfying $0 < E(0) < \eta$ and the inequality

$$\int_t^T E(s)F^{-1}(E(s))\,\mathrm{d}s \leq T_0 E(S)\,, \quad \forall\, 0 \leq t \leq T\,, \tag{47}$$

then E satisfies the following estimate:

$$E(t) \leq F\left(\frac{1}{\psi_r^{-1}(\frac{t}{T_0})}\right), \quad \forall t \geq \frac{T_0}{F^{-1}(r)}\,, \tag{48}$$

where r is any real such that

$$\frac{1}{T_0}\int_0^{+\infty} E(\tau)F^{-1}(E(\tau))\,\mathrm{d}\tau \leq r \leq \eta\,.$$

Thus, one can apply the above result to E with $\eta = r_0^2$ and show that $\lim t \to +\infty E(t) = 0$, the decay rate being given by estimate (48).

If g is polynomial close to zero, one gets back that the energy $E(t)$ decays as $t^{\frac{-2}{p-1}}$ at infinity. If $g(v)$ behaves as $\exp(-1/|v|)$ close to zero, then $E(t)$ decays as $1/(\ln(t))^2$ at infinity.

The usefulness of convexity arguments has been first pointed out by Lasiecka and Tataru [78] using Jensen's inequality and then in different ways by Martinez [93] (the weight function does not depend on the energy) and W. Liu and Zuazua [89] and Eller Lagnese and Nicaise [56]. Optimal decay rates have been obtained by Alabau–Boussouira [5,6] using a weight function determined through the theory of convex conjugate functions and Young's (named also as Fenchel–Moreau's) inequality. This argument was also used by W. Liu and Zuazua [89] in a slightly different way and combined to a Lyapunov technique. Optimality of estimates in [5] is proved in one-dimensional situation and for boundary dampings applying optimality results of Vancostenoble [119] (see also Martinez and Vancostenoble [118]).

Indirect Damping for Coupled Systems

Many complex phenomena are modeled through coupled systems. In stabilizing (or controlling) energies of the vector state, one has very often access only to some components of this vector either due to physical constraints or to cost considerations. In this case, the situation is to stabilize a full system of coupled equation through a reduced number of feedbacks. This is called indirect damping. This notion has been introduced by Russell [109] in 1993.

As an example, we consider the following system:

$$\begin{cases} \partial_t^2 u - \Delta u + \partial_t u + \alpha v = 0 \\ \partial_t^2 v - \Delta v + \alpha u = 0 \end{cases}$$
$$\text{in } \Omega \times \mathbb{R}\,, \quad u = 0 = v \quad \text{on } \partial\Omega \times \mathbb{R}\,. \tag{49}$$

Here, the first equation is damped through a linear distributed feedback, while no feedback is applied to the second equation. The question is to determine if this coupled system inherits any kind of stability for nonzero values of the coupling parameter α from the stabilization of the first equation only.

In the finite dimensional case, stabilization (or control) of coupled ODE's can be analyzed thanks to a powerful rank type condition named Kalman's condition. The situation is much more involved in the case of coupled PDE's.

One can show first show that the above system fails to be exponentially stable (see also [66] for related results). More generally, one can study the stability of the system

$$\begin{cases} u'' + A_1 u + Bu' + \alpha v = 0 \\ v'' + A_2 v + \alpha u = 0 \end{cases} \tag{50}$$

in a separable Hilbert space H with norm $|\cdot|$, where A_1, A_2 and B are self-adjoint positive linear operators in H. Moreover, B is assumed to be a bounded operator. So, our analysis applies to systems with internal damping supported in the whole domain Ω such as (49); the reader is referred to [1,2] for related results concerning boundary stabilization problems (see also Beyrath [23,24] for localized indirect dampings).

In light of the above observations, system (50) fails to be exponentially stable, at least when H is infinite dimensional and A_1 has a compact resolvent as in (49). Indeed it is shown in Alabau, Cannarsa and Komornik [8] that the total energy of sufficiently smooth solutions of (50) decays polynomially at infinity whenever $|\alpha|$ is small enough but nonzero. From this result we can also deduce that any solution of (50) is strongly stable regardless of its smoothness: this fact follows by a standard density argument since the semigroup associated with (50) is a contraction semigroup.

A brief description of the key ideas of the approach developed in [2,8] is as follows. Essentially, one uses a finite iteration scheme and suitable multipliers to obtain an estimate of the form

$$\int_0^T E(u(t), v(t))\mathrm{d}t \leq c \sum_{k=0}^{j} E(u^{(k)}(0), v^{(k)}(0)) \qquad \forall\, T \geq 0\,, \quad (51)$$

where j is a positive integer and E denotes the total energy of the system

$$E(u, v) = \frac{1}{2}\left(|A_1^{1/2}u|^2 + |u'|^2\right) + \frac{1}{2}\left(|A_2^{1/2}v|^2 + |v'|^2\right) + \alpha\langle u, v\rangle\,.$$

Once (51) is proved, an abstract lemma due to Alabau [1,2] shows that $E(u(t), v(t))$ decays polynomially at ∞. This abstract lemma can be stated as follows.

Let A be the infinitesimal generator of a continuous semi-group $\exp(tA)$ on an Hilbert space $\mathcal{H}$, and $D(A)$ its domain. For U^0 in $\mathcal{H}$ we set in all the sequel $U(t) = \exp(tA)U^0$ and assume that there exists a functional E defined on $C([0, +\infty), \mathcal{H})$ such that for every U^0 in $\mathcal{H}$, $E(\exp(.A))$ is a non-increasing, locally absolutely continuous function from $[0, +\infty)$ on $[0, +\infty)$. Assume moreover that there exist an integer $k \in \mathbb{N}^\star$ and nonnegative constants c_p for $p = 0, \ldots k$ such that

$$\int_S^T E(U(t))\mathrm{d}t \leq \sum_{p=0}^{k} c_p E(U^{(p)}(S)) \qquad \forall 0 \leq S \leq T\,, \forall U^0 \in D(A^k)\,. \quad (52)$$

Then the following inequalities hold for every U^0 in $D(A^{kn})$ and all $0 \leq S \leq T$ where n is any positive integer:

$$\int_S^T E(U(\tau))\frac{(\tau - S)^{n-1}}{(n-1)!}\mathrm{d}\tau \leq c \sum_{p=0}^{kn} E(U^{(p)}(S))\,, \quad (53)$$

and

$$E(U(t)) \leq c \sum_{p=0}^{kn} E(U^{(p)}(0))t^{-n} \qquad \forall t > 0\,, \quad \forall U^0 \in D(A^{kn})\,,$$

where c is a constant which depends on n.

First (53) is proved by induction on n. For $n = 1$, it reduces to the hypothesis (52). Assume now that (53) holds for n and let U^0 be given in $D(A^{k(n+1)})$. Then we have

$$\int_S^T \int_t^T E(U(\tau))\frac{(\tau - t)^{n-1}}{(n-1)!}\mathrm{d}\tau\mathrm{d}t \leq c \sum_{p=0}^{kn} \int_S^T E(U^{(p)}(t))\mathrm{d}t \qquad \forall\, 0 \leq S \leq T\,, \forall\, U^0 \in D(A^{kn}).$$

Since U^0 is in $D(A^{k(n+1)})$ we deduce that $U^{(p)}(0) = A^p U^0$ is in $D(A^k)$ for $p \in \{0, \ldots kn\}$. Hence we can apply the assumption (52) to the initial data $U^{(p)}(0)$. This together with Fubini's Theorem applied on the left hand side of the above inequality give (53) for $n + 1$. Using the property that $E(U(t))$ is non increasing in (53) we easily obtain the last desired inequality.

Applications to wave-wave, wave-Petrowsky equations and various concrete examples hold.

The above results have been studied later on by Batkai, Engel, Prüss and Schnaubelt [18] using very interesting resolvent and spectral criteria for polynomial stability of abstract semigroups. The above abstract lemma in [2] has also been generalized using interpolation theory. One should note that this integral inequality involving higher order energies of solutions is not of differential nature, unlike Haraux's and Komornik's integral inequalities. Another approach based on decoupling techniques and for slightly different abstract systems have been introduced by Ammar Khodja Bader and Ben Abdallah [12].

Spectral conditions have also been studied by Z. Liu [88] and later on by Z. Liu and Rao [90], Loreti and Rao [92] for peculiar abstract systems and in general for coupled equations only of the *same* nature (wave-wave for instance), so that a dispersion relation for the eigenvalues of the coupled system can be derived. Also these last results are given for internal stabilization only. Because of the above limitations, Z. Liu–Rao and Loreti–Rao's results are less powerful in generality than the ones given by Alabau, Cannarsa and Komornik [8] and Alabau [2]. Moreover results through energy type estimates and integral inequalities can be generalized to include nonlinear indirect dampings as shown in [7]. On the other side spectral methods are very useful to obtain optimal decay rates provided that one can determine at which speed the eigenvalues approach the imaginary axis for high frequencies.

Memory Dampings

We consider the following model problem

$$\begin{cases} u_{tt}(t,x) - \Delta u(t,x) + \int_0^t \beta(t-s)\,\Delta u(s,x)\,\mathrm{d}s = \\ \qquad\qquad\qquad\qquad |u(t,x)|^{\gamma} u(t,x) \\ u(t,\cdot)_{|\partial\Omega} = 0 \\ (u(0,\cdot), u_t(0,\cdot)) = (u_0, u_1) \end{cases} \tag{54}$$

where $0 < \gamma \leq \frac{2}{N-2}$ holds. The second member is a source term. The damping

$$\int_0^t \beta(t-s)\Delta u(s,x)\,\mathrm{d}s$$

is of memory type.

The energy is defined by

$$\begin{aligned} E_u(t) = {} & \frac{1}{2}\|u_t(t)\|^2_{L^2(\Omega)}\,\mathrm{d}x \\ & + \frac{1}{2}\left(1 - \int_0^t \beta(s)\,\mathrm{d}s\right)\|\nabla u(t)\|^2_{L^2(\Omega)} \\ & - \frac{1}{\gamma+2}\|u(t)\|^{\gamma+2}_{L^{\gamma+2}(\Omega)} \\ & + \frac{1}{2}\int_0^t \beta(t-s)\|\nabla u(t) - \nabla u(s)\|^2_{L^2(\Omega)}\,\mathrm{d}s \end{aligned}$$

The damping term produces dissipation of the energy, that is (for strong solutions)

$$\begin{aligned} E'_u(t) = {} & -\frac{1}{2}\beta(t)\|\nabla u(t)\|^2 \\ & + \frac{1}{2}\int_0^t \beta'(t)\|\nabla u(s) - \nabla u(t)\|^2\,\mathrm{d}s \leq 0 \end{aligned}$$

One can consider more general abstract equations of the form

$$u''(t) + Au(t) - \int_0^t \beta(t-s)Au(s)\,\mathrm{d}s = \nabla F(u(t)) \qquad t \in (0,\infty) \tag{55}$$

in a Hilbert space X, where $A\colon D(A) \subset X \to X$ is an accretive self-adjoint linear operator with dense domain, and ∇F denotes the gradient of a Gâteaux differentiable functional $F\colon D(A^{1/2}) \to \mathbb{R}$. In particular, equation (54) fits into this framework as well as several other classical equations of mathematical physics such as the linear elasticity system.

We consider the following assumptions.

Assumptions (H1)

1. A is a self-adjoint linear operator on X with dense domain $D(A)$, satisfying

$$\langle Ax, x\rangle \geq M\|x\|^2 \quad \forall x \in D(A) \tag{56}$$

for some $M > 0$.

2. $\beta\colon [0,\infty) \to [0,\infty)$ is a locally absolutely continuous function such that

$$\int_0^\infty \beta(t)\mathrm{d}t < 1 \; \beta(0) > 0 \quad \beta'(t) \leq 0 \qquad \text{for a.e. } t \geq 0\,.$$

3. $F\colon D(A^{1/2}) \to \mathbb{R}$ is a functional such that
   1. F is Gâteaux differentiable at any point $x \in D(A^{1/2})$;
   2. for any $x \in D(A^{1/2})$ there exists a constant $c(x) > 0$ such that

   $$|DF(x)(y)| \leq c(x)\|y\|, \quad \text{for any } y \in D(A^{1/2}),$$

   where $DF(x)$ denotes the Gâteaux derivative of F in x; consequently, $DF(x)$ can be extended to the whole space X (and we will denote by $\nabla F(x)$ the unique vector representing $DF(x)$ in the Riesz isomorphism, that is, $\langle \nabla F(x), y\rangle = DF(x)(y)$, for any $y \in X$);
   3. for any $R > 0$ there exists a constant $C_R > 0$ such that

   $$\|\nabla F(x) - \nabla F(y)\| \leq C_R\|A^{1/2}x - A^{1/2}y\|$$

   for all $x, y \in D(A^{1/2})$ satisfying $\|A^{1/2}x\|\,,\; \|A^{1/2}y\| \leq R$.

Assumptions (H2)

1. There exist $p \in (2,\infty]$ and $k > 0$ such that

$$\beta'(t) \leq -k\beta^{1+\frac{1}{p}}(t) \quad \text{for a.e. } t \geq 0$$

(here we have set $\frac{1}{p} = 0$ for $p = \infty$).

2. $F(0) = 0$, $\nabla F(0) = 0$, and there is a strictly increasing continuous function $\psi\colon [0,\infty) \to [0,\infty)$ such that $\psi(0) = 0$ and

$$|\langle \nabla F(x), x\rangle| \leq \psi(\|A^{1/2}x\|)\|A^{1/2}x\|^2 \quad \forall x \in D(A^{1/2})\,.$$

Under these assumptions, global existence for sufficiently small (resp. all) initial data in the energy space can be proved for nonvanishing (resp. vanishing) source terms.

It turns out that the above energy methods based on multiplier techniques combined with linear and nonlinear integral inequalities can be extended to handle memory dampings and applied to various concrete examples such as wave, linear elastodynamic and Petrowsky equations for instance. This allows to show in [10] that exponential as well as polynomial decay of the energy holds if the kernel decays respectively exponentially or polynomially at infinity.

The method is as follows. One evaluates expressions of the form

$$\int_t^T \langle u''(s) + Au(s) - \int_0^t \beta * Au(s) - \nabla F(u(s), Mu\rangle \,\mathrm{d}s$$

where the multipliers Mu are of the form $\phi(s)(c_1(\beta * u)(s) + c_2(s)u)$ with ϕ which is a differentiable, nonincreasing and nonnegative function, and c_1 being a suitable constant, whereas c_2 may be chosen dependent on β.

Integrating by parts the resulting relations and performing some involved estimates, one can prove that for all $t_0 > 0$ and all $T \geq t \geq t_0$

$$\int_t^T \phi(s)E(s)\,\mathrm{d}s \leq C\phi(0)E(t) + \int_t^T \phi(s) \times \int_0^s \beta(s-\tau)\left\|A^{1/2}u(s) - A^{1/2}u(\tau)\right\|^2 \,\mathrm{d}\tau\,\mathrm{d}s\,,$$

If $p = \infty$, that is if the kernel β decays exponentially, one can easily bound the last term of the above estimate by $cE(t)$ thanks to the dissipation relation.

If $p \in (2, \infty)$, one has to proceed differently since the term

$$\int_t^T \phi(s)\int_0^s \beta(s-\tau)\left\|A^{1/2}u(s) - A^{1/2}u(\tau)\right\|^2 \,d\tau\,\mathrm{d}s$$

cannot be directly estimated thanks to the dissipation relation. To bound this last term, one can generalize an argument of Cavalcanti and Oquendo [37] as follows. Define, for any $m \geq 1$,

$$\varphi_m(t) := \int_0^t \beta^{1-\frac{1}{m}}(t-s)\|A^{1/2}u(s) - A^{1/2}u(t)\|^2\mathrm{d}s\,, \quad t \geq 0\,. \tag{57}$$

Then, we have for any $T \geq S \geq 0$

$$\int_S^T E_u^{\frac{m}{p}}(t)\int_0^t \beta(t-s)\|A^{1/2}u(s) - A^{1/2}u(t)\|^2\mathrm{d}s\mathrm{d}t \leq CE_u^{\frac{p}{p+m}}(S)\left(\int_S^T E_u^{1+\frac{m}{p}}(t)\varphi_m(t)\mathrm{d}t\right)^{\frac{m}{p+m}} \tag{58}$$

for some constant $C > 0$. Suppose that, for some $m \geq 1$, the function φ_m defined in (57) is bounded. Then, for any $S_0 > 0$ there is a positive constant C such that

$$\int_S^\infty E_u^{1+\frac{m}{p}}(t)\mathrm{d}t \leq C\Big(E_u^{\frac{m}{p}}(0) + \|\varphi_m\|_\infty^{\frac{m}{p}}\Big)E_u(S) \quad \forall\, S \geq S_0\,. \tag{59}$$

One uses this last result first with $m = 2$ noticing that φ_2 is bounded and $\phi = E^{2/p}$. This gives a first energy decay rate as $(t+1)^{-p/2}$. This estimate shows that φ_1 is bounded. Then one applies once again the last result with $m = 1$ and $\phi = E^{1/p}$. One deduces then that E decays as $(t+1)^{-p}$ which is the *optimal* decay rate expected.

Bibliographical Comments

For an introduction to the multiplier method, we refer the interested reader to the books of J.-L. Lions [86], Komornik [74] and the references therein. The celebrated result of Bardos Lebeau and Rauch is presented in [86]. A general abstract presentation of control problems for hyperbolic and parabolic equations can be found in the book of Lasiecka and Triggiani [80,81]. Results on spectral methods and the frequency domain approach can be found in the book of Z. Liu [88]. There also exists an interesting approach developed for bounded feedback operators by Haraux and extended to the case of unbounded feedbacks by Ammari and Tucsnak [11]. In this approach, the polynomial (or exponential) stability of the damped system is proved thanks to the corresponding observability for the undamped (conservative) system. Such observability results for weakly coupled undamped systems have been obtained for instance in [3].

Many other very interesting issues have been studied in connection to semilinear wave equations, see [34,123] and the references therein, and damped wave equations with nonlinear source terms [39].

Well-posedness and asymptotic properties for PDE's with memory terms have first been studied by Dafermos [53,54] for convolution kernels with past history (convolution up to $t = -\infty$), by Prüss [103] and Prüss and Propst [102] in which the efficiency of different models of dampings are compared to experiments (see also Londen Petzeltova and Prüss [91]). Decay estimates for the energy of solutions using multiplier methods combined with Lyapunov type estimates for an equivalent energy are proved in Munoz Rivera [97], Munoz Rivera and Salvatierra [96], Cavalcanti and Oquendo [37] and Giorgi Naso and Pata [67] and many other papers.

Optimal Control

As for positional control, also for optimal control problems it is convenient to adopt the abstract formulation introduced in Sect. "Abstract Evolution Equations". Let the state space be represented by the Hilbert space H, and the state equation be given in the form (12), that is

$$\begin{cases} u'(t) = Au(t) + Bf(t) & t \in [0, T] \\ u(0) = u_0 \, . \end{cases} \tag{60}$$

Recall that A is the infinitesimal of a strongly continuous semigroup, e^{tA}, in H, B is a (bounded) linear operator from F (the control space) to H, and u_f stands for the unique (mild) solution of (60) for a given control function $f \in L^2(0, T; H)$.

A typical optimal control problem of interest for PDE's is the *Bolza problem* which consists in

$$\begin{cases} \text{minimizing the } \textit{cost functional} \\ \quad J(f) \doteq \int_0^T L(t, u_f(t), f(t))\mathrm{d}t + \ell\left(u_f(T)\right) \\ \text{over all controls} \quad f \in L^2(0, T; F) \, . \end{cases} \tag{61}$$

Here, T is a positive number, called the *horizon*, whereas L and ℓ are given functions, called the *running cost* and *final cost*, respectively. Such functions are usually assumed to be *bounded below*.

A control function $f_* \in L^2(0, T; F)$ at which the above minimum is attained is called an *optimal control* for problem (61) and the corresponding solution u_{f_*} of (60) is said to be an *optimal trajectory*. Alltogether, $\{u_{f_*}, f_*\}$ is called an *optimal (trajectory/control) pair*.

For problem (61) the following issues will be addressed in the sections below:

- *the existence* of controls minimizing functional J;
- *necessary conditions* that a candidate solution must satisfy;
- *sufficient conditions for optimality* provided by the dynamic programming method.

Other problems of particular interest to CT for PDE's are problems with an *infinite horizon* ($T = \infty$), problems with a *free horizon* T and a final *target*, and problems with constraints on both control variables and state variables. Moreover, the study of nonlinear variants of (60), including semilinear problems of the form

$$\begin{cases} u'(t) = Au(t) + h(t, u(t), f(t)) & t \in [0, T] \\ u(0) = u_0 \, , \end{cases} \tag{62}$$

is strongly motivated by applications. The discussion of all these variants, however, will not be here pursued in detail.

Traditionally, in optimal control theory, state variables are denoted by the letters $x, y, \ldots$, whereas $u, v, \ldots$ are reserved for control variables. For notational consistency, in this section $u(\cdot)$ will still denote the state of a given system and $f(\cdot)$ a control function, while ϕ will stand for a fixed element of control space F.

Existence of Optimal Controls

From the study of finite dimensional optimization it is a familiar fact that the two essential ingredients to guarantee the existence of minima are compactness and lower semicontinuity. Therefore, it is clear that, in order to obtain a solution of the optimal control problem (60)–(61), one has to make assumptions that allow to recover such properties. The typical hypotheses that are made for this purpose are the following:

- *coercivity*: there exist constants $c_0 > 0$ and $c_1 \in \mathbb{R}$ such that

$$\ell(\phi) \geq c_1 \quad \text{and} \quad L(t, u, \phi) \geq c_0\|\phi\|^2 + c_1 \qquad \forall (t, u, \phi) \in [0, T] \times H \times F \tag{63}$$

- *convexity*: for every $(t, u) \in [0, T] \times H$

$$\phi \mapsto L(t, u, \phi) \quad \text{is convex on} \quad F \, . \tag{64}$$

Under the above hypotheses, assuming lower semicontinuity of ℓ and of the map $L(t, \cdot, \phi)$, it is not hard to show that problem (60)–(61) has at least one solution. Indeed, assumption (63) allows to show that any minimizing sequence of controls $\{f_k\}$ is bounded in $L^2(0, T; H)$. So, it admits a subsequence, still denoted by $\{f_k\}$ which converges weakly in $L^2(0, T; H)$ to some function f. Then, by linearity, $u_{f_k}(t)$ converges to $u_f(t)$ for every $t \in [0, T]$. So, using assumption (64), it follows that f is a solution of (60)–(61).

The problem becomes more delicate when the Tonelli type coercivity condition (63) is relaxed, or the state equation is nonlinear as in (62). Indeed, the convergence of $u_{f_k}(t)$ is no longer ensured, in general. So, in order to recover compactness, one has to make further assumptions, such as the compactness of e^{tA}, or structural properties of L and h. For further reading, one may consult the monographs [22,85], and [79], for problems where the running and final costs are given by quadratic forms (the so-called Linear Quadratic problem), or [84] and [59] for more general optimal control problems.

Necessary Conditions

Once the existence of a solution to problem (60)–(61) has been established, the next important step is to provide conditions to detect a candidate solution, possibly showing that it is, in fact, optimal. By and large the optimality conditions of most common use are the ones known as Pontryagin's Maximum Principle named after the Russian mathematician L.S. Pontryagin who greatly contributed to the development of control theory, see [100,101].

So, suppose $\{u_*, f_*\}$, where $u_* = u_{f_*}$ is a candidate optimal pair and consider the so-called adjoint system

$$\begin{cases} -p'(t) = A^* p(t) + \partial_u L(t, u_*(t), f_*(t)) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad t \in [0,T] \text{ a.e.} \\ p(T) = \partial \ell(u_*(T)) , \end{cases}$$

where $\partial_u L(t,u,\phi)$ and $\partial\ell(u)$ denote the Fréchet gradients of the maps $L(t,\cdot,\phi)$ and ℓ at u, respectively. Observe that the above is a backward linear Cauchy problem with terminal condition, which can obviously be reduced to a forward one by the change of variable $t \to T - t$. So, it admits a unique mild solution, labeled p_*, which is called the *adjoint state* associated with $\{u_*, f_*\}$.

Pontryagin's Maximum Principle states that, if $\{u_*, f_*\}$ is optimal, then

$$\langle p_*(t), Bf_*(t)\rangle + L(t, u_*(t), f_*(t)) = \min_{\phi\in F}\left[\langle p_*(t), B\phi\rangle + L(t, u_*(t), \phi)\right] \qquad t \in [0,T] \text{ a.e.} \tag{65}$$

The name Maximum Principle rather than Minimum Principle, as it would be more appropriate, is due to the fact that, traditionally, attention was focussed on the *maximization*—instead of minimization—of the functional in (61). Even today, in most models from economics, one is interested in maximizing payoffs, such as revenues, utility, capital and so on. In that case, (65) would still be true, with a "max" instead of a "min".

At first glance, it might be hard to understand the revelance of (65) to problem (61). To explain this, introduce the function, called the *Hamiltonian*,

$$\mathcal{H}(t,u,p) = \min_{\phi\in F}\left[\langle p, B\phi\rangle + L(t,u,\phi)\right] \qquad (t,u,p) \in [0,T]\times H \times H . \tag{66}$$

Then, Fermat's rule yields $B^* p + \partial_\phi L(t,u,\phi) = 0$ at every $\phi \in F$ at which the minimum in (66) is attained. Therefore, from (65) it follows that

$$B^* p_*(t) + \partial_\phi L(t, u_*(t), f_*(t)) = 0 \qquad t \in [0,T] \text{ a.e.} \tag{67}$$

which provides a much-easier-to-use optimality condition.

There is a vast literature on necessary conditions for optimality for distributed parameter systems. The set-up that was considered above can be generalized in several ways: one can consider nonlinear state equations as in (62), nonsmooth running and finals costs, constraints on both state and control, problems with infinite horizon or exit times. Further reading and useful references on most of these extensions can be found in the aforementioned monographs [22,79,84,85], and in [59] which is mainly concerned with time optimal control problems.

Dynamic Programming

Though useful as it may be, Pontryagin's Maximum Principle remains a necessary condition. So, without further information, it does not suffice to prove the optimality of a give trajectory/control pair. Moreover, even when the map $\phi \mapsto \partial_\phi L(t,u,\phi)$ turns out to be invertible, the best result identity (67) can provide, is a representation of $f_*(t)$ in terms of $u_*(t)$ and $p_*(t)$: not enough to determine $f_*(t)$, in general.

This is why other methods to construct optimal controls have been proposed over the years. One of the most interesting ones is the so-called *dynamic programming* method (abbreviated, DP), initiated by the work of R. Bellman [20]. Such a method will be briefly described below in the set-up of distributed parameter systems.

Fix $T > 0$, s such that $0 \le s \le T$, and consider the optimal control problem

to minimize

$$J^{s,\nu}(f) = \int_s^T L\left(t, u_f^{s,\nu}(t), f(t)\right) \mathrm{d}t + \ell\left(u_f^{s,\nu}(T)\right) \tag{68}$$

over all control functions $f \in L^2(s,T;F)$, where $u_f^{s,\nu}(t)$ is the solution of the controlled system

$$\begin{cases} u'(t) = Au(t) + Bf(t) & t \in [s,T] \\ u(s) = \nu . \end{cases} \tag{69}$$

The *value function* U associated to (68)-(69) is the real-valued function defined by

$$U(s,\nu) = \inf_{f\in L^2(s,T;F)} J^{s,\nu}(f) \quad \forall (s,\nu) \in [0,T]\times H . \tag{70}$$

A fundamental step of DP is the following result, known as Bellman's *optimality principle*.

Theorem 5 *For any* $(s, v) \in [0, T] \times H$ *and any* $f \in L^2(s, T; F)$

$$U(s,v) \leq \int_s^r L\left(t, u_f^{s,v}(t), f(t)\right) dt + U\left(r, u_f^{s,v}(r)\right) \quad \forall r \in [s,T]\,.$$

Moreover, $f^*(\cdot)$ *is optimal if and only if*

$$U(s,v) = \int_s^r L\left(t, u_f^{s,v}(t), f(t)\right) dt + U\left(r, u_f^{s,v}(r)\right) \quad \forall r \in [s,T]\,.$$

The connection between DP and optimal control is based on the properties of the value function. Indeed, applying Bellman's optimality principle, one can show that, if U is Fréchet differentiable, then

$$\begin{cases} \partial_s U(s,v) + \langle Av, \partial_v U(s,v)\rangle + \mathcal{H}(s,v,\partial_v U(s,v)) = 0 \\ \qquad\qquad (s,v) \in (0,T) \times D(A) \\ U(T,v) = \ell(v) \quad v \in H \end{cases}$$

where $\mathcal{H}$ is the Hamiltonian defined in (66). The above equation is the celebrated *Hamilton–Jacobi equation* of DP. To illustrate its connections with the original optimal control problem, a useful formal argument—that can, however, be made rigorous—is the following. Consider a sufficiently smooth solution W of the above problem and let $(s, v) \in (0, T) \times D(A)$. Then, for any trajectory/control pair $\{u, f\}$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} W(t,u(t)) &= \partial_s W(t,u(t)) + \langle \partial_v W(t,u(t)), Au(t) \\ &\quad + Bf(t)\rangle \\ &= \langle \partial_v W(t,u(t)), Bf(t)\rangle \\ &\quad - \mathcal{H}(t,u(t),\partial_v W(t,u(t))) \\ &\geq -L(t,u(t),f(t)) \end{aligned} \tag{71}$$

by the definition of $\mathcal{H}$. Therefore, integrating from s to T,

$$\ell(u(T)) - W(s,v) \geq -\int_s^T L(t,u(t),f(t))\mathrm{d}t,$$

whence $J^{s,v}(f) \geq W(s,v)$. Thus, taking the infimum over all $f \in L^2(s,T;F)$,

$$W(s,v) \leq U(s,v) \quad \forall (s,v) \in (0,T) \times D(A)\,. \tag{72}$$

Now, suppose there is a control function $f_* \in L^2(s,T;F)$ such that, for all $t \in [s,T]$,

$$\begin{aligned} \langle \partial_v W(t,u_*(t)), Bf_*(t)\rangle &+ L(t,u_*(t),f_*(t)) \\ &= \mathcal{H}(t,u_*(t),\partial_v W(t,u_*(t)))\,, \end{aligned} \tag{73}$$

where $u_*(\cdot) = u_{f_*}^{s,v}(\cdot)$. Then, from (71) and (73) it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t,u_*(t)) = -L(t,u(t),f(t))\,,$$

whence

$$W(s,v) = J^{s,v}(f_*) \geq U(s,v)\,.$$

From the above inequality and (72) it follows that $W(s,v) = U(s,v)$ for all $(s,v) \in (0,T) \times D(A)$, hence for all $(s,v) \in (0,T) \times H$ since $D(A)$ is dense in H. So, f_* is an **optimal control**.

Note 2 The above considerations lead to the following procedure to obtain optimal an optimal trajectory:

- find a smooth solution of the Hamilton–Jacobi equation;
- for every $(t,v) \in (0,T) \times D(A)$ provide a feedback $f(t,v)$ such that

$$\begin{aligned} \langle \partial_v W(t,v), Bf(t,v)\rangle &+ L(t,v,f(t,v)) \\ &= \mathcal{H}(t,v,\partial_v W(t,v)) \end{aligned}$$

- solve the so-called *closed loop equation*

$$\begin{cases} u'(t) = Au(t) + Bf(t,u(t)) & t \in [s,T] \\ u(s) = v \end{cases}$$

Notice that not only is trajectory u optimal, but the corresponding control f is given in feedback form as well.

Linear Quadratic Optimal Control

One of the most successful applications of DP is the so-called Linear Quadratic optimal control problem. Consider problem (68)–(69) with costs L and ℓ given by

$$L(t,u,\phi) = \langle M(t)u,u\rangle + \langle N(t)\phi,\phi\rangle \quad \forall (T,u,\phi) \in [0,T] \times H \times F$$

and

$$\ell(u) = \langle Du,u\rangle \quad \forall u \in H\,,$$

where

- $M\colon [0,T] \to \mathcal{L}(H)$ is continuous, $M(t)$ is symmetric and $\langle M(t)u,u\rangle \geq 0$ for every $(t,u) \in [0,T] \times H$;
- $N\colon [0,T] \to F$ is continuous, $N(t)$ is symmetric and $\langle N(t)\phi,\phi\rangle \geq c_0|\phi|^2$ for every $(t,\phi) \in [0,T] \times F$ and some constant $c_0 > 0$;

- $D \in \mathcal{L}(H)$ is symmetric and $\langle Du, u\rangle \geq 0$ for every $u \in H$.

Then, assumptions (63) and (64) are satisfied. So, a solution to (68)–(69) does exist. Moreover, it is unique because of the strict convexity of functional $J^{s,v}$.

In order to apply DP, one computes the Hamiltonian

$$\mathcal{H}(t,u,p) = \min_{\phi \in F} \Big[\langle p, B\phi\rangle + \langle M(t)u,u\rangle + \langle N(t)\phi,\phi\rangle\Big] = \langle M(t)u,u\rangle - \frac{1}{4}\langle BN^{-1}(t)B^* p, p\rangle \,,$$

where the above minimum is attained at

$$\phi_*(t,p) = -\frac{1}{2}N^{-1}(t)B^* p \,. \tag{74}$$

Therefore, the Hamilton–Jacobi equation associated to the problem is

$$\begin{cases} \partial_s W(s,v) + \langle Av, \partial_v W(s,v)\rangle + \langle M(s)v,v\rangle \\ \quad -\frac{1}{4}\langle BN^{-1}(s)B^* \partial_v W(s,v), \partial_v W(s,v)\rangle = 0 \\ \qquad\qquad\qquad\qquad \forall (s,v) \in (0,T)\times D(A) \\ w(T,v) = \langle Dv,v\rangle \quad \forall v \in H \end{cases}$$

It is quite natural to search a solution of the above problem in the form

$$W(s,v) = \langle P(s)v,v\rangle \quad \forall (s,v) \in [0,T]\times H\,,$$

with $P\colon [0,T] \to \mathcal{L}(H)$ continuous, symmetric and such that $\langle P(t)u,u\rangle \geq 0$. Substituting into the Hamilton–Jacobi equation yields

$$\begin{cases} \langle P'(s)v,v\rangle + \langle [A^*P(s) + P(s)A]v,v\rangle + \langle M(s)v,v\rangle \\ \quad -\langle BN^{-1}(s)B^*P(s)v, P(s)v\rangle = 0 \\ \qquad\qquad\qquad\qquad \forall (s,v) \in (0,T)\times D(A) \\ \langle P(T)v,v\rangle = \langle Dv,v\rangle \quad \forall v \in H \end{cases}$$

Therefore, P must be a solution of the so-called *Riccati equation*

$$\begin{cases} P'(s) + A^*P(s) + P(s)A + M(s) \\ \quad -P(s)BN^{-1}(s)B^*P(s) = 0 \qquad \forall s \in (0,T) \\ P(T) = D \end{cases}$$

Once a solution $P(\cdot)$ of the Riccati equation is known, the procedure described in Note 2 can be applied. Indeed, recalling (74) and the fact that $\partial_v W(t,v) = 2P(t)v$, one concludes that $f(t,v) = -N^{-1}(t)B^*P(t)v$ is a feedback law. So, solving the closed loop equation

$$\begin{cases} u'(t) = [A - BN^{-1}(t)B^*P(t)]u(t) \quad t \in (s,T) \\ u(s) = v \end{cases}$$

one obtains the unique optimal trajectory of problem (68)–(69).

In sum, by DP one reduces the original Linear Quadratic optimal control problem to the problem of finding the solution of the Riccati equation, which is easier to solve than the Hamilton–Jacobi equation.

Bibliographical Comments

Different variants of the Riccati equation have been successfully studied by several authors in connection with different state equations and cost functionals, including boundary control problems and problems for other functional equations, see [22,79] and the references therein. Sometimes, the solution of the Riccati equation related to a linearized model provides feedback stabilization for nonlinear problems as in [104].

Unfortunately, the DP method is hard to implement for general optimal control problems, because of several obstructions: nonsmoothness of solutions to Hamilton–Jacobi equations, selection problems that introduce discontinuities, unboundedness of the coefficients, numerical complexity. Besides the Linear Quadratic case, the so-called Linear Convex case is the other example that can be studied by DP under fairly general conditions, see [14]. For nonlinear optimal control problems some of the above difficulties have been overcome extending the notion of viscosity solutions to infinite dimensional spaces, see [45,46,47,48,49], see also [28,29,30,31,32,33] and [112]. Nevertheless, finding additional ideas to make a generalized use of DP for distributed parameter systems possible, remains a challenging problem for the next generations.

Future Directions

In addition to all considerations spread all over this article on promising developments of recent—as well as established—research lines, a few additional topics deserve to be mentioned.

The one subject that has received the highest attention, recently, is that of numerical approximation of control problems, from the point of view of both controllability and optimal control. Here the problem is that, due to high frequency spurious numerical solutions, stable algorithms for solving initial-boundary value problems do not necessarily yield convergent algorithms for computing controls. This difficulty is closely related to the existence of concentrated numerical solutions that escape the observation mechanisms. Nevertheless, some interesting results have been obtained so far, see, e. g., [124,125].

Several interesting results for nonlinear control problems have been obtained by the *return method*, developed initially by Coron [42] for a stabilization problem. This and other techniques have then been applied to fluid models ([68,69]), the Korteweg–de Vries equation ([105,106,107]), and Schrödinger type equations ([19]), see also [43] and the references therein. It seems likely that these ideas, possibly combined with other techniques like Carleman estimates as in [70], will lead to new exiting results in the years to come.

A final comment on null controllability for *degenerate parabolic equations* is in order. Indeed, many problems that are relevant for applications are described by parabolic equation equations in divergence form

$$\partial_t u = \nabla\cdot(A(x)\nabla u) + b(x)\cdot\nabla u + c(t,x)u + f \quad \text{in} \quad Q_T\,,$$

or in the general form

$$\partial_t u = \text{Tr}\,[A(x)\nabla^2 u] + b(x)\cdot\nabla u + c(t,x)u + f \quad \text{in} \quad Q_T\,,$$

where $A(x)$ is a symmetric matrix, positive definite in Ω but possibly singular on Γ. For instance, degenerate parabolic equations arise in fluid dynamics as suitable transformations of the Prandtl equations, see, e. g., [94]. They can also be obtained as Kolmogorov equations of diffusions processes on domains that are invariant for stochastic flows, see, e. g., [52]. The latter interpretation explains why they have been applied to biological problems, such as gene frequency models for population genetics (see, e. g., the Wright–Fischer model studied in [111]).

So far, null controllability properties of degenerate parabolic equations have been fully understood only in dimension one: for some kind of degeneracy, null controllability holds true (see [36] and [9]), but, in general, one can only expect regional null controllability (see [35]). Since very little is known on null controllability for degenerate parabolic equations in higher space dimensions, it is conceivable that such a topic will provide interesting problems for future developments.

Bibliography

1. Alabau F (1999) Stabilisation frontière indirecte de systèmes faiblement couplés. C R Acad Sci Paris Sér I 328:1015–1020
2. Alabau F (2002) Indirect boundary stabilization of weakly coupled systems. SIAM J Control Optim 41(2):511–541
3. Alabau-Boussouira F (2003) A two-level energy method for indirect boundary observability and controllability of weakly coupled hyperbolic systems. SIAM J Control Optim 42(3):871–906
4. Alabau-Boussouira F (2004) A general formula for decay rates of nonlinear dissipative systems. C R Math Acad Sci Paris 338:35–40
5. Alabau-Boussouira F (2005) Convexity and weighted integral inequalities for energy decay rates of nonlinear dissipative hyperbolic systems. Appl Math Optim 51(1):61–105
6. Alabau-Boussouira F (2006) Piecewise multiplier method and nonlinear integral inequalities for Petrowsky equations with nonlinear dissipation. J Evol Equ 6(1):95–112
7. Alabau-Boussouira F (2007) Asymptotic behavior for Timoshenko beams subject to a single nonlinear feedback control. NoDEA 14(5–6):643–669
8. Alabau F, Cannarsa P, Komornik V (2002) Indirect internal damping of coupled systems. J Evol Equ 2:127–150
9. Alabau-Boussouira F, Cannarsa P, Fragnelli G (2006) Carleman estimates for degenerate parabolic operators with applications to null controllability. J Evol Equ 6:161–204
10. Alabau-Boussouira F, Cannarsa P, Sforza D (2008) Decay estimates for second order evolution equations with memory. J Funct Anal 254(5):1342–1372
11. Ammari K, Tucsnak M (2001) Stabilization of second order evolution equations by a class of unbounded feedbacks. ESAIM Control Optim Calc Var 6:361–386
12. Ammar-Khodja F, Bader A, Benabdallah A (1999) Dynamic stabilization of systems via decoupling techniques. ESAIM Control Optim Calc Var 4:577–593
13. Barbu V (2003) Feedback stabilization of Navier–Stokes equations. ESAIM Control Optim Calc Var 9:197–206
14. Barbu V Da Prato G (1983) Hamilton Jacobi equations in Hilbert spaces. Pitman, London
15. Barbu V, Lasiecka I, Triggiani R (2006) Tangential boundary stabilization of Navier–Stokes equations. Mem Amer Math Soc 181(852):128
16. Barbu V, Triggiani R (2004) Internal stabilization of Navier–Stokes equations with finite-dimensional controllers. Indiana Univ Math J 53(5):1443–1494
17. Bardos C, Lebeau G, Rauch R (1992) Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. SIAM J Control Optim 30:1024–1065
18. Bátkai A, Engel KJ, Prüss J, Schnaubelt R (2006) Polynomial stability of operator semigroups. Math Nachr 279(13–14):1425–1440
19. Beauchard K (2005) Local controllability of a 1-D Schödinger equation. J Math Pures Appl 84(7):851–956
20. Bellman R (1957) Dynamic Programming. Princeton University Press, Princeton
21. Benabdallah A, Dermenjian Y, Le Rousseau J (2007) Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications to controllability and an inverse problem. J Math Anal Appl 336(2):865–887
22. Bensoussan A, Da Prato G, Delfour MC, Mitter SK (1993) Representation and Control of Infinite Dimensional Systems. Systems and Control: Foundations and applications, Birkhäuser, Boston
23. Beyrath A (2001) Stabilisation indirecte localement distribué de systèmes faiblement couplés. C R Acad Sci Paris Sér I Math 333(5):451–456
24. Beyrath A (2004) Indirect linear locally distributed damping of coupled systems. Bol Soc Parana Mat 22(2):17–34
25. Burq N, Gérard P (1997) Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes. C R Acad Sci Paris Sér I Math 325(7):749–752

26. Burq N, Hitrik M (2007) Energy decay for damped wave equations on partially rectangular domains. Math Res Lett 14(1):35–47
27. Burq N, Lebeau G (2001) Mesures de défaut de compacité, application au système de Lamé. Ann Sci École Norm Sup (4) 34(6):817–870
28. Cannarsa P (1989) Regularity properties of solutions to Hamilton–Jacobi equations in infinite dimensions and nonlinear optimal control. Diff Integral Equ 2:479–493
29. Cannarsa P, Da Prato G (1990) Some results on non-linear optimal control problems and Hamilton–Jacobi equations in infinite dimensions. J Funct Anal 90:27–47
30. Cannarsa P, Gozzi F, Soner HM (1991) A boundary value problem for Hamilton–Jacobi equations in Hilbert spaces. Applied Math Optim 24:197–220
31. Cannarsa P, Gozzi F, Soner HM (1993) A dynamic programming approach to nonlinear boundary control problems of parabolic type. J Funct Anal 117:25–61
32. Cannarsa P, Di Blasio G (1995) A direct approach to infinite dimensional Hamilton–Jacobi equations and applications to convex control with state constraints. Differ Integral Equ 8:225–246
33. Cannarsa P, Tessitore ME (1996) Infinite dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type. SIAM J Control Optim 34:1831–1847
34. Cannarsa P, Komornik V, Loreti P (2002) One-sided and internal controllability of semilinear wave equations with infinitely iterated logarithms. Discret Contin Dyn Syst 8(3):745–756
35. Cannarsa P, Martinez P, Vancostenoble J (2004) Persistent regional null controllability for a class of degenerate parabolic equations. Commun Pure Appl Anal 3(4):607–635
36. Cannarsa P, Martinez P, Vancostenoble (2005) Null Controllability of degenerate heat equations. Adv Differ Equ 10(2):153–190
37. Cavalcanti MM, Oquendo HP (2003) Frictional versus viscoelastic damping in a semilinear wave equation. SIAM J Control Optim 42:1310–1324
38. Chen G (1981) A note on boundary stabilization of the wave equation. SIAM J Control Optim 19:106–113
39. Chueshov I, Lasiecka I, Toundykov D (2008) Long-term dynamics of semilinear wave equation with nonlinear localized interior damping and a source term of critical exponent. Discret Contin Dyn Syst 20(3):459–509
40. Conrad F, Rao B (1993) Decay of solutions of the wave equation in a star-shaped domain with nonlinear boundary feedback. Asymptot Anal 7:159–177
41. Conrad F, Pierre (1994) Stabilization of second order evolution equations by unbounded nonlinear feedback. Ann Inst H Poincaré Anal Non Linéaire 11(5):485–515, Asymptotic Anal 7:159–177
42. Coron JM (1992) Global asymptotic stabilization for controllable systems without drift. Math Control Signals Syst 5(3):295–312
43. Coron JM (2007) Control and nonlinearity. Mathematical surveys and monographs vol 136, Providence, RI: xiv+426
44. Coron JM, Trélat E (2004) Global steady-state controllability of one-dimensional semilinear heat equations. SIAM J Control Optim 43(2):549–569
45. Crandall MG, Lions PL (1985) Hamilton Jacobi equation in infinite dimensions I: Uniqueness of viscosity solutions. J Funct Anal 62:379–396
46. Crandall MG, Lions PL (1986) Hamilton Jacobi equation in infinite dimensions II: Existence of viscosity solutions. J Funct Anal 65:368–425
47. Crandall MG, Lions PL (1986) Hamilton Jacobi equation in infinite dimensions III. J Funct Anal 68:214–247
48. Crandall MG, Lions PL (1990) Hamilton Jacobi equation in infinite dimensions IV: Hamiltonians with unbounded linear terms. J Funct Anal 90:237–283
49. Crandall MG, Lions PL (1991) Hamilton Jacobi equation in infinite dimensions V: Unbounded linear terms and *B*-continuous solutions. J Funct Anal 97:417–465
50. Curtain RF, Weiss G (1989) Well posedness of triples of operators (in the sense of linear systems theory). Control and Estimation of Distributed Parameter Systems (Vorau, 1988). Internat. Ser. Numer. Math., vol. 91, Birkhäuser, Basel
51. Curtain RF, Zwart H (1995) An introduction to infinite-dimensional linear systems theory. Texts in Applied Mathematics, vol 21. Springer, New York
52. Da Prato G, Frankowska H (2007) Stochastic viability of convex sets. J Math Anal Appl 333(1):151–163
53. Dafermos CM(1970) Asymptotic stability inviscoelasticity. Arch Ration Mech Anal 37:297–308
54. Dafermos CM (1970) An abstract Volterra equation with applications to linear viscoelasticity. J Differ Equ 7:554–569
55. Engel KJ, R Nagel R (2000) One-parameter semigroups for linear evolution equations. Springer, New York
56. Eller M, Lagnese JE, Nicaise S (2002) Decay rates for solutions of a Maxwell system with nonlinear boundary damping. Comp Appl Math 21:135–165
57. Fabre C, Puel J-P, Zuazua E (1995) Approxiamte controllability of the semilinear heat equation. Proc Roy Soc Edinburgh Sect A 125(1):31–61
58. Fattorini HO, Russell DL (1971) Exact controllability theorems for linear parabolic equations in one space dimension. Arch Rat Mech Anal 4:272–292
59. Fattorini HO (1998) Infinite Dimensional Optimization and Control theory. Encyclopedia of Mathematics and its Applications, vol 62. Cambridge University Press, Cambridge
60. Fattorini HO (2005) Infinite dimensional linear control systems. North-Holland Mathematics Studies, vol 201. Elsevier Science B V, Amsterdam
61. Fernández-Cara E, Zuazua E (2000) Null and approximate controllability for weakly blowing up semilinear heat equations. Ann Inst H Poincaré Anal Non Linéaire 17(5):583–616
62. Fernández-Cara E, Zuazua E (2000) The cost approximate controllability for heat equations: The linear case. Adv Differ Equ 5:465–514
63. Fernández-Cara E, Guerrero S, Imanuvilov OY, Puel J-P (2004) Local exact controllability of the Navier–Stokes system. J Math Pures Appl 83(9–12):1501–1542
64. Fursikov A (2000) Optimal control of distributed systems. Theory and applications. Translation of Mathematical Monographs, vol 187. American Mathematical Society, Providence
65. Fursikov A, Imanuvilov OY (1996) Controllability of evolution equations. Lecture Notes, Research Institute of Mathematics, Seoul National University, Seoul
66. Gibson JS (1980) A note on stabilization of infinite dimensional linear oscillators by compact linear feedback. SIAM J Control Optim 18:311–316

67. Giorgi C, Naso MG, Pata V (2005) Energy decay of electromagnetic systems with memory. Math Models Methods Appl Sci 15(10):1489–1502
68. Glass O (2000) Exact boundary controllability of 3-D Euler equation. ESAIM Control Optim Calc Var 5:1–44
69. Glass O (2007) On the controllability of the 1-D isentropic Euler equation. J Eur Math Soc (JEMS) 9(3):427–486
70. Guerrero S (2007) Local controllability to the trajectories of the Navier–Stokes system with nonlinear Navier-slip boundary conditions. ESAIM Control Optim Calc Var 12(3):484–544
71. Haraux A (1978) Semi-groupes linéaires et équations d'évolution linéaires périodiques. Publication du Laboratoire d'Analyse Numérique no 78011. Université Pierre et Marie Curie, Paris
72. Haraux A (1989) Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps. Portugal Math 46(3):245–258
73. Ho LF (1986) Observabilité frontière de l'équation des ondes. C R Acad Sci Paris Sér I Math 302(12):443–446
74. Komornik V (1994) Exact Controllability and Stabilization. The Multiplier Method. Collection RMA, vol 36. Masson–John Wiley, Paris–Chicester
75. Komornik V, Loreti P (2005) Fourier series in control theory. Springer, New York
76. Komornik V, Zuazua E (1990) A direct method for the boundary stabilization of the wave equation. J Math Pures Appl 69:33–54
77. Lasiecka I, Lions J-L, Triggiani R (1986) Nonhomogeneous boundary value problems for second order hyperbolic operators. J Math Pures Appl 65(2):149–192
78. Lasiecka I, Tataru D (1993) Uniform boundary stabilization of semilinear wave equation with nonlinear boundary damping. Differ Integral Equ 8:507–533
79. Lasiecka I, Triggiani R (1991) Differential and algebraic Riccati equations with application to boundary/point control problems: continuous theory and approximation theory. Lecture Notes in Control & Inform Sci, vol 164. Springer, Berlin
80. Lasiecka I, Triggiani R (2000) Control theory for partial differential equations: continuous and approximation theories. I. Encyclopedia of Mathematics and its Applications, vol 74. Cambridge University Press, Cambridge
81. Lasiecka I, Triggiani R (2000) Control theory for partial differential equations: continuous and approximation theories. II. Encyclopedia of Mathematics and its Applications, vol 75. Cambridge University Press, Cambridge
82. Le Rousseau J (2007) Carleman estimates and controllability results for the one-dimensional heat equation with BV coefficients. J Differ Equ 233(2):417–447
83. Lebeau G, Robbiano L (1995) Exact control of the heat equation. Comm Partial Differ Equ 20(1–2):335–356
84. Li X, Yong J (1995) Optimal control of infinite dimensional systems. Systems & Control: Foundations & Applications. Birkhäuser, Boston
85. Lions J-L (1971) Optimal control of systems governed by partial differential equations. Springer, New-York
86. Lions J-L (1988) Contrôlabilité exacte et stabilisation de systèmes distribués I-II. Masson, Paris
87. Liu K (1997) Locally distributed control and damping for the conservative systems. SIAM J Control Optim 35:1574–1590
88. Liu Z, Zheng S (1999) Semigroups associated with dissipative systems. Chapman Hall CRC Research Notes in Mathematics, vol 398. Chapman Hall/CRC, Boca Raton
89. Liu WJ, Zuazua E (1999) Decay rates for dissipative wave equations. Ric Mat 48:61–75
90. Liu Z, Rao R (2007) Frequency domain approach for the polynomial stability of a system of partially damped wave equations. J Math Anal Appl 335(2):860–881
91. Londen SO, Petzeltová H, Prüss J (2003) Global well-posedness and stability of a partial integro-differential equation with applications to viscoelasticity. J Evol Equ 3(2):169–201
92. Loreti P, Rao B (2006) Optimal energy decay rate for partially damped systems by spectral compensation. SIAM J Control Optim 45(5):1612–1632
93. Martinez P (1999) A new method to obtain decay rate estimates for dissipative systems with localized damping. Rev Mat Complut 12:251–283
94. Martinez P, Raymond J-P, Vancostenoble J (2003) Regional null controllability for a linearized Crocco type equation. SIAM J Control Optim 42(2):709–728
95. Miller L (2002) Escape function conditions for the observation, control, and stabilization of the wave equation. SIAM J Control Optim 41(5):1554–1566
96. Muñoz Rivera JE, Peres Salvatierra A (2001) Asymptotic behaviour of the energy in partially viscoelastic materials. Quart Appl Math 59:557–578
97. Muñoz Rivera JE (1994) Asymptotic behaviour in linear viscoelasticity. Quart Appl Math 52:628–648
98. Nakao M (1996) Decay of solutions of the wave equation with a local nonlinear dissipation. Math Ann 305:403–417
99. Pazy A (1968) Semigroups of linear operators and applications to partial differential equations. Springer Berlin
100. Pontryagin LS (1959) Optimal regulation processes. Uspehi Mat Nauk 14(1):3–20
101. Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko EF (1962) The mathematical theory of optimal processes. Interscience Publishers John Wiley & Sons, Inc., New York–London, translated from the Russian by Trirogoff KN, edited by Neustadt LW
102. Propst G, Prüss J (1996) On wave equation with boundary dissipation of memory type. J Integral Equ Appl 8:99–123
103. Prüss J (1993) Evolutionary integral equations and applications. Monographs in Mathematics, vol 87. Birkhäuser Verlag, Basel
104. Raymond J-P (2006) Feedback boundary stabilization of the two-dimensional Navier–Stokes equations. SIAM J Control Optim 45(3):790–828
105. Rosier L (1997) Exact boundary controllability for the Korteweg–de Vries equation on a bounded domain. ESAIM Control Optim Calc Var 2:33–55
106. Rosier L (2000) Exact boundary controllability for the linear Korteweg–de Vries equation on the half-line. SIAM J Control Optim 39(2):331–351
107. Rosier L, Zhang BY (2006) Global stabilization of the generalized Korteweg–de Vries equation posed on a finite domain. SIAM J Control Optim 45(3):927–956
108. Russell DL (1978) Controllability and stabilizability theorems for linear partial differential equations: recent progress and open questions. SIAM Rev 20(4):639–739
109. Russell DL (1993) A general framework for the study of indirect damping mechanisms in elastic systems. J Math Anal Appl 173(2):339–358

110. Salamon D (1987) Infinite-dimensional linear systems with unbounded control and observation: a functional analytic approach. Trans Am Math Soc 300(2):383–431
111. Shimakura N (1992) Partial differential operators of elliptic type. Translations of Mathematical Monographs, vol 99-American Mathematical Society, Providence
112. Tataru D (1992) Viscosity solutions for the dynamic programming equations. Appl Math Optim 25:109–126
113. Tataru D (1994) A-priori estimates of Carleman's type in domains with boundary. J Math Pures Appl 75:355–387
114. Tataru D (1995) Boundary controllability for conservative P.D.E. Appl Math Optim 31:257–295
115. Tataru D (1996) Carleman estimates and unique continuation near the boundary for P.D.E.'s. J Math Pures Appl 75:367–408
116. Tataru D (1997) Carleman estimates, unique continuation and controllability for anizotropic PDE's. Optimization methods in partial differential equations, South Hadley MA 1996. Contemp Math vol 209. Am Math Soc pp 267–279, Providence
117. de Teresa L (2000) Insensitizing controls for a semilinear heat equation. Comm Partial Differ Equ 25:39–72
118. Vancostenoble J, Martinez P (2000) Optimality of energy estimates for the wave equation with nonlinear boundary velocity feedbacks. SIAM J Control Optim 39:776–797
119. Vancostenoble J (1999) Optimalité d'estimation d'énergie pour une équation des ondes amortie. C R Acad Sci Paris, 328, série I, pp 777–782
120. Vitillaro E (2002) Global existence for the wave equation with nonlinear boundary damping and source terms. J Differ Equ 186(1):259–298
121. Zabczyk J (1992) Mathematical control theory: an introduction. Birkhäuser, Boston
122. Zuazua E (1989) Uniform stabilization of the wave equation by nonlinear feedbacks. SIAM J Control Optim 28:265–268
123. Zuazua E (1990) Exponential decay for the semilinear wave equation with locally distributed damping. Comm Partial Differ Equ 15:205–235
124. Zuazua E (2006) Control and numerical approximation of the heat and wave equations. In: Sanz-Solé M, Soria J, Juan Luis V, Verdera J (eds) Proceedings of the International Congress of Mathematicians, vol I, II, III, European Mathematical Society, Madrid, pp 1389–1417
125. Zuazua E (2006) Controllability and observability of partial differential equations: Some results and open problems. In: Dafermos CM, Feireisl E (eds) Handbook of differential equations: evolutionary differential equations, vol 3. Elsevier/North-Holland, Amsterdam, pp 527–621

Cooperative Games

ROBERTO SERRANO[1,2]
[1] Department of Economics, Brown University, Providence, USA
[2] IMDEA-Social Sciences, Madrid, Spain

Article Outline

Glossary

Game theory Discipline that studies strategic situations.

Cooperative game Strategic situation involving coalitions, whose formation assumes the existence of binding agreements among players.

Characteristic or coalitional function The most usual way to represent a cooperative game.

Solution concept Mapping that assigns predictions to each game.

Core Solution concept that assigns the set of payoffs that cannot be improved upon by any coalition.

Shapley value Solution concept that assigns the average of marginal contributions to coalitions.

Definition of the Subject

Cooperative game theory It is one of the two counterparts of game theory. It studies the interactions among coalitions of players. Its main question is this: Given the sets of feasible payoffs for each coalition, what payoff will be awarded to each player? One can take a positive or normative approach to answering this question, and different solution concepts in the theory lean towards one or the other.

Core It is a solution concept that assigns to each cooperative game the set of payoffs that no coalition can improve upon or block. In a context in which there is unfettered coalitional interaction, the core arises as a good positive answer to the question posed in cooperative game theory. In other words, if a payoff does not belong to the core, one should not expect to see it as the prediction of the theory if there is full cooperation.

Shapley value It is a solution that prescribes a single payoff for each player, which is the average of all marginal contributions of that player to each coalition he or she is a member of. It is usually viewed as a good normative answer to the question posed in cooperative game theory. That is, those who contribute more to the groups that include them should be paid more.

Although there were some earlier contributions, the official date of birth of game theory is usually taken to be 1944, year of publication of the first edition of the *Theory* *of Games and Economic Behavior*, by John von Neumann and Oskar Morgenstern [42]. The core was first proposed by Francis Ysidro Edgeworth in 1881 [13], and later reinvented and defined in game theoretic terms in [14]. The Shapley value was proposed by Lloyd Shapley in his 1953 Ph D dissertation [37]. Both the core and the Shapley value have been applied widely, to shed light on problems in different disciplines, including economics and political science.

Introduction

Game theory is the study of games, also called strategic situations. These are decision problems with multiple decision makers, whose decisions impact one another. It is divided into two branches: non-cooperative game theory and cooperative game theory. The actors in non-cooperative game theory are individual players, who may reach agreements only if they are self-enforcing. The non-cooperative approach provides a rich language and develops useful tools to analyze games. One clear advantage of the approach is that it is able to model how specific details of the interaction among individual players may impact the final outcome. One limitation, however, is that its predictions may be highly sensitive to those details. For this reason it is worth also analyzing more abstract approaches that attempt to obtain conclusions that are independent of such details. The cooperative approach is one such attempt, and it is the subject of this article.

The actors in cooperative game theory are coalitions, that is, groups of players. For the most part, two facts, that coalitions can form and that each coalition has a feasible set of payoffs available to its members, are taken as given. Given the coalitions and their sets of feasible payoffs as primitives, the question tackled is the identification of final payoffs awarded to each player. That is, given a collection of feasible sets of payoffs, one for each coalition, can one predict or recommend a payoff (or set of payoffs) to be awarded to each player? Such predictions or recommendations are embodied in different solution concepts.

Indeed, one can take several approaches to answering the question just posed. From a positive or descriptive point of view, one may want to get a prediction of the likely outcome of the interaction among the players, and hence, the resulting payoff be understood as the natural consequence of the forces at work in the system. Alternatively, one can take a normative or prescriptive approach, set up a number of normative goals, typically embodied in axioms, and try to derive their logical implications. Although authors sometimes disagree on the classification of the different solution concepts according to these two criteria – as we shall see, the understanding of each solution concept is enhanced if one can view it from very distinct approaches –, in this article we shall exemplify the positive approach with the core and the normative approach with the Shapley value. While this may oversimplify the issues, it should be helpful to a reader new to the subject.

The rest of the article is organized as follows. Sect. "Cooperative Games" introduces the basic model of a cooperative game, and discusses its assumptions as well as the notion of solution concepts. Sect. "The Core" is devoted to the core, and Sect. "The Shapley Value" to the Shapley value. In each case, some of the main results for each of the two are described, and examples are provided. Sect. "Future Directions" discusses some directions for future research.

Cooperative Games

Representations of Games. The Characteristic Function

Let us begin by presenting the different ways to describe a game. The first two are the usual ways employed in non-cooperative game theory.

The most informative way to describe a game is called its *extensive form*. It consists of a game tree, specifying the timing of moves for each player and the information available to each of them at the time of making a move. At the end of each path of moves, a final outcome is reached and a payoff vector is specified. For each player, one can define a *strategy*, i. e., a complete contingent plan of action to play the game. That is, a strategy is a function that specifies a feasible move each time a player is called upon to make a move in the game.

One can abstract from details of the interaction (such as timing of moves and information available at each move), and focus on the concept of strategies. That is, one can list down the set of strategies available to each player, and arrive at the *strategic* or *normal form* of the game. For two players, for example, the normal form is represented in a bimatrix table. One player controls the rows, and the other the columns. Each cell of the bimatrix is occupied with an ordered pair, specifying the payoff to each player if each of them chooses the strategy corresponding to that cell.

One can further abstract from the notion of strategies, which will lead to the *characteristic function form* of representing a game. From the strategic form, one makes assumptions about the strategies used by the complement of a coalition of players to determine the feasible payoffs for the coalition (see, for example, the derivations in [7,42]). This is the representation most often used in cooperative game theory.

Thus, here are the primitives of the basic model in cooperative game theory. Let $N = \{1, \dots, n\}$ be a finite set of players. Each non-empty subset of N is called a *coalition*. The set N is referred to as the *grand coalition*. For each coalition S, we shall specify a set $V(S) \subset \mathbb{R}^{|S|}$ containing $|S|$-dimensional payoff vectors that are feasible for coalition S. This is called the characteristic function, and the pair (N, V) is called a *cooperative game*. Note how a reduced form approach is taken because one does not explain what strategic choices are behind each of the payoff vectors in $V(S)$. In addition, in this formulation, it is implicitly assumed that the actions taken by the complement coalition (those players in $N \setminus S$) cannot prevent S from achieving each of the payoff vectors in $V(S)$. There are more general models in which these sorts of externalities across coalitions are considered, but we shall ignore them in this article.

Assumptions on the Characteristic Function

Some of the most common technical assumptions made on the characteristic function are the following:

(1) For each $S \subseteq N$, $V(S)$ is closed. Denote by $\partial V(S)$ the boundary of $V(S)$. Hence, $\partial V(S) \subseteq V(S)$.

(2) For each $S \subseteq N$, $V(S)$ is comprehensive, i. e., for each $x \in V(S)$, $\{x\} - \mathbb{R}^{|S|}_{+} \subseteq V(S)$.

(3) For each $x \in \mathbb{R}^{|S|}$,

$$\partial V(S) \cap \left(\{x\} + \mathbb{R}^{|S|}_{+}\right)$$

is bounded.

(4) For each $S \subseteq N$, there exists a continuously differentiable representation of $V(S)$, i. e., a continuously differentiable function $g_S : \mathbb{R}^{|S|} \to \mathbb{R}$ such that

$$V(S) = \{x \in \mathbb{R}^{|S|} | g_S(x) \leq 0\} .$$

(5) For each $S \subseteq N$, $V(S)$ is non-leveled, i. e., for every $x \in \partial V(S)$, the gradient of g_S at x is positive in all its coordinates.

With the assumptions made, $\partial V(S)$ is its Pareto frontier, i. e., the set of vectors $x_S \in V(S)$ such that there does not exist $y_S \in V(S)$ satisfying that $y_i \geq x_i$ for all $i \in S$ with at least one strict inequality.

Other assumptions usually made relate the possibilities available to different coalitions. Among them, a very important one is *balancedness*, which we shall define next:

A collection $\mathcal{T}$ of coalitions is balanced if there exists a set of weights $w(S) \in [0, 1]$ for each $S \in \mathcal{T}$ such that for every $i \in N$, $\sum_{S \in \mathcal{T}, S \supseteq \{i\}} w(S) = 1$. One can think of these weights as the fraction of time that each player devotes to each coalition he is a member of, with a given coalition representing the same fraction of time for each player. The game (N, V) is balanced if $x_N \in V(N)$ whenever $(x_S) \in V(S)$ for every S in a balanced collection $\mathcal{T}$. That is, the grand coalition can always implement any "time-sharing arrangement" that the different subcoalitions may come up with.

The characteristic function defined so far is often referred to as a *non-transferable utility (NTU)* game. A particular case is the *transferable utility (TU)* game case, in which for each coalition $S \subseteq N$, there exists a real number $v(S)$ such that

$$V(S) = \left\{ x \in \mathbb{R}^{|S|} : \sum_{i \in S} x_i \leq v(S) \right\} .$$

Abusing notation slightly, we shall denote a TU game by (N, v). In the TU case there is an underlying nummeraire – money – that can transfer utility or payoff at a one-to-one rate from one player to any other. Technically, the theory of NTU games is far more complex: it uses convex analysis and fixed point theorems, whereas the TU theory is based on linear inequalities and combinatorics.

Solution Concepts

Given a characteristic function, i. e., a collection of sets $V(S)$, one for each S, the theory formulates its predictions on the basis of different *solution concepts*. We shall concentrate on the case in which the grand coalition forms, that is, cooperation is totally successful. Of course, solution concepts can be adapted to take care of the case in which this does not happen.

A *solution* is a mapping that assigns a set of payoff vectors in $V(N)$ to each characteristic function game (N, V). Thus, a solution in general prescribes a set, which can be empty, or a singleton (when it assigns a unique payoff vector as a function of the fundamentals of the problem). The leading set-valued cooperative solution concept is the core, while one of the most used single-valued ones is the Shapley value for TU games.

There are several criteria to evaluate the reasonableness or appeal of a cooperative solution. As outlined above, in a normative approach, one can propose axioms, abstract principles that one would like the solution to satisfy, and the next step is to pursue their logical consequences. Historically, this was the first argument to justify the Shapley value. Alternatively, one could start by defending a solution on the basis of its definition alone. In the case of the core, this will be especially natural: in a context in which players can freely get together in groups, the prediction should be payoff vectors that cannot be improved upon by any coalition. One can further enhance one's positive understanding of the solution concept by proposing games in extensive form or in normal form played non-cooperatively by players whose self-enforcing agreements lead to a given solution. This is simply to provide non-cooperative foundations or non-cooperative implementation to the cooperative solution in question, and it is an important research agenda initiated by John Nash in [25], referred to as the Nash program (see [34] for a recent survey). Today, there are interesting results of these different kinds for many solution concepts, which include axiomatic characterizations and non-cooperative foundations. Thus, one can evaluate the appeal of the axioms and the non-cooperative procedures behind each solution to defend a more normative or positive interpretation in each case.

The Core

The idea of agreements that are immune to coalitional deviations was first introduced to economic theory by Edgeworth in [13], which defined the set of coalitionally stable allocations of an economy under the name "final settlements." Edgeworth envisioned this concept as an alternative to *competitive equilibrium* [43], of central importance in economic theory, and was also the first to investigate the connections between the two concepts. Edgeworth's notion, which today we refer to as the *core*, was rediscovered and introduced to game theory in [14]. The origins of the core were not axiomatic. Rather, its simple and appealing definition appropriately describes stable outcomes in a context of unfettered coalitional interaction.

The core of the game (N, V) is the set of payoff vectors

$$C(N, V) = \{x \in V(N): \ \nexists S \subseteq N, x_S \in V(S) \setminus \partial V(S)\}\,.$$

In words, it is the set of feasible payoff vectors for the grand coalition that no coalition can upset. If such a coalition S exists, we shall say that S can improve upon or block x, and x is deemed unstable. That is, in a context where any coalition can get together, when S has a blocking move, coalition S will form and abandon the grand coalition and its payoffs x_S in order to get to a better payoff for each of the members of the coalition, a plan that is feasible for them.

Non-Emptiness

The core can prescribe the empty set in some games. A game with an empty core is to be understood as a situation of strong instability, as any payoffs proposed to the grand coalition are vulnerable to coalitional blocking.

Example Consider the following simple majority 3-player TU game, in which the votes of at least two players makes the coalition winning. That is, we represent the situation by the following characteristic function: $v(S) = 1$ for any S containing at least two members, $v(\{i\}) = 0$ for all $i \in N$. Clearly, $C(N, v) = \emptyset$. Any feasible payoff agreement proposed to the grand coalition will be blocked by at least one coalition.

An important sufficient condition for the non-emptiness of the core of NTU games is balancedness, as shown in [32]:

Theorem 1 (Scarf [32]) *Let the game (N, V) be balanced. Then $C(N, V) \neq \emptyset$.*

For the TU case, balancedness is not only sufficient, but it becomes also necessary for the non-emptiness of the core:

Theorem 2 (Bondareva [9]; Shapley [39]) *Let (N, v) be a TU game. Then, (N, v) is balanced if and only if $C(N, v) \neq \emptyset$.*

The Connections with Competitive Equilibrium

In economics, the institution of markets and the notion of prices are essential to the understanding of the allocation of goods and the distribution of wealth among individuals. For simplicity in the presentation, we shall concentrate on exchange economies, and disregard production aspects. That is, we shall assume that the goods in question have already been produced in some fixed amounts, and now they are to be allocated to individuals to satisfy their consumption needs.

An *exchange economy* is a system in which each agent i in the set N has a consumption set $Z_i \subseteq \mathbb{R}^l_+$ of commodity bundles, as well as a preference relation over Z_i and an initial endowment $\omega_i \in Z_i$ of the commodities. A feasible *allocation* of goods in the economy is a list of bundles $(z_i)_{i \in N}$ such that $z_i \in Z_i$ and $\sum_{i \in N} z_i \leq \sum_{i \in N} \omega_i$. An allocation is *competitive* if it is supported by a *competitive equilibrium*. A competitive equilibrium is a price-allocation pair $(p, (z_i)_{i \in N})$, where $p \in \mathbb{R}^l \setminus \{0\}$ is such that

- for every $i \in N$, z_i is top-ranked for agent i among all bundles z satisfying that $pz \leq p\omega_i$,
- and $\sum_{i \in N} z_i = \sum_{i \in N} \omega_i$.

In words, this is what the concept expresses. First, at the equilibrium prices, each agent demands z_i, i. e., wishes to purchase this bundle among the set of affordable bundles, the budget set. And second, these demands are such that all markets clear, i. e., total demand equals total supply.

Note how the notion of a competitive equilibrium relies on the principle of private ownership (each individual owns his or her endowment, which allows him or her to access markets and purchase things). Moreover, each agent is a price-taker in all markets. That is, no single individual can affect the market prices with his or her actions; prices are fixed parameters in each individual's consumption decision. The usual justification for the price-taking assumption is that each individual is "very small" with respect to the size of the economy, and hence, has no market power.

One difficulty with the competitive equilibrium concept is that it does not explain where prices come from. There is no single agent in the model responsible for coming up with them. Walras in [43] told the story of an auctioneer calling out prices until demand and supply coincide, but in many real-world markets there is no auctioneer. More generally, economists attribute the equilibrium prices to the workings of the forces of demand and supply, but this appears to be simply repeating the definition. So, is there a different way one can explain competitive equilibrium prices?

As it turns out, there is a very robust result that answers this question. We refer to it as the *equivalence principle* (see, e. g., [6]), by which, under certain regularity conditions, the predictions provided by different game-theoretic solution concepts, when applied to an economy with a large enough set of agents, tend to converge to the set of competitive equilibrium allocations. One of the first results in this tradition was provided by Edgeworth in 1881 for the core. Note how the core of the economy can be defined in the space of allocations, using the same definition as above. Namely, a feasible allocation is in the core if it cannot be blocked by any coalition of agents when making use of the coalition's endowments.

Edgeworth's result was generalized later by Debreu and Scarf in [11] for the case in which an exchange economy is replicated an arbitrary number of times (Anderson studies in [1] the more general case of arbitrary sequences of economies, not necessarily replicas). An informal statement of the Debreu–Scarf theorem follows:

Theorem 3 (Debreu and Scarf [11]) *Consider an exchange economy. Then,*

- *(i) The set of competitive equilibrium allocations is contained in the core.*
- *(ii) For each non-competitive core allocation of the original economy, there exists a sufficiently large replica of the economy for which the replica of the allocation is blocked.*

The first part states a very appealing property of competitive allocations, i. e., their coalitional stability. The second part, known as the core convergence theorem, states that the core "shrinks" to the set of competitive allocations as the economy grows large.

In [3], Aumann models the economy as an atomless measure space, and demonstrates the following core equivalence theorem:

Theorem 4 (Aumann [3]) *Let the economy consists of an atomless continuum of agents. Then, the core coincides with the set of competitive allocations.*

For readers who wish to pursue the topic further, [2] provides a recent survey.

Axiomatic Characterizations

The axiomatic foundations of the core were provided much later than the concept was proposed. These characterizations are all inspired by Peleg's work. They include [26,27], and [36] – the latter paper also provides an axiomatization of competitive allocations in which core convergence insights are exploited.

In all these characterizations, the key axiom is that of *consistency*, also referred to as the reduced game property. Consistency means that the outcomes prescribed by a solution should be "invariant" to the number of players in the game. More formally, let (N, V) be a game, and let σ be a solution. Let $x \in \sigma(N, V)$. Then, the solution is consistent if for every $S \subseteq N$, $x_S \in \sigma(S, V_{xS})$, where (S, V_{xS}) is the reduced game for S given payoffs x, defined as follows. The feasible set for S in this reduced game is the projection of $V(N)$ at $x_{N\setminus S}$, i. e., what remains after paying those outside of S:

$$V_{xS}(S) = \{y_S : (y_S, x_{N\setminus S}) \in V(N)\} .$$

However, the feasible set of $T \subset S$, $T \neq S$, allows T to make deals with any coalition outside of S, provided that those services are paid at the rate prescribed by $x_{N\setminus S}$:

$$V_{xS}(T) = \{y_T \in \cup_{Q \subseteq N\setminus S}(y_T, x_Q) \in V(T \cup Q)\} .$$

It can be shown that the core satisfies consistency with respect to this reduced game. Moreover, consistency is the central axiom in the characterization of the core, which, depending on the version one looks at, uses a host of other axioms; see [26,27,36].

Non-cooperative Implementation

To obtain a non-cooperative implementation of the core, the procedure must embody some feature of anonymity, since the core is usually a large set and it contains payoffs where different players are treated very differently. For instance, if the procedure always had a fixed set of moves, typically the prediction would favor the first mover, making it impossible to obtain an implementation of the entire set of payoffs.

The model in [30] builds in this anonymity by assuming that negotiations take place in continuous time, so that anyone can speak at the beginning of the game, and at any point in time, instead of having a fixed order. The player that gets to speak first makes a proposal consisting of naming a coalition that contains him and a feasible payoff for that coalition. Next, the players in that coalition get to respond. If they all accept the proposal, the coalition leaves and the game continues among the other players. Otherwise, a new proposal may come from any player in N. It is shown that, if the TU game has a non-empty core (as well as any of its subgames), a class of stationary self-enforcing predictions of this procedure coincide with the core. If a core payoff is proposed to the grand coalition, there are no incentives for individual players to reject it. Conversely, a non-core payoff cannot be sustained because any player in a blocking coalition has an incentive to make a proposal to that coalition, who will accept it (knowing that the alternative, given stationarity, would be to go back to the non-core status quo). [24] offers a discrete-time version of the mechanism: in this work, the anonymity required is imposed on the solution concept, by looking at the order-independent equilibria of the procedure.

The model in [33] sets up a market to implement the core. The anonymity of the procedure stems from the random choice of broker. The broker announces a vector $(x_1, \ldots, x_n)$, where the components add up to $v(N)$. One can interpret x_i as the price for the productive asset held by player i. Following an arbitrary order, the remaining players either accept or reject these prices. If player i accepts, he sells his asset to the broker for the price x_i and leaves the game. Those who reject get to buy from the broker, at the called out prices, the portfolio of assets of their choice if the broker still has them. If a player rejects, but does not get to buy the portfolio of assets he would like because someone else took them before, he can always leave the market with his own asset. The broker's payoff is the worth of the final portfolio of assets that he holds, plus the net monetary transfers that he has received. It is shown in [33] that the prices announced by the broker will always be his top-ranked vectors in the core. If the TU game is such that gains from cooperation increase with the size of coalitions, a beautiful theorem of Shapley in [41] is used to prove that the set of all equilibrium payoffs of this procedure will coincide with the core. Core payoffs are here understood as those price vectors where all arbitrage opportunities in the market have been wiped out. Also, procedures in [35] implement the core, but do not rely on the TU assumption, and they use a procedure in which the order of moves can be endogenously changed by players. Finally, yet another way to build anonymity in the procedure is by allowing the proposal to be made by brokers outside of the set N, as done in [28].

An Application

Consider majority games within a parliament. Suppose there are 100 seats, and decisions are made by simple majority so that 51 votes are required to pass a piece of legislation.

In the first specification, suppose there is a very large party – player 1 –, who has 90 seats. There are five small parties, with 2 seats each. Given the simple majority rules, this problem can be represented by the following TU characteristic function: $v(S) = 1$ if S contains player 1, and $v(S) = 0$ otherwise. The interpretation is that each winning coalition can get the entire surplus – pass the desired proposal. Here, a coalition is winning if and only if player 1 is in it. For this problem, the core is a singleton: the entire unit of surplus is allocated to player 1, who has all the power. Any split of the unit surplus of the grand coalition ($v(N) = 1$) that gives some positive fraction of surplus to any of the small parties can be blocked by the coalition of player 1 alone.

Consider now a second problem, in which player 1, who continues to be the large party, has 35 seats, and each of the other five parties has 13 seats. Now, the characteristic function is as follows: $v(S) = 1$ if and only if S either contains player 1 and two small parties, or it contains four of the small parties; $v(S) = 0$ otherwise. It is easy to see that now the core is empty: any split of the unit surplus will be blocked by at least one coalition. For example, the entire unit going to player 1 is blocked by the coalition of all five small parties, which can award 0.2 to each of them. But this arrangement, in which each small party gets 0.2 and player 1 nothing, is blocked as well, because player 1 can bribe two of the small parties (say, players 2 and 3) and promise them 1/3 each, keeping the other third for itself, and so on. The emptiness of the core is a way to describe the fragility of any agreement, due to the inherent instability of this coalition formation game.

The Shapley Value

Now consider a transferable utility or TU game in characteristic function form. The number $v(S)$ is referred to as the worth of S, and it expresses S's initial position (e. g., the maximum total amount of surplus in nummeraire – money, or power – that S initially has at its disposal.

Axiomatics

Shapley in [37] is interested in solving in a fair and unique way the problem of distribution of surplus among the players, when taking into account the worth of each coalition. To do this, he restricts attention to single-valued solutions and resorts to the axiomatic method. He proposes the following axioms on a single-valued solution:

(i) Efficiency: The payoffs must add up to $v(N)$, which means that all the grand coalition surplus is allocated.
(ii) Symmetry: If two players are substitutes because they contribute the same to each coalition, the solution should treat them equally.
(iii) Additivity: The solution to the sum of two TU games must be the sum of what it awards to each of the two games.
(iv) Dummy player: If a player contributes nothing to every coalition, the solution should pay him nothing.

(To be precise, the name of the first axiom should be different. In an economic sense, the statement does imply efficiency in superadditive games, i. e., when for every pair of disjoint coalitions S and T, $v(S) + v(T) \leq v(S \cup T)$. In the absence of superadditivity, though, forming the grand coalition is not necessarily efficient, because a higher aggregate payoff can be obtained from a different coalition structure.)

The surprising result in [37] is this:

Theorem 5 (Shapley [37]) *There is a unique single-valued solution to TU games satisfying efficiency, symmetry, additivity and dummy. It is what today we call the Shapley value, the function that assigns to each player i the payoff*

$$Sh_i(N,v) = \sum_{S, i \in S} \frac{(|S|-1)!(|N|-|S|)!}{|N|!}[v(S) - v(S \setminus \{i\})] .$$

That is, the Shapley value awards to each player the average of his marginal contributions to each coalition. In taking this average, all orders of the players are considered to be equally likely. Let us assume, also without loss of generality, that $v(\{i\}) = 0$ for each player i.

What is especially surprising in Shapley's result is that nothing in the axioms (with the possible exception of the dummy axiom) hints at the idea of marginal contributions, so marginality in general is the outcome of all the axioms, including additivity or linearity. Among the axioms utilized by Shapley, additivity is the one with a lower normative content: it is simply a mathematical property to justify simplicity in the computation of the solution. Young in [45] provides a beautiful counterpart to Shapley's theorem. He drops additivity (as well as the dummy player axiom), and instead, uses an axiom of marginality. Marginality means that the solution should pay the same to a pleyar in two games if his or her marginal contributions to coalitions is the same in both games. Marginality is an idea with a strong tradition in economic theory. Young's result is "dual" to Shapley's, in the sense that marginality is assumed and additivity derived as the result:

Theorem 6 (Young [45]) *There exists a unique single-valued solution to TU games satisfying efficiency, symmetry and marginality. It is the Shapley value.*

Apart from these two, [19] provides further axiomatizations of the Shapley value using the idea of potential and the concept of consistency, as described in the previous section.

There is no single way to extend the Shapley value to the class of NTU games. There are three main extensions that have been proposed: the Shapley λ-transfer value [40], the Harsanyi value [16], and the Maschler–Owen consistent value [23]. They were axiomatized in [5,10,17], respectively.

The Connections with Competitive Equilibrium

As was the case for the core, there is a value equivalence theorem. The result holds for the TU domain (see [4,8,38]). It can be shown that the Shapley value payoffs can be supported by competitive prices. Furthermore, in large enough economies, the set of competitive payoffs "shrinks" to approximate the Shapley value. However, the result cannot be easily extended to the NTU domain. While it holds for the λ-transfer value, it need not obtain for the other extensions. For further details, the interested reader is referred to [18] and the references therein.

Non-cooperative Implementation

Reference [15] was the first to propose a procedure that provided some non-cooperative foundations of the Shapley value. Later, other authors have provided alternative procedures and techniques to the same end, including [20,21,29,44].

We shall concentrate on the description of the procedure proposed by Hart and Mas-Colell in [20]. Generalizing an idea found in [22], which studies the case of $\delta = 0$ – see below –, Hart and Mas-Colell propose the following non-cooperative procedure. With equal probability, each player $i \in N$ is chosen to publicly make a feasible proposal to the others: $(x_1, \ldots, x_n)$ is such that the sum of its components cannot exceed $v(N)$. The other players get to respond to it in sequence, following a prespecified order. If all accept, the proposal is implemented; otherwise, a random device is triggered. With probability $0 \leq \delta < 1$, the same game continues being played among the same n players (and thus, a new proposer will be chosen again at random among them), but with probability $1 - \delta$, the proposer leaves the game. He is paid 0 and his resources are removed, so that in the next period, proposals to the remaining $n - 1$ players cannot add up to more than $v(N \setminus \{i\})$. A new proposer is chosen at random among the set $N \setminus \{i\}$, and so on.

As shown in [20], there exists a unique stationary self-enforcing prediction of this procedure, and it actually coincides with the Shapley value payoffs for any value of δ. (Stationarity means that strategies cannot be history dependent). As $\delta \to 1$, the Shapley value payoffs are also obtained not only in expectation, but with independence of who is the proposer. One way to understand this result, as done in [20], is to check that the rules of the procedure and stationary behavior in it are in agreement with Shapley's axioms. That is, the equilibrium relies on immediate acceptances of proposals, stationary strategies treat substitute players similarly, the equations describing the equilibrium have an additive structure, and dummy players will have to receive 0 because no resources are destroyed if they are asked to leave. It is also worth stressing the important role in the procedure of players' marginal contributions to coalitions: following a rejection, a proposer incurs the risk of being thrown out and the others of losing his resources, which seem to suggest a "price" for them.

In [21], the authors study the conditions under which stationarity can be removed to obtain the result. Also, [29] uses a variant of the Hart and Mas-Colell procedure, by replacing the random choice of proposers with a bidding stage, in which players bid to obtain the right to make proposals.

An Application

Consider again the class of majority problems in a parliament consisting of 100 seats. As we shall see, the Shapley value is a good way to understand the power that each party has in the legislature.

Let us begin by considering again the problem in which player 1 has 90 seats, while each of the five small parties has 2 seats. It is easy to see that the Shapley value, like the core in this case, awards the entire unit of surplus to player 1: effectively, each of the small parties is a dummy player, and hence, the Shapley value awards zero to each of them.

Consider a second problem, in which player 1 is a big party with 35 seats, and there are 5 small parties, with 13 seats each. The Shapley value awards 1/3 to the large party, and, by symmetry, 2/15 to each of the small parties. To see this, we need to see when the marginal contributions of player 1 to any coalition are positive. Recall that there are 6! possible orders of players. Note how, if player 1 arrives first or second in the room in which the coalition is forming, his marginal contribution is zero: the coalition was losing before he arrived and continues to be a losing coalition after his arrival. Similarly, his marginal contribution is also zero if he arrives fifth or sixth to the coalition; indeed, in this case, before he arrives the coalition is already winning, so he adds nothing to it. Thus, only when he arrives third or fourth, which happens a third of the times, does he change the nature of the coalition, from losing to winning. This explains his Shapley value share of 1/3. In this game, the Shapley value payoffs roughly correspond to the proportion of seats that each party has.

Next, consider a third problem in which there are two large parties, while the other four parties are very small. For example, let each of the large parties have 48 seats (say, players 1 and 2), while each of the four small parties has only one seat. Now, the Shapley value payoffs are 0.3 to each of the two large parties, and 0.1 to each of the small ones. To see this, note that the marginal contribution of a small party is only positive when he comes fourth in line, and out of the preceding three parties in the coalition, exactly one of them is a large party, i. e., 72 orders out of the 5! orders in which he is fourth. That is, $(72/5!) \times (1/6) = 1/10$. In this case, the competition between the large parties for the votes of the small parties increases the power of the latter quite significantly, with respect to the proportion of seats that each of them holds.

Finally, consider a fourth problem with two large parties (players 1 and 2) with 46 seats each, one mid-size party (player 3) with 5 seats, and three small parties, each with one seat. First, note that each of the three small parties has become a dummy player: no winning coalition where he belongs becomes losing if he leaves the coalition, and so players 4, 5 and 6 are paid zero by the Shapley value. Now, note that, despite the substantial difference of seats between each large party and the mid-size party, each of them is identical in terms of marginal contributions to a winning coalition. Indeed, for $i = 1, 2, 3$, player i's marginal contribution to a coalition is positive only if he arrives second or third or fourth or fifth (and out of the preceding players in the coalition, exactly one is one of the non-dummy players). Note how the Shapley value captures nicely the changes in the allocation of power due to each different political scenario. In this case, the fierce competition between the large parties for the votes of player 3, the swinging party to form a majority, explains the equal share of power among the three.

Future Directions

This article has been a first approach to cooperative game theory, and has emphasized two of its most important solution concepts. The literature on these topics is vast, and the interested reader is encouraged to consult the general references listed below. For the future, one should expect to see progress of the theory into areas that have been less explored, including games with asymmetric information and games with coalitional externalities. In both cases, the characteristic function model must be enriched to take care of the added complexities.

Relevant to this encyclopedia are issues of complexity. The complexity of cooperative solution concepts has been studied (see, for instance, [12]). In terms of computational complexity, the Shapley value seems to be easy to compute, while the core is harder, although some classes of games have been identified in which this task is also simple.

Finally, one should insist on the importance of novel and fruitful applications of the theory to shed new light on concrete problems. In the case of the core, for example, the insights of core stability in matching markets have been successfully applied by Alvin Roth and his collaborators to the design of matching markets in the "real world" (e. g., the job market for medical interns and hospitals, the allocation of organs from doners to patients, and so on) – see [31].

Bibliography

Primary Literature

1. Anderson RM (1978) An elementary core equivalence theorem. Econometrica 46:1483–1487
2. Anderson RM (2008) Core convergence. In: Durlauff S, Blume L (eds) The New Palgrave Dictionary of Economics, 2nd edn. McMillan, London
3. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
4. Aumann RJ (1975) Values of markets with a continuum of traders. Econometrica 43:611–646
5. Aumann RJ (1985) An axiomatization of the non-transferable utility value. Econometrica 53:599–612
6. Aumann RJ (1987) Game theory. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave Dictionary of Economics, Norton, New York
7. Aumann RJ, Peleg B (1960) Von Neumann–Morgenstern solutions to cooperative games without side payments. Bull Am Math Soc 66:173–179
8. Aumann RJ, Shapley LS (1974) Values of Non-Atomic Games. Princeton University Press, Princeton
9. Bondareva ON (1963) Some applications of linear programming methods to the theory of cooperative games (in Russian). Problemy Kibernetiki 10:119 139
10. de Clippel G, Peters H, Zank H (2004) Axiomatizing the Harsanyi solution, the symmetric egalitarian solution and the consistent solution for NTU-games. I J Game Theory 33:145–158
11. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. Int Econ Rev 4:235–246
12. Deng X, Papadimitriou CH (1994) On the complexity of cooperative solution concepts. Math Oper Res 19:257–266
13. Edgeworth FY (1881) Mathematical Psychics. Kegan Paul Publishers, London. reprinted in 2003) Newman P (ed) F. Y. Edgeworth's Mathematical Psychics and Further Papers on Political Economy. Oxford University Press, Oxford
14. Gillies DB (1959) Solutions to general non-zero-sum games. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games IV. Princeton University Press, Princeton, pp 47–85
15. Gul F (1989) Bargaining foundations of Shapley value. Econometrica 57:81–95
16. Harsanyi JC (1963) A simplified bargaining model for the n-person cooperative game. Int Econ Rev 4:194–220
17. Hart S (1985) An axiomatization of Harsanyi's non-transferable utility solution. Econometrica 53:1295–1314
18. Hart S (2008) Shapley value. In: Durlauff S, Blume L (eds) The New Palgrave Dictionary of Economics, 2nd edn. McMillan, London
19. Hart S, Mas-Colell A (1989) Potencial, value and consistency. Econometrica 57:589–614
20. Hart S, Mas-Colell A (1996) Bargaining and value. Econometrica 64:357–380
21. Krishna V, Serrano R (1995) Perfect equilibria of a model of n-person non-cooperative bargaining. I J Game Theory 24:259–272
22. Mas-Colell A (1988) Algunos comentarios sobre la teoria cooperativa de los juegos. Cuadernos Economicos 40:143–161
23. Maschler M, Owen G (1992) The consistent Shapley value for games without side payments. In: Selten R (ed) Rational Interaction: Essays in Honor of John Harsanyi. Springer, New York
24. Moldovanu B, Winter E (1995) Order independent equilibria. Games Econ Behav 9:21–34
25. Nash JF (1953) Two person cooperative games. Econometrica 21:128–140
26. Peleg B (1985) An axiomatization of the core of cooperative games without side payments. J Math Econ 14:203–214
27. Peleg B (1986) On the reduced game property and its converse. I J Game Theory 15:187–200
28. Pérez-Castrillo D (1994) Cooperative outcomes through noncooperative games. Games Econ Behav 7:428–440
29. Pérez-Castrillo D, Wettstein D (2001) Bidding for the surplus: a non-cooperative approach to the Shapley value. J Econ Theory 100:274–294
30. Perry M, Reny P (1994) A non-cooperative view of coalition formation and the core. Econometrica 62:795–817
31. Roth AE (2002) The economist as engineer: game theory, experimentation and computation as tools for design economics. Econometrica 70:1341–1378
32. Scarf H (1967) The core of an *N* person game. Econometrica 38:50–69

33. Serrano R (1995) A market to implement the core. J Econ Theory 67:285–294
34. Serrano R (2005) Fifty years of the Nash program, 1953–2003. Investigaciones Económicas 29:219–258
35. Serrano R, Vohra R (1997) Non-cooperative implementation of the core. Soc Choice Welf 14:513–525
36. Serrano R, Volij O (1998) Axiomatizations of neoclassical concepts for economies. J Math Econ 30:87–108
37. Shapley LS (1953) A value for *n*-person games. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games II. Princeton University Press, Princeton, pp 307–317
38. Shapley LS (1964) Values of large games VII: a general exchange economy with money. Research Memorandum 4248-PR. RAND Corporation, Santa Monica
39. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 14:453–460
40. Shapley LS (1969) Utility comparison and the theory of games. In La Décision: Agrégation et Dynamique des Ordres de Préférence. CNRS, Paris
41. Shapley LS (1971) Cores of convex games. I J Game Theory 1:11–26
42. von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton
43. Walras L (1874) Elements of Pure Economics, or the Theory of Social Wealth. English edition: Jaffé W (ed) Reprinted in 1984 by Orion Editions, Philadelphia
44. Winter E (1994) The demand commitment bargaining and snowballing of cooperation. Econ Theory 4:255–273
45. Young HP (1985) Monotonic solutions of cooperative games. I J Game Theory 14:65–72

Books and Reviews

Myerson RB (1991) Game Theory: An Analysis of Conflict. Harvard University Press, Cambridge

Osborne MJ, Rubinstein A (1994) A Course in Game Theory. MIT Press, Cambridge

Peleg B, Sudholter P (2003) Introduction to the Theory of Cooperative Games. Kluwer, Amsterdam. 2nd edn. Springer, Berlin

Roth AE, Sotomayor M (1990) Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge University Press, Cambridge

Cooperative Games (Von Neumann–Morgenstern Stable Sets)

Jun Wako[1], Shigeo Muto[2]

[1] Department of Economics, Gakushuin University,
Tokyo, Japan

[2] Graduate School of Decision Science and Technology,
Tokyo, Institute of Technology, Tokyo, Japan

Article Outline

Glossary

Characteristic function form game
A characteristic function form game consists of a set of players and a characteristic function that gives each group of players, called a coalition, a value or a set of payoff vectors that they can gain by themselves. It is a typical representation of cooperative games. For characteristic function form games, several solution concepts are defined such as von Neumann – Morgenstern stable set, core, bargaining set, kernel, nucleolus, and Shapley value.

Abstract game An abstract game consists of a set of outcomes and a binary relation, called domination, on the outcomes. Von Neumann and Morgenstern presented this game form for general applications of stable sets.

Strategic form game A strategic form game consists of a player set, each player's strategy set, and each player's payoff function. It is usually used to represent non-cooperative games.

Imputation An imputation is a payoff vector in a characteristic function form game that satisfies group rationality and individual rationality. The former means that the players divide the amount that the grand coalition of all players can gain, and the latter says that each player is assigned at least the amount that he/she can gain by him/herself.

Domination Domination is a binary relation defined on the set of imputations, outcomes, or strategy combinations, depending on the form of a given game. In characteristic function form games, an imputation is said to dominate another imputation if there is a coalition of players such that they can realize their payoffs in the former by themselves, and make each of them better off than in the latter. Domination given a priori in abstract games can be also interpreted in the same way. In strategic form games, domination is defined on the basis of commonly beneficial changes of strategies by coalitions.

Internal stability A set of imputations (outcomes, strategy combinations) satisfies internal stability if there is no domination between any two imputations in the set.

External stability A set of imputations (outcomes, strategy combinations) satisfies external stability if any imputation outside the set is dominated by some imputation inside the set.

von Neumann–Morgenstern stable set A set of imputations (outcomes, strategy combinations) is a von Neumann–Morgenstern stable set if it satisfies both internal and external stability.

Farsighted stable set A farsighted stable set is a more sophisticated stable set concept mainly defined for strategic form games. Given two strategy combinations x and y, we say that x indirectly dominates y if there exist a sequence of coalitions $S^1, \dots, S^p$ and a sequence of strategy combinations $y = x^0, x^1, \dots, x^p = x$ such that each coalition S^j can induce strategy combination x^j by a joint move from x^{j-1}, and all members of S^j end up with better payoffs at $x^p = x$ compared to the payoffs at x^{j-1}. A farsighted stable set is a set of strategy combinations that is stable both internally and externally with respect to the indirect domination. A farsighted stable can be defined in abstract games and characteristic function form games.

Definition of the Subject

The von Neumann–Morgenstern stable set for solution (hereafter stable set) is the first solution concept in cooperative game theory defined by J. von Neumann and O. Morgenstern. Though it was defined cooperative games in characteristic function form, von Neumann and Morgenstern gave a more general definition of a stable set in abstract games. Later, J. Greenberg and M. Chwe cleared a way to apply the stable set concept to the analysis of non-cooperative games in strategic and extensive forms. Though general existence of stable sets in characteristic function form games was denied by a 10-person game presented by W.F. Lucas, stable sets exist in many important games. In voting games, for example, stable sets exist, and they indicate what coalitions can be formed in detail. The core, on the other hand, can be empty in voting games, though it is one of the best known solution concept in cooperative game theory. The analysis of stable sets is not necessarily straightforward, since it can reveal a variety of possibilities. However, stable sets give us deep insights into players' behavior in economic, political and social situations such as coalition formation among players.

Introduction

For studies of economic or social situations where players can take cooperative behavior, the stable set was defined by von Neumann and Morgenstern [31] as a solution concept for characteristic function form cooperative games. They also defined the stable set in abstract games so that one can apply the concept to more general games including non-cooperative situations. Greenberg [9] and Chwe [3] cleared a way to apply the stable set concept to the analysis of non-cooperative games in strategic and extensive forms.

The stable set is a set of outcomes satisfying two stability conditions: internal and external stability. The internal stability means that between any two outcomes in the set, there is no group of players such that all of its members prefer one to the other and they can realize the preferred outcome. The external stability means that for any outcome outside the set, there is a group of players such that all of its members have a commonly preferred outcome in the set and they can realize it. Though general existence was denied by Lucas [18] and Lucas and Rabie [21], the stable set has revealed many interesting behavior of players in economic, political, and social systems.

Von Neumann and Morgenstern (also Greenberg) assumed only a single move by a group of players. Harsanyi [13] first pointed out that stable sets in characteristic function form games may fail to cover "farsighted" behavior of players. Harsanyi's work inspired Chwe's [3] contribution to the formal study of foresight in social environments.

Chwe paid attention to a possible chain of moves such that a move of a group of players will bring about a new move of another group of players, which will further cause a third group of players to move, and so on. Then the group of players moving first should take into account a sequence of moves that may follow, and evaluate their profits obtained at the end. By incorporating such a sequence of moves, Chwe [3] defined a more sophisticated stable set, which we call a farsighted stable set in what follows. Recent work by Suzuki and Muto [51,52] showed that the farsighted stable set provides more reasonable outcomes than the original (myopic) stable set in important classes of strategic form games.

The rest of the chapter is organized as follows. Section "Stable Sets in Abstract Games" presents the definition of a stable set in abstract games. Section "Stable Set and Core" shows basic relations between the two solution concepts of stable set and core. Section "Stable Sets in Characteristic Function form Games" gives the definition of a stable set in characteristic function form games. Section "Applications of Stable Sets in Abstract and Characteristic Function Form Games" first discusses general properties of stable sets in characteristic function form games, and then presents applications of stable sets in abstract and characteristic function games to political and economic systems. Section "Stable Sets and Farsighted Stable Sets in Strategic Form Games" gives the definitions of a stable set and a farsighted stable set in strategic form games. Section "Applications of Farsighted Stable Sets in Strategic Form Games" discusses properties of farsighted stable sets and their applications to social and economic situations. Section "Future Directions" ends the chapter with remarks. Section "Bibliography" offers a list of references.

Stable Sets in Abstract Games

An *abstract game* is a pair $(W, \succ)$ of a set of outcomes W and an irreflexive binary relation $\succ$ on W, where irreflexivity means that $x \succ x$ holds for no element $x \in W$. The relation $\succ$ is interpreted as follows: if $x \succ y$ holds, then there must exist a set of players such that they can induce x from y by themselves and all of them are better off in x.

A subset K of W is called a *stable set* of abstract game $(W, \succ)$ if the following two conditions are satisfied:

1. Internal stability: For any two elements $x, y \in K$, $x \succ y$ never holds.
2. External stability: For any element $z \notin K$, there must exist $x \in K$ such that $x \succ z$.

We explain more in detail what the external and internal stability conditions imply in the definition of a stable set. Suppose players have common understanding that each outcome inside a stable set is "stable" and that each outcome outside the set is "unstable". Here the "stability" means that no group of players has an incentive to deviate from it, and the "instability" means that there is at least one group of players that has an incentive to deviate from it. Then the internal and external stability conditions guarantee that the common understanding is never disproved, and thus continues to prevail.

In fact, suppose the set is both internally and externally stable, and pick any outcome in the set. Then by internal stability, no group of players can be better off by deviating from it and inducing an outcome inside the set. Thus no group of players reaches an agreement to deviate, which makes each outcome inside the set remain stable. Deviating players may be better off by inducing an outcome outside the set; but outcomes outside the set are commonly considered to be unstable. Thus deviating players can never expect that such an outcome will continue. Next pick any outcome outside the set. Then by external stability, there exists at least one group of players who can become better off by deviating from it and inducing an outcome inside the set. The induced outcome is considered to be stable since it is in the set. Hence the group of players will deviate. Hence each outcome outside the set remains unstable.

Stable Set and Core

Another solution concept that is widely known is a core. For a given abstract game $G = (W, \succ)$, a subset C of W is called the *core* of G if $C = \{x \in W |$ there is no $y \in W$ with $y \succ x\}$. From the definition, the core satisfies internal stability. Thus the core C of G is contained in any stable set of G if the latter exists. To see this, suppose that $C \not\subset K$ for a stable set K, and C is non-empty, i. e., $C \neq \emptyset$. (If $C = \emptyset$, then clearly $C \subset K$.) Pick any element $x \in C \backslash K$. Since $x \notin K$, by external stability there exists $y \in K$ with $y \succ x$, which contradicts $x \in C$.

When the core of a game satisfies external stability, it has very strong stability. It is called the *stable core*. The stable core is the unique stable set of the game.

Stable Sets in Characteristic Function form Games

An n-person *game in characteristic function form* with transferable utility is a pair (N, v) of a player set $N = \{1, 2, \ldots, n\}$ and a characteristic function v on the set 2^N of all subsets of N such that $v(\emptyset) = 0$. Each subset of N is called a *coalition*. The game (N, v) is often called a TU-game. A characteristic function form game without transferable utility is called an NTU-game: its characteristic function gives each coalition a set of payoff vectors to players. For NTU-games and their stable sets, refer to Aumann and Peleg [1] and Peleg [35]. In this section, hereafter, we deal with only TU characteristic function form games, and refer to them simply as characteristic function form games.

Let (N, v) be a characteristic function form game. The characteristic function v assigns a real number $v(S)$ to each coalition $S \subseteq N$. The value $v(S)$ indicates the worth that coalition S can achieve by itself.

An n-dimensional vector $x = (x_1, x_2, \ldots, x_n)$ is called a payoff vector. A payoff vector x is called an *imputation* if the following two conditions are satisfied:

1. Group rationality: $\sum_{i=1}^{n} x_i = v(N)$,
2. Individual rationality: $x_i \geq v(\{i\})$ for each $i \in N$.

The first condition says that all players cooperate and share the worth $v(N)$ that they can produce. The second condition says that each player must receive at least the amount that he/she can gain by him/herself. Let A be the set of all imputations.

Let x, y be any imputations and S be any coalition. We say that x *dominates* y *via* S and write this as $x \operatorname{dom}_S y$ if the following two conditions are satisfied:

1. Coalitional rationality: $x_i > y_i$ for each $i \in S$,
2. Effectivity: $\sum_{i \in S} x_i \leq v(S)$.

The first condition says that every member of coalition S strictly prefers x to y. The second condition says that coalition S can guarantee the payoff x_i for each member $i \in S$ by themselves. We say that x *dominates* y (denoted by $x \operatorname{dom} y$) if there exists at least one coalition S such that $x \operatorname{dom}_S y$.

It should be noted that a pair (A, dom) is an abstract game defined in Sect. "Stable Sets in Abstract Games". It is easily seen that "dom" is an irreflexive binary relation on A. A stable set and the core of game (N, v) are defined to be a stable set and the core of the associated abstract game (A, dom), respectively.

Since von Neumann and Morgenstern defined the stable set, its general existence had been one of the most important problems in game theory. The problem was eventually solved in a negative way. Lucas [18] found the following 10-person characteristic function form game in which no stable set exists.

A game with no stable set: Consider the following 10-person game:

$$\begin{aligned}
N &= \{1, 2, \ldots, 10\}\,, \\
v(N) &= 5\,, \quad v(\{1,3,5,7,9\}) = 4\,, \\
v(\{3,5,7,9\}) &= v(\{1,5,7,9\}) = v(\{1,3,7,9\}) = 3\,, \\
v(\{1,2\}) &= v(\{3,4\}) = v(\{5,6\}) = v(\{7,8\}) \\
&= v(\{9,10\}) = 1\,, \\
v(\{3,5,7\}) &= v(\{1,5,7\}) = v(\{1,3,7\}) = v(\{3,5,9\}) \\
&= v(\{1,5,9\}) = v(\{1,3,9\}) = 2\,, \\
v(\{1,4,7,9\}) &= v(\{3,6,7,9\}) = v(\{5,2,7,9\}) = 2 \text{ and} \\
v(S) &= 0 \text{ for all other } S \subset N\,.
\end{aligned}$$

Though this game has no stable set, it has a nonempty core. A game with no stable set and an empty core was also found by Lucas and Rabie [21].

We remark on a class of games in which a stable core exists. As mentioned before, if a stable set exists, it always contains the core, which is of course true also in characteristic function form games. Furthermore, in characteristic function form games, there is an interesting class, called convex games, in which the core satisfies external stability. That is, the core is a stable core. A characteristic function form game (N, v) is a *convex game* if for any $S, T \subseteq N$ with $S \subset T$ and for any $i \notin T$, $v(S \cup \{i\}) - v(S) \leq v(T \cup \{i\}) - v(T)$, i. e., the bigger coalition a player joins, the larger the player's contribution becomes. In convex games, the core is large and satisfies external stability. For the details, refer to Shapley [45].

Though general existence is denied, the stable set provides us with very useful insights into many economic, political, and social issues. In the following, we will present some stable set analyzes applied to those issues.

Applications of Stable Sets in Abstract and Characteristic Function Form Games

Symmetric Voting Games

This section deals with applications of stable sets to voting situations. Let us start with a simple example.

Example 1 Suppose there is a committee consisting of three players 1, 2 and 3. Each player has one vote. Decisions are done according to a simple majority rule. That is, to pass a bill, at least two votes are necessary.

Before analyzing players' behavior, we first formulate the situation as a characteristic function form game. Let the player set be $N = \{1, 2, 3\}$. Since a coalition of a simple majority of players can pass any bill, we give value 1 to such coalitions. Other coalitions can pass no bill. We thus give them value 0. Hence the characteristic function is given by

$$v(S) = \begin{cases} 1 & \text{if } |S| \geq 2\,, \\ 0 & \text{if } |S| \leq 1\,, \end{cases}$$

where $|S|$ denotes the number of players in coalition S. The set of imputations is given by

$$A = \{x = (x_1, x_2, x_3) | x_1 + x_2 + x_3 = 1, x_1, x_2, x_3 \geq 0\}\,.$$

One stable set of this game is given by the set K consisting of three imputations, $(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2)$. A brief proof is the following. Since each of the three imputations has only two numbers 1/2 and 0, internal stability is trivial. To show external stability, take any imputation $x = (x_1, x_2, x_3)$ from outside K. Suppose first $x_1 < 1/2$. Since $x \notin K$, at least one of x_2 and x_3 is less than 1/2. We assume $x_2 < 1/2$. Then $(1/2, 1/2, 0)$ dominates x via coalition $\{1, 2\}$. Next suppose $x_1 = 1/2$. Since $x \notin K, 0 < x_2, x_3 < 1/2$. Thus $(0, 1/2, 1/2)$ dominates x via coalition $\{2, 3\}$. Finally suppose $x_1 > 1/2$. Then $x_2, x_3 < 1/2$, and thus $(0, 1/2, 1/2)$ dominates x via coalition $\{2, 3\}$. Thus the proof of external stability is complete. This three-point stable set indicates that a two-person coalition is formed, and that players in the coalition share equally the outcome obtained by passing a bill.

This game has another three types of stable sets. First, any set $K_c^1 = \{x \in A | x_1 = c\}$ with $0 \leq c < 1/2$ is a stable set. The internal stability of each K_c^1 is trivial. To show external stability, take any imputation $x = (x_1, x_2, x_3) \notin K_c^1$. Suppose $x_1 > c$. Define by $y_1 = c, y_2 = x_2 + (x_1 - c)/2, y_3 = x_3 + (x_1 - c)/2$. Then $y \in K_c^1$ and $y \operatorname{dom}_{\{2,3\}} x$. Next suppose $x_1 < c$. Notice that at least one of x_2 and x_3 is less than $1-c$ since $c < 1/2$. Suppose without loss of generality $x_2 < 1 - c$. Since $c < 1/2$, we have $(c, 1-c, 0) \in K_c^1$ and $(c, 1-c, 0) \operatorname{dom}_{\{1,2\}} x$. Thus external stability holds. This stable set indicates that player 1 gets a fixed amount c and players 2 and 3 negotiate for how to allocate the rest $1 - c$. Similarly, any sets $K_c^2 = \{x \in A | x_2 = c\}$ and $K_c^3 = \{x \in A | x_3 = c\}$ with $0 \leq c < 1/2$ are stable sets.

The three-person game of Example 1 has no other stable set. See von Neumann and Morgenstern [31]. The former stable set is called a *symmetric* (or *objective*) stable set, while the latter types are called *discriminatory* stable sets.

As a generalization of the above result, symmetric stable sets are found in general n-person simple majority voting games. An n-person characteristic function form game (N, v) with $N = \{1, 2, \ldots, n\}$ is called a *simple majority voting game* if

$$v(S) = \begin{cases} 1 & \text{if } |S| > n/2\,, \\ 0 & \text{if } |S| \leq n/2\,. \end{cases}$$

A coalition S with $v(S) = 1$, i. e., with $|S| > n/2$, is called a *winning* coalition. A winning coalition including no smaller winning coalitions is called a *minimal winning* coalition. In simple majority voting games, a minimal winning coalition means a coalition of $(n + 1)/2$ players if n is odd, or $(n + 2)/2$ players if n is even. The following theorem holds. See Bott [2] for the proof.

Theorem 1 *Let (N, v) be a simple majority voting game. Then the following hold.*

(1) If n is odd, then the set

$$K = \langle \underbrace{2/(n+1), \ldots, 2/(n+1)}_{\frac{n+1}{2}}, \underbrace{0, \ldots, 0}_{\frac{n-1}{2}} \rangle$$

is a stable set where the symbol $\langle x \rangle$ denotes the set of all imputations obtained from x through permutations of its components.

(2) If n is even, the set

$$K = \langle \{x \in A |\, \underbrace{x_1 = \ldots = x_{n/2}}_{\frac{n}{2}} \geq \underbrace{x_{(n/2)+1} = \ldots = x_n}_{\frac{n}{2}}\} \rangle$$

is a stable set, where

$$A = \left\{ x = (x_1, \ldots, x_n) \,\middle|\, \sum_{i=1}^{n} x_i = 1, x_1, \ldots, x_n \geq 0 \right\}$$

and $\langle Y \rangle = \bigcup_{x \in Y} \langle x \rangle\,.$

It should be noted from (1) of Theorem 1 that when the number of players is odd, a minimal winning coalition is formed. The members of the coalition share equally the total profit. On the other hand, when the number of players is even, (2) of Theorem 1 shows that every player may gain a positive profit. This implies that the grand coalition of all players is formed. In negotiating for how to share the profit, two coalitions, each with $n/2$ players, are formed and profits are shared equally within each coalition. Since at least $n/2 + 1$ players are necessary to win when n is even, an $n/2$-player coalition is the smallest coalition that can prevent its complement from winning. Such a coalition is called a *minimal blocking* coalition. When n is odd, an $(n + 1)/2$-player minimal winning coalition is also a minimal blocking coalition.

General Voting Games

In this section, we present properties of stable sets and cores in general (not necessarily symmetric) voting games. A characteristic function from game (N, v) is called a *simple game* if $v(S) = 1$ or 0 for each nonempty coalition $S \subseteq N$. A coalition S with $v(S) = 1$ (resp. $v(S) = 0$) is a winning coalition (resp. losing coalition). A simple game is called a *voting game* if it satisfies (1) $v(N) = 1$, (2) if $S \subseteq T$, then $v(S) \leq v(T)$, and (3) if S is winning, then $N - S$ is losing. The first condition implies that the grand coalition N is always winning. The second condition says that a superset of a winning coalition is also winning. The third condition says that there are no two disjoint winning coalitions. It is easily shown that the simple majority voting game studied in the previous section satisfies these conditions. A player has a veto if he/she belongs to every winning coalition. As for cores of voting games, the following theorem holds.

Theorem 2 *Let (N, v) be a voting game. Then the core of (N, v) is nonempty if and only if there exists a player with veto.*

Thus the core is not a useful tool for analyzing voting situations with no veto player. In simple majority voting games, no player has a veto, and thus the core is empty. The following theorem shows that stable sets always exist.

Theorem 3 *Let (N, v) be a voting game. Let S be a minimal winning coalition and define a set K by*

$$K = \left\{ x \in A \middle| \sum_{i \in S} x_i = 1, x_i = 0 \forall i \notin S \right\} .$$

Then K is a stable set.

Thus in voting games, a minimal winning coalition is always formed, and they gain all the profit. For the proofs of these theorems, see Owen [34]. Further results on stable sets in voting games are found in Bott [2], Griesmer [12], Heijmanns [16], Lucas et al. [20], Muto [26,28], Owen [32], Rosenmüller [36], Shapley [43,44].

Production Market Games

Let us start with a simple example.

Example 2 There are four players, each having one unit of a raw material. Two units of the raw material are necessary for producing one unit of an indivisible commodity. One unit of the commodity is sold at p dollars.

The situation is formulated as the following characteristic function form game. The player set is $N = \{1, 2, 3, 4\}$. Since two units of the raw material are necessary to produce one unit of the commodity, the characteristic function v is given by

$$v(S) = 2p \text{ if } |S| = 4 , \quad v(S) = p \text{ if } |S| = 3, 2 , \quad v(S) = 0 \text{ if } |S| = 1, 0 .$$

The set of imputations is

$$A = \{x = (x_1, x_2, x_3, x_4) | x_1 + x_2 + x_3 + x_4 = 2p, \quad x_1, x_2, x_3, x_4 \geq 0\} .$$

The following set K is one of the stable sets of the game:

$$K = \langle \{x = (x_1, x_2, x_3, x_4) \in A | x_1 = x_2 = x_3 \geq x_4\} \rangle .$$

To show internal stability, take two imputations $x = (x_1, x_2, x_3, x_4)$ with $x_1 = x_2 = x_3 \geq x_4$ and $y = (y_1, y_2, y_3, y_4)$ in K. Suppose x dominates y. Since $x_1 = x_2 = x_3 \geq p/2 \geq x_4$, the domination must hold via coalition $\{i, 4\}$ with $i = 1, 2, 3$. Then we have a contradiction $2p = \sum_{i=1}^4 x_i > \sum_{i=1}^4 y_i = 2p$, since $y \in K$ implies that the largest three elements of y are equal. To show external stability, take $z = (z_1, z_2, z_3, z_4) \notin K$. Suppose $z_1 \geq z_2 \geq z_3 \geq z_4$. Then $z_1 > z_3$. Define $y = (y_1, y_2, y_3, y_4)$ by

$$y_i = \begin{cases} z_3 + \dfrac{z_1 + z_2 - 2z_3}{4} & \text{for } i = 1, 2, 3 , \\ z_4 + \dfrac{z_1 + z_2 - 2z_3}{4} & \text{for } i = 4 . \end{cases}$$

Then $y \in K$ and $y \operatorname{dom}_{\{3,4\}} z$, since $y_3 > z_3$, $y_4 > z_4$ and $y_3 + y_4 \leq p = v(\{3, 4\})$.

This stable set shows that in negotiating for how to share the profit of $2p$ dollars, three players form a coalition and share equally the gain obtained through collaboration. At least two players are necessary to produce the commodity. Thus a three-player coalition is the smallest coalition that can prevent its complement from producing the commodity, i. e., a minimal blocking coalition. We would claim that in the market a minimal blocking coalition is formed and that profits are shared equally within the coalition.

An extension of the model was given by Hart [14] and Muto [27]. Hart considered the following production market with n players, each holding one unit of a raw material. To produce one unit of an indivisible commodity, k units of raw materials are necessary. The associated production market game is defined by the player set $N = \{1, 2, \ldots, n\}$ and the characteristic function v given by

$$v(S) = \begin{cases} 0 & \text{if } 0 \leq |S| < k , \\ p & \text{if } k \leq |S| < 2k , \\ \vdots & \\ jp & \text{if } jk \leq |S| < (j+1)k , \\ \vdots & \\ hp & \text{if } hk \leq |S| < n , \end{cases} ,$$

where $n = hk + r$ and h, r are integers such that $h \geq 1$ and $0 \leq r \leq k - 1$. When $h = 1$,

$$v(S) = \begin{cases} 0 & \text{if } |S| < k , \\ p & \text{if } |S| \geq k . \end{cases}$$

The following theorem holds.

Theorem 4 *Suppose $h = 1$. Let $t = n - k + 1$ and $n = tu + w$ where u, w are integers such that $u \geq 1$ and $0 \leq w \leq t - 1$. Then the following set K is a stable set.*

$$\begin{aligned} K = \langle \{x = (x_1, \ldots, x_n) \in A | \\ x_1 = \ldots = x_t \geq x_{t+1} = \ldots = x_{2t} \\ \geq \ldots \geq x_{tu+1} = \ldots = x_n = 0 \rangle , \end{aligned}$$

where

$$A = \left\{ x = (x_1, \ldots, x_n) \middle| \sum_{i=1}^n x_i = p, x_1, \ldots, x_n \geq 0 \right\}$$

is the set of imputations.

The theorem shows that in negotiating for how to share the profit, minimal blocking coalitions, i. e., coalitions of

$n-k+1$ players, are formed and within each coalition, profits are shared equally. Players failing to form a coalition gain nothing. When $h \geq 2$, the following theorem holds.

Theorem 5 *Suppose $h \geq 2$. Let*

$$K = \left\langle \left\{ x = (x_1, \ldots, x_n) \in A \right| \right. \\ \left. \left. x_1 = \cdots = x_{n-k+1} \geq \frac{p}{k} \geq x_{n-k+2} = \cdots = x_n \right\} \right\rangle ,$$

where

$$A = \left\{ x = (x_1, \ldots, x_n) \middle| \sum_{i=1}^{n} x_i = hp, x_1, \ldots, x_n \geq 0 \right\} .$$

Then K is a stable set if and only if

$$n \geq (h+1)(k-1) .$$

Therefore if n is large or k is small, then a minimal blocking coalition is formed and the rest of the players also form a coalition. Within each coalition, profits are shared equally.

The next example deals with the case in which more than one raw materials are necessary to produce a commodity.

Example 3 Two types of raw materials P and Q are needed, one unit each, to produce one unit of an indivisible commodity, which is sold at p dollars. Player 1 holds one unit of raw material p, and each of players 2 and 3 holds one unit of raw material Q.

This situation is formulated as the following characteristic function form game. The player set is $N = \{1, 2, 3\}$. Since one unit of raw materials P and Q are necessary to produce the commodity, the characteristic function ν is given by

$$\nu(N) = p , \\ \nu(\{1,2\}) = \nu(\{1,3\}) = p , \quad \nu(\{2,3\}) = 0 , \\ \nu(\{1\}) = \nu(\{2\}) = \nu(\{3\}) = 0 , \quad \nu(\emptyset) = 0 .$$

The set of imputations is

$$A = \{x = (x_1, x_2, x_3) | x_1 + x_2 + x_3 = p, \ x_1, x_2, x_3 \geq 0\} .$$

The following set K is one of the stable sets in this game:

$$K = \{x = (x_1, x_2, x_3) \in A | x_2 = x_3\} .$$

To show internal stability, take two imputations $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ in K and suppose x dominates y. Then the domination must hold via coalitions $\{1,2\}, \{1,3\}$ since values of other coalitions (except $\{1,2,3\}$) are 0. If $x \operatorname{dom}_{\{1,2\}} y$, then $x_1 > y_1$ and $x_2 > y_2$ hold. Thus we have a contradiction $p = \sum_{i=1}^{3} x_i > \sum_{i=1}^{3} y_i = p$. The domination via $\{1,3\}$ leads to the same contradiction. To show external stability, take any imputation $z = (z_1, z_2, z_3) \notin K$. Then $z_2 \neq z_3$. Without loss of generality, let $z_2 < z_3$. Define $y = (y_1, y_2, y_3)$ by

$$y_i = \begin{cases} z_1 + \dfrac{z_3 - z_2}{3} & \text{for } i = 1 , \\ z_2 + \dfrac{z_3 - z_2}{3} & \text{for } i = 2 , \\ z_2 + \dfrac{z_3 - z_2}{3} & \text{for } i = 3 . \end{cases}$$

Then $y \in K$ and $y \operatorname{dom}_{\{1,2\}} z$, since $y_1 > z_1, y_2 > z_2$ and $y_1 + y_2 < \nu(\{1,2\})$. This stable set shows that in negotiating for how to share the profit p dollars, players 2 and 3 form a coalition against player 1 and share equally the gain obtained through collaboration.

There exist other stable sets in which players 2 and 3 collaborate but they do not share equally the profit. More precisely, the following set

$$K = \{x = (x_1, x_2, x_3) \in A | \\ x_2 \text{ and } x_3 \text{ move towards the same direction}\}$$

is a stable set, where "move towards the same direction" means that if x_2 increases then x_3 increases, and if x_2 decreases then x_3 decreases.

A generalization of the results above is given by the following theorem due to Shapley [42]. Shapley's original theorem is more complicated and holds in more general markets.

Theorem 6 *Suppose there are m players, $1, \ldots, m$, each holding one unit of raw material P, and n players, $m+1, \ldots, m+n$, each holding one unit of raw material Q. To produce one unit of an indivisible commodity, one unit of each of raw materials Pand Q is necessary. One unit of commodity is sold at p dollars. In this market, the following set K is a stable set.*

$$K = \{x = (x_1, x_2, \ldots, x_{m+n}) \in A | \\ x_1 = \cdots = x_m, x_{m+1} = \cdots = x_{m+n}\} .$$

where

$$A = \left\{ x = (x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}) \right| \\ \left. \sum_{i=1}^{m+n} x_i = p \times \min(m,n), \ x_1, \ldots, x_{m+n} \geq 0 \right\} ,$$

is the set of imputations of this game.

This theorem shows that players holding the same raw material form a coalition and share equally the profit gained through collaboration.

For further results on stable sets in production market games, refer to Hart [14], Muto [27], Owen [34]. Refer also to Lucas [19], Owen [33], Shapley [41], for further general studies on stable sets.

Assignment Games

The following two sections deal with markets in which indivisible commodities are traded between sellers and buyers, or bartered among agents. The first market is the assignment market originally introduced by Shapley and Shubik [47].

An assignment market consists of a set of $n(\geq 1)$ buyers $B = \{1, \dots, n\}$ and a set of n sellers $F = \{1', \dots, n'\}$. Each seller $k' \in F$ is endowed with one indivisible commodity to sell, which is called object k'. Thus F also denotes the set of n objects in the market. The objects are differentiated. Each buyer $i \in B$ wants to buy at most one of the objects, and places a nonnegative monetary valuation $u_{ik'}(\geq 0)$ for each object $k' \in F$. The matrix $U = (u_{ik'})_{(i,k') \in B \times F}$ is called the valuation matrix. The sellers place no valuation for any objects. An *assignment market* is denoted by $M(B, F, U)$. We remark that an assignment market with $|B| \neq |F|$ can be transformed into the market with $|B| = |F|$ by adding dummy buyers resp. sellers, and zero rows resp. columns correspondingly to valuation matrix U.

For each coalition $S \subseteq B \cup F$ with $S \cap B \neq \emptyset$ and $S \cap F \neq \emptyset$, we define assignment problem $P(S)$ as follows:

$$\begin{aligned} P(S)\colon \overline{m}(S) = \max_x & \sum_{(i,k') \in (S\cap B) \times (S \cap F)} u_{ik'} x_{ik'} \\ s.t. \sum_{k' \in S \cap F} x_{ik'} &\leq 1 \quad \text{for all } \ i \in S \cap B\,, \\ \sum_{i \in S \cap B} x_{ik'} &\leq 1 \quad \text{for all } \ k' \in S \cap F\,, \\ x_{ik'} &\geq 0 \quad \text{for all } \ (i, k') \in (S \cap B) \times (S \cap F)\,. \end{aligned}$$

Assignment problem $P(S)$ has at least one optimal integer solution (see Simonnard [49]), which gives an optimal matching between sellers and buyers in S that yields the highest possible surplus in S. Without loss of generality, we assume that the rows and columns of valuation matrix U are arranged so that the diagonal assignment x^* with $x^*_{ii'} = 1, i = 1, \dots, n$, is one of the optimal solutions to $P(B \cup F)$.

For a given assignment market $M(B, F, U)$, we define the associated *assignment game* G to be the characteristic function form game $(B \cup F, v)$. The player set of G is $B \cup F$. The characteristic function v is defined as follows: $v(S) = \overline{m}(S)$ for each $S \subseteq F \cup B$ with $S \cap B \neq \emptyset$ and $S \cap F \neq \emptyset$. For coalitions only of sellers or buyers, they cannot produce surplus from trade. Thus $v(S) = 0$ for each S with $S \subseteq B$, $S \subseteq F$, or $S = \emptyset$. The imputation set of G is

$$A = \left\{(w, p) \in \Re^B_+ \times \Re^F_+ \,\middle|\, \sum_{i \in B} w_i + \sum_{k' \in F} p_{k'} = v(B \cup F)\right\}.$$

Shapley and Shubik [47] proved that for any assignment game G, the core C is given by the set of optimal solutions to the dual problem of assignment problem $P(B \cup F)$, i. e.,

$$\begin{aligned} C = \{(w, p) \in A | w_i + p_{k'} \geq u_{ik'} = v(\{i, k'\}) \\ \text{for each } \ (i, k') \in B \times F\}\,, \end{aligned}$$

and thus the core is nonempty. They also showed that sellers' core payoff vector p gives market prices of the respective objects at which the demand and supply equilibrates for each object.

The general existence of stable sets in assignment games is still unsolved. However, as mentioned in Sect. "Stable Set and Core", if a game has the stable core, i. e., the core with external stability, then it is the unique stable set of the game. Thus we consider when an assignment game has the stable core.

Given an assignment market $M(B, F, U)$, we say that valuation matrix U satisfies the *dominant diagonal* condition if all of its diagonal entries are row and column maximums, i. e.,

$$\begin{aligned} u_{ii'} &= \max\{u_{ik'} | k' \in F\} \\ &= \max\{u_{ji'} | j \in B\} \quad \text{for each } \ i = 1, \dots, n\,. \end{aligned}$$

The dominant diagonal condition implies that each buyer i can yield the maximum surplus by purchasing the object of seller i'. Thus the players do not have to compete for partners. They come to be more concerned with the bargaining with his/her best matched partner. Then it is proved that valuation matrix U satisfies the dominant diagonal condition if and only if the core of G includes the imputations $(\underline{w}, \overline{p})$ and $(\overline{w}, \underline{p})$, where $\underline{w}_i = 0, \overline{p}_{i'} = u_{ii'}, \overline{w}_i = u_{ii'}$, and $\underline{p}_{i'} = 0$ for each $i \in N$. Furthermore, the following theorem holds.

Theorem 7 *Let $M(B, F, U)$ be any assignment market with $|B| = |F|$. The associated assignment game G has the stable core if and only if valuation matrix U satisfies the dominant diagonal condition.*

It is also proved that an assignment game G is convex if and only if valuation matrix U satisfies that $u_{ik'} = 0$ for each $(i, k') \in B \times F$ with $i \neq k$. This implies that the core has the von Neumann–Morgenstern stability in a larger class of assignment games including convex assignment games. For more details, refer to Solymosi and Raghavan [50].

House Barter Games

In this section, we consider a market in which only indivisible commodities are bartered. This market was originally considered by Shapley and Scarf [46].

The market we consider has $n(\geq 2)$ players, each endowed with one indivisible commodity, e. g., a house. Let $N = \{1, \ldots, n\}$ be the set of players. The n indivisible commodities are differentiated. The commodity initially owned by player i is called *house i*. Thus N also denotes the set of houses in the market. We assume that each player wants to own exactly one house, and no player disposes any house. Each player i has a complete, reflexive, and transitive preference relation R_i on N. Here, $jR_i h$ denotes that player i prefers house j at least as well as house h. Let $jP_i h$ denote that player i strictly prefers j to h, and $jI_i h$ denote that player i is indifferent between j and h. Defining players' preferences this way, we assume that each player strictly prefers owning a house to owning no house. The bundle $R = (R_i)_{i \in N}$ of players' preference relations is called a *preference profile*.

There is no divisible good such as money in the market. The players only exchange their houses to make a mutually beneficial trade. Thus an *allocation* of the market is defined to be a bijection x from N onto N, where $x(i)$ denotes the house assigned to player i in x. An allocation can be regarded as a permutation of N. An allocation x is also indicated by the vector $x = (x_1, \ldots, x_n)$ with $x_i = x(i)$ for each $i \in N$. Let A be the set of allocations. The market above is referred to as house barter market $M(N, R)$, or briefly market M.

Let x, y be any allocations of market M. For each coalition $S \subseteq N$, let $x(S)$ be the set of houses assigned to the members of S in x, i. e.,

$$x(S) = \{j \in N | j = x(i) \quad \text{for some} \quad i \in S\} .$$

We say that x *weakly dominates* y (denoted by x wdom y) if there exists a coalition S satisfying the following conditions:

1. $x(i)R_i y(i)$ for each $i \in S$ with P_i holding for at least one $i \in S$,
2. $x(S) = S$.

The second condition is the effectivity condition, which requires that each player i in S can obtain house $x(i)$ by exchanging their own endowments. We say that x *strongly dominates* y (denoted by x sdom y) if $x(i)P_i y(i)$ for *each* $i \in S$, and $x(S) = S$. We use the notations x wdom$_S$ y and x sdom$_S$ y to indicate the associated coalition S.

An allocation x is said to be *individually rational* if $x(i)R_i i$ for each player $i \in N$. An allocation x is *Pareto efficient* if there exists no allocation $y \in A$ with y wdom$_N$ x. If there is no $y \in A$ with y sdom$_N$ x, then x is *weakly Pareto efficient*. The three sets of individually rational, Pareto efficient, and weakly Pareto efficient allocations are denoted by IR, PA, and WPA, respectively.

We define cores and stable sets of market M by cores and stable sets of the associated *house barter games* (A, wdom) and (A, sdom), which are the abstract games with the outcome sets given by the allocation set A, and the binary relations on A given by the weak and strong dominations, respectively.

A nonempty subset of A is referred to as a *wdom stable set* of market M if it is a stable set of abstract game (A, wdom). A nonempty subset of A is referred to as a *sdom stable set stable set* of market M if it is a stable set of abstract game (A, sdom). The wdom and sdom stable sets are the stable sets defined by the weak and strong dominations, respectively. A subset of A, which may be empty, is called the *strict core* of market M if it is the core of abstract game (A, wdom). The *core* of market M is the core of abstract game (A, sdom). The strict core and the core of market M are the cores defined by the weak and strong dominations, respectively.

From the definitions above, a wdom stable set is a subset of PA, and an sdom stable set is a subset of WPA. However, both wdom and sdom stable sets may not be subsets of IR. The strict core is a subset of $PA \cap IR$. The core is a subset of $WPA \cap IR$.

Shapley and Scarf [46] proved that the core is nonempty for any house barter market $M(N, R)$. However, since external stability is not imposed on the core, the core does not necessarily coincide with an sdom stable set. In fact, the following example shows that there is a house barter market with no sdom stable set.

Example 4 Let M_1 be the market with the player set $N = \{1, 2, 3\}$ and the following preference profile:

1) $2 \; P_1 \; 3 \; P_1 \; 1$,
2) $3 \; P_2 \; 1 \; P_2 \; 2$,
3) $1 \; P_3 \; 2 \; P_3 \; 3$,

Market M_1 has six allocations: $x^1 = (2, 3, 1)$, $x^2 = (2, 1, 3)$, $x^3 = (1, 3, 2)$, $x^4 = (3, 2, 1)$, $x^5 = (3, 1, 2)$, $x^6 =$

$(1, 2, 3)$. From the preference profile, x^1 is clearly a core allocation. Allocation x^1 strongly dominates x^5 and x^6. Let $X = \{x^2, x^3, x^4\}$. We note that x^1 does not strongly dominate any $x^k \in X$. Here, suppose market M_1 has an sdom stable set K. Then the external stability of K implies $x^1 \in K$ and $K \cap X \neq \emptyset$. Note that $x^2 \text{ sdom}_{\{1,2\}}\, x^4$, $x^4 \text{ sdom}_{\{1,3\}}\, x^3$, and $x^3 \text{ sdom}_{\{2,3\}}\, x^2$. This together with the internal stability of K implies that K can contain only one allocation $x^k \in X$. However, the allocation x^k strongly dominates only one allocation in $X \backslash \{x^k\}$. This means that K does not have external stability, which is a contradiction. Thus, there exists no sdom stable set in market M_1.

Market M_1 however has a nice feature: the singleton $\{x^1\}$ is the strict core, and x^1 weakly dominates all the other allocations $x^2, \ldots, x^6$. In addition, every player shows only strict preferences. Noticing these facts, Roth and Postlewaite [38] proved that for any house barter market $M(N, R)$, if each player has a strict preference relation, then the strict core is a singleton, and it is the unique wdom stable set, i. e., the wdom stable core. Wako [54] proved that this property is extended as follows:

Theorem 8 *For any house barter market $M(N, R)$, if the strict core SC is nonempty, then it is the unique wdom stable set. Furthermore, for any $x, y \in SC$, $x(i) I_i y(i)$ for each $i \in N$.*

Even if some players do not have strict preference relations, the strict core is characterized as the wdom stable core as far as it is nonempty. The wdom stable core SC of market M has nice properties. First, each allocation $x \in SC$ is individually rational, Pareto efficient, and not weakly dominated by any other allocations since x is a strict core allocation. Secondly, any allocation outside SC is weakly dominated by some allocation in SC, since SC is a wdom stable set. Thirdly, even if SC contains different allocations, we may choose any one of them, since they are indifferent for each player.

However, the strict core can be empty when indifferences are allowed in preference relations. Shapley and Scarf [46] already discussed this point by the following example.

Example 5 Let M_2 be the market with the player set $N = \{1, 2, 3\}$ and the following preference profile:

1) $2 \; P_1 \; 3 \; I_1 \; 1$,
2) $1 \; I_2 \; 3 \; P_2 \; 2$,
3) $2 \; P_3 \; 1 \; I_3 \; 3$,

It can be verified that the strict core of M_2 is empty, and that the sets $V^1 = \{(2, 3, 1), (2, 1, 3)\}$ and $V^2 = \{(1, 3, 2), (3, 1, 2)\}$ are both wdom stable sets of M_2. Thus Theorem 8 does not carry over to the cases with the strict core being empty.

Quint and Wako [40] then gave a necessary and sufficient condition for the strict core to be nonempty. For each player $i \in N$ and each nonempty coalition$S \subseteq N$, let $B_i(S)$ be the set of player i's most-preferred house in S, i. e., $B_i(S) = \{h \in S | h R_i j \text{ for each } j \in S\}$. We call a partition $T = \{T_1, \ldots, T_m\}$ of N a *partition by minimal self-mapped sets (PMSS)* if each $T_k \in T$ satisfies the following conditions:

$$T_k = \bigcup_{i \in T_k} B_i\big(N \backslash \cup_{l=1}^{k-1} T_l\big) \text{ and there}$$
$$\text{is no } \; S \subset T_k \; \text{ with } \; S = \bigcup_{i \in S} B_i\big(N \backslash \cup_{l=1}^{k-1} T_l\big) \,.$$

There exists at least one PMSS for any market M. When there are more than one PMSSs, each PMSS consists of the same sets with only the order of some sets being different. We say that group $T_k \in T$ is a lower (resp. higher) group of $T_l \in T$ if $k > l$ (resp. $k < l$). The fact that $T = \{T_1, \ldots, T_m\}$ is a PMSS means that for each player i of any given group $T_k \in T$, the houses most preferred by i among the endowments of groups T_k and lower are all owned in his/her group T_k. The following theorem holds.

Theorem 9 *Let $T = \{T_1, \ldots, T_m\}$ be a PMSS of a given house barter market $M(N, R)$. Then the strict core is nonempty if and only if there exists an allocation $x \in A$ such that*

$$x(T_k) = T_k \;\; \text{and} \;\; x(i) \in B_i\big(N \backslash \cup_{l=1}^{k-1} T_l\big)$$
$$\text{for each } \; i \in T_k \; \text{ and each } \; T_k \in T \,.$$

The necessary and sufficient condition above requires that in any group $T_k \in T$, each player $i \in T_k$ can obtain his/her most-preferred houses (among those owned in groups T_k and lower) through a feasible exchange within his/her own group T_k. We refer to this condition as *segmentability*. Suppose a house barter market has segmentability, and that some players in a group T_k of PMSS T have more preferable houses in higher groups. However, such houses can be exchanged within the higher groups in a mutually beneficial way, and by the definition of a PMSS, no player in the higher groups has an incentive to trade with lower groups. In this sense, the market is segmented into distinct groups. It follows from Theorems 8 and 9 that in house barter markets, segmentability is necessary and sufficient for the existence of the wdom stable core. Quint and Wako [40] gave a polynomial-time algorithm to examine segmentability of a given house barter market. The following example shows a house barter market with segmentability.

Example 6 Let M_3 be the market with the player set $N = \{1, 2, 3, 4, 5, 6\}$ and the following preference profile:

$$\begin{array}{ll}
1) & 2\ P_1\ 3\ P_1\ 5\ P_1\ 4\ P_1\ 1\ P_1\ 6\,, \\
2) & 1\ I_2\ 3\ P_2\ 4\ P_2\ 6\ P_2\ 5\ P_2\ 2\,, \\
3) & 1\ P_3\ 2\ P_3\ 3\ P_3\ 4\ P_3\ 5\ P_3\ 6\,, \\
4) & 2\ P_4\ 5\ P_4\ 6\ P_4\ 3\ P_4\ 4\ P_4\ 1\,, \\
5) & 1\ I_5\ 4\ P_5\ 5\ P_5\ 3\ P_5\ 6\ P_5\ 2\,, \\
6) & 3\ P_6\ 6\ P_6\ 1\ P_6\ 2\ P_6\ 4\ P_6\ 5\,.
\end{array}$$

Although market M_3 has two PMSSs, $T = \{T_1 = \{1, 2, 3\}, T_2 = \{4, 5\}, T_3 = \{6\}\}$ and $T' = \{T'_1 = \{1, 2, 3\}, T'_2 = \{6\}, T'_3 = \{4, 5\}\}$, the differences are only the order of the groups. The wdom stable core of M_3 is the set $K = \{(2, 3, 1, 5, 4, 6)\}$.

The house barter market was also discussed by Moulin [25] from a wide perspective of cooperative microeconomics and game theory. Ehlers [5] initiated a study on stable sets of two-sided matching games, which were originally studied by Gale and Shapley [8]. For a comprehensive study on two-sided matching games, refer to Roth and Sotomayor [39].

Stable Sets and Farsighted Stable Sets in Strategic Form Games

We first defined a stable set in an abstract game in Sect. "Stable Sets in Abstract Games". This means that we can also define a stable set of a strategic form game. In this section we introduce a more sophisticated stable set concept: a *farsighted stable set* of a strategic form game.

Let $G = (N, \{X_i\}_{i \in N}, \{u_i\}_{i \in N})$ be an *n-person strategic form game*, where $N = \{1, 2, \ldots, n\}$ is the set of players, X_i is the set of *strategies* of player i, and u_i is player i's *payoff function*, $u_i\colon X = X_1 \times X_2 \times \cdots \times X_n \to \Re$ (the set of real numbers).

For any two strategy combinations $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in X$, we say that x *is induced from* y *via coalition* $S \subseteq N$ if $y_i = x_i$ for each $i \in N \backslash S$, that is, the combination x is reached from y by moves only of players in S. It is easily seen from the definition that if x is induced from y via coalition S, y is induced from x via coalition S. Thus we write this inducement relation as $x \leftrightarrow_S y$. We say that x *indirectly dominates* y (denoted by $x \succ y$) if there exist a sequence of strategy combinations $y = x^0, x^1, \ldots, x^{p-1}, x^p = x$ and a sequence of coalitions $S^1, \ldots, S^p$ such that for each $j = 1, \ldots, p$, $x^{j-1} \leftrightarrow_{S^j} x^j$ and $u_i(x) > u_i(x^{j-1})$ for each $i \in S^j$. We sometimes say that x *indirectly dominates* y *starting with* S^1 (denoted by $x \succ_{S^1} y$) to specify the set of players which deviates first from y.

We hereupon remark that in the definition of indirect domination we implicitly assume that joint moves by groups of players are neither once-and-for-all nor binding, i. e., some players in a deviating group may later make another move with players within or even outside the group. It should be noted that the indirect domination defined above is borrowed from Chwe [3]. Though Harsanyi [13] first proposed the notion of indirect domination, his definition was given in characteristic function form games.

When $p = 1$ in the definition of indirect domination, we simply say that x *directly dominates* y, which is denoted by $x \succ^d y$. When we want to specify a deviating coalition, we say that x directly dominates y via coalition S, which is denoted by $x \succ^d_S y$. This direct domination in strategic form games was defined by Greenberg [9].

Let pairs $(X, \succ)$ and $(X, \succ^d)$ be the abstract games associated with game G. A *farsighted stable set* of G is defined to be a stable set of abstract game $(X, \succ)$ with indirect domination. A stable set of abstract game $(X, \succ^d)$ with direct domination is simply called a *stable set* of G .

Applications of Farsighted Stable Sets in Strategic Form Games

Existence of farsighted stable sets in strategic form games remains unsolved. Nevertheless, it has turned out through applications that farsighted stable sets give much sharper insights into players' behavior in economic, political, and social situations than (myopic) stable sets with direct dominations. In what follows, we show the analyses of farsighted stable sets in the prisoner's dilemma and two types of duopoly market games in strategic form.

Prisoner's Dilemma

To make discussion as clear as possible, we will focus on the following particular version of the prisoner' s dilemma shown below. Similar results hold in general prisoner' s dilemma games.

Prisoner's Dilemma:

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	4, 4	0, 5
	Defect	5, 0	1, 1

$$\begin{tikzcd}
\text{(Cooperate, Cooperate)} \arrow[r, "2"] \arrow[d, "1"'] & \text{(Cooperate, Defect)} \arrow[d, "1"'] \\
\text{(Defect, Cooperate)} \arrow[r, "2"] & \text{(Defect, Defect)} \arrow[ul, "12"']
\end{tikzcd}$$

Cooperative Games (Von Neumann–Morgenstern Stable Sets), Figure 1

Here → denotes a direct domination. For example, "(Cooperate, Cooperate)$\xrightarrow{2}$(Cooperate,Defect) means (Cooperate, Defect) $\succ^d_{\{2\}}$ (Cooperate,Cooperate). "(Cooperate, Cooperate) $\overset{12}{\leftarrow}$ (Defect, Defect)" means (Cooperate, Cooperate) $\succ^d_{\{1,2\}}$ (Defect, Defect)

We first present a farsighted stable set derived when two players use only pure strategies. In this case, the set of strategy combinations is $X = \{$(Cooperate, Cooperate), (Cooperate, Defect), (Defect, Cooperate), (Defect, Defect)$\}$, where in each combination, the former (resp. the latter) is player 1's (resp. 2's) strategy.

The direct domination relation of this game is summarized as in Fig. 1.

From Fig. 1, no stable set (with direct domination) exists in the prisoner's dilemma. However, since (Cooperate, Cooperate) $\succ$ (Cooperate, Defect), (Defect,Cooperate) and there is no other indirect domination, the singleton $\{$(Cooperate, Cooperate)$\}$ is the unique farsighted stable set with respect to $\succ$. Hence if the two players are farsighted and make a joint but not binding move, the farsighted stable set succeeds in showing that cooperation of the players results in the unique stable outcome.

We now study stable outcomes in the mixed extension of the prisoner's dilemma, i. e., the prisoner's dilemma with mixed strategies played. Let $X_1 = X_2 = [0, 1]$ be the sets of mixed strategies of players 1 and 2, respectively, and let $t_1 \in X_1$ (resp. $t_2 \in X_2$) denote the probability that player 1 (resp. 2) plays "Cooperate". It is easily seen that the minimax payoffs to players 1 and 2 are both 1 in this game. We call a strategy combination that gives both players at least (resp. more than) their minimax payoffs, an *individually rational* (resp. a *strictly individually rational*) strategy combination. We then have the following theorem.

Theorem 10 *Let*

$$T = \{(t_1, t_2) | 1/4 < t_1 \le 1\ , t_2 = 1\} \cup \\ \{(t_1, t_2) | t_1 = 1, 1/4 < t_2 \le 1\}\ ,$$

and define the singleton $K^1(t_1, t_2) = \{(t_1, t_2)\}$ for each $(t_1, t_2) \in T$. Let $K^2 = \{(0, 0), (1, 1/4)\}$ and $K^3 = \{(0, 0), (1/4, 1)\}$. Then the sets K^2, K^3, and any $K^1(t_1, t_2)$ with $(t_1, t_2) \in T$ are farsighted stable sets of the mixed extension of the prisoner's dilemma. There are no other types of farsighted stable sets in the mixed extension of the prisoner's dilemma.

This theorem shows that if the two players are farsighted and make a joint but not binding move in the prisoner's dilemma, then essentially a single Pareto efficient and strictly individually rational strategy combination results as a stable outcome. i. e., $K^1(t_1, t_2)$. We, however, have two exceptional cases K^2, K^3 that (Defect, Defect) could be stable together with one Pareto efficient point at which one player gains the same payoff as in (Defect, Defect).

n-Person Prisoner's Dilemma

We consider an n-person prisoner's dilemma. Let $N = \{1, \dots, n\}$ be the player set. Each player i has two strategies: C (Cooperate) and D (Defect). Let $X_i = \{C, D\}$ for each $i \in N$. Hereafter we refer to a strategy combination as a *state*. The set of states is $X = \prod_{i \in N} X_i$. For each coalition $S \subset N$, let $X_S = \prod_{i \in S} X_i$ and $X_{-S} = \prod_{i \in S^c} X_i$, where S^c denotes the complement of S with respect to N. Let x_s and x_{-s} denote generic elements of X_S and X_{-S}, respectively. Player i's payoff depends not only on his/her strategy but also on the number of other cooperators. Player i's payoff function $u_i : X \to \Re$ is given by $u_i(x) = f_i(x_i, h)$, where $x \in X, x_i \in X_i$ (player i's choice in x), and h is the number of players other than i playing C. We call the strategic form game thus defined an *n-person prisoner's dilemma game*.

For simplifying discussion, we assume that all players are homogeneous and each player has an identical payoff function. That is, f_i's are identical, and simply written as f unless any confusion arises. We assume the following properties on the function f

Assumption 1

(1) $f(D, h) > f(C, h)$ for all $h = 0, 1, \dots, n - 1$
(2) $f(C, n - 1) > f(D, 0)$
(3) $f(C, h)$ and $f(D, h)$ are strictly increasing in h.

Property (1) means that every player prefers playing D to playing C regardless of which strategies other players play. Property (2) means that if all players play C, then each of them gains a payoff higher than the one in $(D, \dots, D)$.

Property (3) means that if the number of cooperators increases, every player becomes better off regardless of which strategy he/she plays.

It holds from Property (1) that $(D, \dots, D)$ is the unique Nash equilibrium of the game. Here for $x, y \in X$, we say that *y is Pareto superior to x* if $u_i(y) \geq u_i(x)$ for all $i \in N$ and $u_i(y) > u_i(x)$ for some $i \in N$. The state $x \in X$ is said to be *Pareto efficient* if there is no $y \in X$ that is Pareto superior to x. By Property (2), $(C, \dots, C)$ is Pareto superior to $(D, \dots, D)$. Together with Property (3), $(C, \dots, C)$ is Pareto efficient.

Given a state x, we say that x is *individually rational* if for all $i \in N, u_i(x) \geq \min_{y_{-i} \in X_{-i}} \max_{y_i \in X_i} u_i(y)$. If a strict inequality holds, we say that x is *strictly individually rational*. It holds from (1), (3) of Assumption 8.1 that $\min_{y_{-i} \in X_{-i}} \max_{y_i \in X_i} u_i(y) = f(D, 0)$.

The following theorem shows that any state that is strictly individually rational and Pareto efficient is itself a singleton farsighted stable set. That is, any strictly individually rational and Pareto efficient outcome is stable if the players are farsighted. Refer to Suzuki and Muto [51,52] for the details.

Theorem 11 *For n-person prisoner's dilemma game, if x is a strictly individually rational and Pareto efficient state, then $\{x\}$ is a farsighted stable set.*

Duopoly Market Games

We consider two types of duopoly: Cournot quantity-setting duopoly and Bertrand price-setting duopoly. For simplifying discussion, we will consider a simple duopoly model in which firms' cost functions and a market demand function are both linear. Similar results, however, hold in more general duopoly models.

There are two firms 1,2, each producing a homogeneous good with the same marginal cost $c > 0$. No fixed cost is assumed.

(1) **Cournot duopoly:** Firms' strategic variables are their production levels. Let x_1 and x_2 be production levels of firms 1 and 2, respectively. The market price $p(x_1, x_2)$ for x_1 and x_2 is given by

$$p(x_1, x_2) = \max(a - (x_1 + x_2), 0) ,$$

where $a > c$. We restrict the domain of production of both firms to $0 \leq x_i \leq a - c$, $i = 1, 2$. This is reasonable since a firm would not overproduce to make a nonpositive profit. When x_1 and x_2 are produced, firm i's profit is given by

$$\pi_i(x_1, x_2) = (p(x_1, x_2) - c)x_i .$$

Thus Cournot duopoly is formulated as the following strategic form game

$$G^C = (N, \{X_i\}_{i=1,2} , \{\pi_i\}_{i=1,2}) ,$$

where the player set is $N = \{1, 2\}$, each player's strategy set is a closed interval between 0 and $a - c$, i. e., $X_1 = X_2 = [0, a - c]$, and their payoff functions are π_i, $i = 1, 2$. Let $X = X_1 \times X_2$. The joint profit of two firms is maximized when $x_1 + x_2 = (a - c)/2$.

(2) **Bertrand duopoly:** Firms' strategic variables are their price levels. Let

$$D(p) = \max(a - p, 0)$$

be the market demand at price p. Then the total profit at p is

$$\prod(p) = (p - c)D(p) .$$

We restrict the domain of price level p of both firms to $c \leq p \leq a$. This assumption is also reasonable since a firm would avoid a negative profit. The total profit $\prod(p)$ is maximized at $p = (a + c)/2$, which is called a *monopoly price*.

Let p_1 and p_2 be prices of firms 1 and 2, respectively. We assume that if firms' prices are equal, then they share equally the total profit, otherwise all sales go to the lower pricing firm of the two. Thus firm i's profit is given by

$$\rho_i(p_i, p_j) = \begin{cases} \prod(p_i) & \text{if } p_i < p_j \\ \prod(p_i)/2 & \text{if } p_i = p_j \\ 0 & \text{if } p_i > p_j \end{cases}$$
$$\text{for } i, j = 1, 2, i \neq j .$$

Hence Bertrand duopoly is formulated as the strategic form game

$$G^B = \left(N, \{Y_i\}_{i=1,2}, \{\rho_i\}_{i=1,2}\right) ,$$

where $N = \{1, 2\}$, $Y_1 = Y_2 = [c, a]$, and $\rho_i (i = 1, 2)$ is i's payoff function. Let $Y = Y_1 \times Y_2$.

It is well-known that a Nash equilibrium is uniquely determined in either market: $x_1 = x_2 = (a - c)/3$ in the Cournot market, and $p_1 = p_2 = c$ in the Bertrand market.

The following theorem holds for the farsighted stable sets in Cournot duopoly.

Theorem 12 *Let $(x_1, x_2) \in X$ be any strategy pair with $x_1 + x_2 = (a - c)/2$. Then the singleton $\{(x_1, x_2)\}$ is a farsighted stable set. Furthermore, every farsighted stable set is of the form $\{(x_1, x_2)\}$ with $x_1 + x_2 = (a - c)/2$ and $x_1, x_2 \geq 0$.*

As mentioned before, any strategy pair (x_1, x_2) with $x_1 + x_2 = (a - c)/2$ and $x_1, x_2 \geq 0$ maximizes two firms' joint profit. This suggests that the von Neumann–Morgenstern stability together with firms' farsighted behavior produce joint profit maximization even if firms' collaboration is not binding.

As for Bertrand duopoly, we have the following theorem, which claims that the monopoly price pair is itself a farsighted stable set, and no other farsighted stable set exists. Therefore the von Neumann–Morgenstern stability together with firms' farsighted behavior attain efficiency (from the standpoint of firms) also in Bertrand duopoly. Refer to Suzuki and Muto [53] for the details.

Theorem 13 *Let $p = (p_1, p_2)$ be the pair of monopoly prices, i. e., $p_1 = p_2 = (a + c)/2$. Then the singleton $\{p\}$ is the unique farsighted stable set.*

For studies of stable sets with direct domination in duopoly market games, refer to Muto and Okada [29,30]. Properties of stable sets and Harsanyi's original farsighted stable sets in pure exchange economies are investigated by Greenberg et al. [10]. For further studies on stable sets and farsighted stable sets in strategic form games, refer to Kaneko [17], Mariotti [24], Xue [55,56], Diamantoudi and Xue [4].

Future Directions

In this paper, we have reviewed applications of von Neumann–Morgenstern stable sets in abstract games, characteristic function form games, and strategic form games to economic, political and social systems.

Stable sets give us insights into coalition formation among players in the systems in question. Farsighted stable sets, especially applied to some economic systems, show that players' farsighted behavior leads to Pareto efficient outcomes even though their collaboration is not binding. The stable set analysis is also applicable to games with infinitely many players. Those analyses show us new approaches to large economic and social systems with infinitely many players. For the details, refer to Hart [15], Einy et al. [7], Einy and Shitovitz [6], Greenberg et al. [11], Shitovitz and Weber [48], and Rosenmüller and Shitovitz [37]. There is also a study on the linkage between common knowledge of Bayesian rationality and achievement of stable sets in generalized abstract games. Refer to Luo [22,23] for the details.

Analyses of social systems by applying the concepts of farsighted stable sets as well as stable sets must further advance theoretical studies on games in which players inherently take both cooperative and non-cooperative behavior. Those studies will in turn have impacts on developments of economics, politics, sociology, and many applied social sciences.

Bibliography

Primary Literature

1. Aumann R, Peleg B (1960) Von Neumann–Morgenstern solutions to cooperative games without side payments. Bull Am Math Soc 66:173–179
2. Bott R (1953) Symmetric solutions to majority games. In: Kuhn HW, Tucker AW (eds) Contribution to the theory of games, vol II. Annals of Mathematics Studies, vol 28. Princeton University Press, Princeton, pp 319–323
3. Chwe MS-Y (1994) Farsighted coalitional stability. J Econ Theory 63:299–325
4. Diamantoudi E, Xue L (2003) Farsighted stability in hedonic games. Soc Choice Welf 21:39–61
5. Ehlers L (2007) Von Neumann–Morgenstern stable sets in matching problems. J Econ Theory 134:537–547
6. Einy E, Shitovitz B (1996) Convex games and stable sets. Games Econ Behav 16:192–201
7. Einy E, Holzman R, Monderer D, Shitovitz B (1996) Core and stable sets of large games arising in economics. J Econ Theory 68:200–211
8. Gale D, Shapley LS (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15
9. Greenberg J (1990) The theory of social situations: an alternative game theoretic approach. Cambridge University Press, Cambridge
10. Greenberg J, Luo X, Oladi R, Shitovitz B (2002) (Sophisticated) stable sets in exchange economies. Games Econ Behav 39:54–70
11. Greenberg J, Monderer D, Shitovitz B (1996) Multistage situations. Econometrica 64:1415–1437
12. Griesmer JH (1959) Extreme games with three values. In: Tucker AW, Luce RD (eds) Contribution to the theory of games, vol IV. Annals of Mathematics Studies, vol 40. Princeton University Press, Princeton, pp 189–212
13. Harsanyi J (1974) An equilibrium-point interpretation of stable sets and a proposed alternative definition. Manag Sci 20:1472–1495
14. Hart S (1973) Symmetric solutions of some production economies. Int J Game Theory 2:53–62
15. Hart S (1974) Formation of cartels in large markets. J Econ Theory 7:453–466
16. Heijmans J (1991) Discriminatory von Neumann–Morgenstern solutions. Games Econ Behav 3:438–452
17. Kaneko M (1987) The conventionally stable sets in noncooperative games with limited observations I: Definition and introductory argument. Math Soc Sci 13:93–128
18. Lucas WF (1968) A game with no solution. Bull Am Math Soc 74:237–239
19. Lucas WF (1990) Developments in stable set theory. In: Ichiishi T et al (eds) Game Theory and Applications, Academic Press, New York, pp 300–316
20. Lucas WF, Michaelis K, Muto S, Rabie M (1982) A new family of finite solutions. Int J Game Theory 11:117–127

21. Lucas WF, Rabie M (1982) Games with no solutions and empty cores. Math Oper Res 7:491–500
22. Luo X (2001) General systems and φ-stable sets: a formal analysis of socioeconomic environments. J Math Econ 36:95–109
23. Luo X (2006) On the foundation of stability. Academia Sinica, Mimeo, available at http://www.sinica.edu.tw/~xluo/pa14.pdf
24. Mariotti M (1997) A model of agreements in strategic form games. J Econ Theory 74:196–217
25. Moulin H (1995) Cooperative Microeconomics: A Game-Theoretic Introduction. Princeton University Press, Princeton
26. Muto S (1979) Symmetric solutions for symmetric constant-sum extreme games with four values. Int J Game Theory 8:115–123
27. Muto S (1982) On Hart production games. Math Oper Res 7:319–333
28. Muto S (1982) Symmetric solutions for (n,k) games. Int J Game Theory 11:195–201
29. Muto S, Okada D (1996) Von Neumann–Morgenstern stable sets in a price-setting duopoly. Econ Econ 81:1–14
30. Muto S, Okada D (1998) Von Neumann–Morgenstern stable sets in Cournot competition. Econ Econ 85:37–57
31. von Neumann J, Morgenstern O (1953) Theory of Games and Economic Behavior, 3rd ed. Princeton University Press, Princeton
32. Owen G (1965) A class of discriminatory solutions to simple n-person games. Duke Math J 32:545–553
33. Owen G (1968) n-Person games with only 1, n-1, and n-person coalitions. Proc Am Math Soc 19:1258–1261
34. Owen G (1995) Game theory, 3rd ed. Academic Press, New York
35. Peleg B (1986) A proof that the core of an ordinal convex game is a von Neumann–Morgenstern solution. Math Soc Sci 11:83–87
36. Rosenmüller J (1977) Extreme games and their solutions. In: Lecture Notes in Economics and Mathematical Systems, vol 145. Springer, Berlin
37. Rosenmüller J, Shitovitz B (2000) A characterization of vNM-stable sets for linear production games. Int J Game Theory 29:39–61
38. Roth A, Postlewaite A (1977) Weak versus strong domination in a market with indivisible goods. J Math Econ 4:131–137
39. Roth A, Sotomayor M (1990) Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge University Press, Cambridge
40. Quint T, Wako J (2004) On houseswapping, the strict core, segmentation, and linear programming, Math Oper Res 29:861–877
41. Shapley LS (1953) Quota solutions of n-person games. In: Kuhn HW, Tucker TW (eds) Contribution to the theory of games, vol II. Annals of Mathematics Studies, vol 28. Princeton University Press, Princeton, pp 343–359
42. Shapley LS (1959) The solutions of a symmetric market game. In: Tucker AW, Luce RD (eds) Contribution to the theory of games, vol IV. Annals of Mathematics Studies, vol 40. Princeton University Press, Princeton, pp 145–162
43. Shapley LS (1962) Simple games: An outline of the descriptive theory. Behav Sci 7:59–66
44. Shapley LS (1964) Solutions of compound simple games. In Tucker AW et al (eds) Advances in Game Theory. Annals of Mathematics Studies, vol 52. Princeton University Press, Princeton, pp 267–305
45. Shapley LS (1971) Cores of convex games. Int J Game Theory 1:11–26
46. Shapley LS, Scarf H (1974) On cores and indivisibilities. J Math Econ 1:23–37
47. Shapley LS, Shubik M (1972) The assignment game I: The core. Int J Game Theory 1:111–130
48. Shitovitz B, Weber S (1997) The graph of Lindahl correspondence as the unique von Neumann–Morgenstern abstract stable set. J Math Econ 27:375–387
49. Simonnard M (1966) Linear programming. Prentice-Hall, New Jersey
50. Solymosi T, Raghavan TES (2001) Assignment games with stable core. Int J Game Theory 30:177–185
51. Suzuki A, Muto S (2000) Farsighted stability in prisoner's dilemma. J Oper Res Soc Japan 43:249–265
52. Suzuki A, Muto S (2005) Farsighted stability in n-person prisoner's dilemma. Int J GameTheory 33:431–445
53. Suzuki A, Muto S (2006) Farsighted behavior leads to efficiency in duopoly markets. In: Haurie A, et al (eds) Advances in Dynamic Games. Birkhauser, Boston, pp 379–395
54. Wako J (1991) Some properties of weak domination in an exchange market with indivisible goods. Jpn Econ Rev 42:303–314
55. Xue L (1997) Nonemptiness of the largest consistent set. J Econ Theory 73:453–459
56. Xue L (1998) Coalitional stability under perfect foresight. Econ Theory 11:603–627

Books and Reviews

Lucas WF (1992) Von Neumann–Morgenstern stable sets. In: Aumann RJ, Hart S (eds) Handbook of Game Theory with Economic Applications, vol 1. North-Holland, Amsterdam, pp 543–590

Shubik M (1982) Game theory in the social sciences: Concepts and solutions. MIT Press, Boston

Cooperative Multi-hierarchical Query Answering Systems

ZBIGNIEW W. RAS[1,2], AGNIESZKA DARDZINSKA[3]

[1] Department of Computer Science,
University of North Carolina, Charlotte, USA

[2] Institute of Computer Science,
Polish Academy of Sciences, Warsaw, Poland

[3] Department of Computer Science,
Białystok Technical University, Białystok, Poland

Article Outline

Glossary

Autonomous information system An autonomous information system is an information system existing as an independent entity.

Intelligent query answering Intelligent query answering is an enhancement of query-answering into a sort of intelligent system (capable or being adapted or molded). Such systems should be able to interpret incorrectly posed questions and compose an answer not necessarily reflecting precisely what is directly referred to by the question, but rather reflecting what the intermediary understands to be the intention linked with the question.

Knowledge base Knowledge base is a collection of rules defined as expressions written in predicate calculus. These rules have a form of associations between conjuncts of values of attributes.

Ontology Ontology is an explicit formal specification of how to represent objects, concepts and other entities that are assumed to exist in some area of interest and relationships holding among them. Systems that share the same ontology are able to communicate about the domain of discourse without necessarily operating on a globally shared theory. A system commits to ontology if its observable actions are consistent with the definitions in the ontology.

Semantics The meaning of expressions written in some language as opposed to their syntax which describes how symbols may be combined independently of their meaning.

Definition of the Subject

One way to make a Query Answering System (QAS) intelligent is to assume the hierarchical structure of their attributes. Such systems have been investigated by Cuppens and Demolombe [3], Gal and Minker [4], and Gaasterland et al. [6], and they are called cooperative. Queries submitted to them are built, in a classical way, from values of attributes describing objects in an information system S and from two-argument functors "and", "or". Instead of "or", we use the symbol "+". Instead of "and", we use the symbol "*". Let us assume that QAS is associated with an information system S. Now, if query q submitted to QAS fails, then any attribute value listed in q can be generalized and the number of objects supporting q in S may increase. In cooperative systems, these generalizations are controlled either by users [4], or by methods based on knowledge discovery [12]. Conceptually, a similar approach has been proposed by Lin [11]. He defines a neighborhood of an attribute value which we can interpret as its generalization (or its parent in the corresponding hierarchical attribute structure). When query fails, then the query answering system is trying to replace values in a query by new values from their corresponding neighborhoods. QAS for S can also collaborate and exchange knowledge with other information systems. In all such cases, it is called intelligent. In papers [14,15] the query answering strategy was based on a guided process of knowledge (rules) extraction and knowledge exchange among systems. Knowledge extracted from information systems collaborating with S was used to construct new attributes in S and/or impute null or hidden values of attributes in S. This way we do not only enlarge the set of queries which QAS can successfully answer but also increase the overall number of retrieved objects and their confidence. Some attributes in S can be distinguished. We usually call them decision attributes. Their values represent concepts which can be defined in terms of the remaining attributes in S, called classification attributes. Query languages for such information systems are built only from values of decision attributes and from two-argument functors "+", "*" [16]. The semantics of queries is defined in terms of semantics of values of classification attributes. Precision and recall of QAS is strictly dependent on the support and confidence of the classifiers used to define queries.

Introduction

Responses by QAS to submitted queries do not always contain the information desired and although they may be logically correct, can sometimes be misleading. Research in the area of intelligent query answering rectifies these problems. The classical approach is based on a cooperative method called relaxation for expanding an information system and related to it queries [3,4]. The relaxation method expands the scope of a query by relaxing the constraints implicit in the query. This allows QAS to return answers related to the original query as well as the literal answers which may be of interest to the user.

This paper concentrates on multi-hierarchical decision systems which are defined as information systems with several hierarchical distinguished attributes called decision attributes. Their values are used to build queries. We give the theoretical framework for modeling such systems and its corresponding query languages. Standard interpretation and the classifier-based interpretation of queries are introduced and used to model the quality (precision, recall) of QAS.

Multi-hierarchical Decision System

In this section we introduce the notion of a multi-hierarchical decision system S and the query language built from atomic expressions containing only values of the decision attributes in S. Classifier-based semantics and the standard semantics of queries in S are proposed. The set of objects X in S is defined as the interpretation domain for both semantics. Standard semantics identifies all objects in X which should be retrieved by a query. Classifier-based semantics gives weighted sets of objects which are retrieved by queries. The notion of precision and recall of QAS in the proposed setting is introduced. We use only rule-based classifiers to define the classifier-based semantics. By improving the confidence and support of the classifiers we improve the precision and recall of QAS.

Definition 1 By a multi-hierarchical decision system we mean a triple $S = (X, A \cup D, V)$, where X is a nonempty, finite set of objects, $D = \{d[i] : 1 \le i \le k\}$ is a set of hierarchical decision attributes, A is a nonempty finite set of classification attributes, and $V = \bigcup\{V_a : a \in A \cup D\}$ is a set of their values.

We assume that: V_a, V_b are disjoint for any $a, b \in A \cup D$, such that $a \neq b$, $a : X \to V_a$ is a partial function for every $a \in A \cup D$.

Definition 2 By a set of decision queries (d-queries) for S we mean a least set T_D such that:

- $0, 1 \in T_D$,
- if $w \in \bigcup\{V_a : a \in D\}$, then $w, \sim w \in T_D$,
- if $t_1, t_2 \in T_D$, then $(t_1 + t_2), (t_1 * t_2) \in T_D$.

Definition 3 Decision query t is called simple if $t = t_1 * t_2 * \ldots * t_n$ and

$$(\forall j \in \{1, 2, \ldots, n\}) \left[\left(t_j \in \bigcup\{V_a : a \in D\}\right) \vee \left(t_j = \sim w \wedge w \in \bigcup\{V_a : a \in D\}\right)\right] .$$

Definition 4 By a set of classification terms (c-terms) for S we mean a least set T_C such that:

- $0, 1 \in T_C$,
- if $w \in \bigcup\{V_a : a \in A\}$, then $w, \sim w \in T_C$,
- $t_1, t_2 \in T_C$, then $(t_1 + t_2), (t_1 * t_2) \in T_C$.

Definition 5 Classification term t is called simple if $t = t_1 * t_2 * \ldots * t_n$ and

$$(\forall j \in \{1, 2, \ldots, n\}) \left[\left(t_j \in \bigcup\{V_a : a \in A\}\right) \vee \left(t_j = \sim w \wedge w \in \bigcup\{V_a : a \in A\}\right)\right] .$$

Definition 6 By a classification rule we mean any expression of the form $[t_1 \to t_2]$, where t_1 is a simple classification term and t_2 is a simple decision query.

Definition 7 Semantics M_S of c-terms in $S = (X, A \cup D, V)$ is defined in a standard way as follows:

- $M_S(0) = 0$, $M_S(1) = X$,
- $M_S(w) = \{x \in X : w = a(x)\}$ for any $w \in V_a$, $a \in A$,
- $M_S(\sim w) = \{x \in X : (\exists v \in V_a)[v = a(x) \,\&\, v \neq w]\}$ for any $w \in V_a$, $a \in A$,
- if t_1, t_2 are terms, then

$$M_S(t_1 + t_2) = M_S(t_1) \cup M_S(t_2) ,$$
$$M_S(t_1 * t_2) = M_S(t_1) \cap M_S(t_2) .$$

Now, we introduce the notation used for values of decision attributes. Assume that the term $d[i]$ also denotes the first granularity level of a hierarchical decision attribute $d[i]$. The set $\{d[i,1], d[i,2], d[i,3], \ldots\}$ represents the values of attribute $d[i]$ at its second granularity level. The set $\{d[i,1,1], d[i,1,2], \ldots, d[i,1,n_i]\}$ represents the values of attribute d at its third granularity level, right below the node $d[i,1]$. We assume here that the value $d[i,1]$ can be refined to any value from $\{d[i,1,1], d[i,1,2], \ldots, d[i,1,n_i]\}$, if necessary. Similarly, the set $\{d[i,3,1,3,1], d[i,3,1,3,2], d[i,3,1,3,3], d[i,3,1,3,4]\}$ represents the values of attribute d at its fourth granularity level which are finer than the value $d[i,3,1,3]$.

Now, let us assume that a rule-based classifier is used to extract rules describing simple decision queries in S. We denote that classifier by **RC**. The definition of semantics N_S of c-terms is **RC** independent whereas the definition of semantics M_S of d-queries is **RC** dependent.

Definition 8 Classifier-based semantics M_S of d-queries in $S = (X, A \cup D, V)$ is defined as follows:

If t is a simple d-query in S and $\{r_j = [t_j \to t] : j \in J_t\}$ is a set of all rules defining t which are extracted from S by classifier **RC**, then $M_S(t) = \{(x, p_x) : (\exists j \in J_t)(x \in M_S(t_j)[p_x = \Sigma\{\text{conf}(j) \cdot \text{sup}(j) : x \in M_S(t_j) \,\&\, j \in J_t\}/\Sigma\{\text{sup}(j) : x \in M_S(t_j) \,\&\, j \in J_t\}]$, where conf($j$), sup($j$) denote the confidence and the support of the rule $[t_j \to t]$, correspondingly.

Definition 9 Attribute value $d[j_1, j_2, \ldots, j_n]$ in $S = (X, A \cup D, V)$ is dependent on $d[i_1, i_2, \ldots, i_k]$ in S, if one of the following conditions hold:

1) $n \le k \,\&\, (\forall m \le n)\,[i_m = j_m]$,
2) $n > k \,\&\, (\forall m \le k)\,[i_m = j_m]$.

Otherwise, $d[j_1, j_2, \ldots, j_n]$ is called independent from $d[i_1, i_2, \ldots, i_k]$ in S.

Example 1 The attribute value $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}]$ is dependent on the attribute value $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}, 5, 3]$. Also, $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}, \mathbf{5}, \mathbf{3}, 2, 4]$ is dependent on $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}, \mathbf{5}, \mathbf{3}]$.

Definition 10 Let $S = \left(X, A \cup \{d[1], d[2], \ldots, d[k]\}, V\right)$, $w \in V_{d[i]}$, and $IV_{d[i]}$ be the set of all attribute values in $V_{d[i]}$ which are independent from w.

Standard semantics N_S of d-queries in S is defined as follows:

- $N_S(0) = 0$, $N_S(1) = X$,
- if $w \in V_{d[i]}$, then $N_S(w) = \{x \in X: d[i](x) = w\}$, for any $1 \leq i \leq k$
- if $w \in V_{d[i]}$, then $N_S(\sim w) = \{x \in X: (\exists v \in IV_{d[i]}) [d[i](x) = v]\}$, for any $1 \leq i \leq k$
- if t_1, t_2 are terms, then

$$N_S(t_1 + t_2) = N_S(t_1) \cup N_S(t_2)\,,$$
$$N_S(t_1 * t_2) = N_S(t_1) \cap N_S(t_2)\,.$$

Definition 11 Let $S = (X, A \cup D, V)$, t is a d-query in S, $N_S(t)$ is its meaning under standard semantics, and $M_S(t)$ is its meaning under classifier-based semantics. Assume that $N_S(t) = X_1 \cup Y_1$, where $X_1 = \{x_i, i \in I_1\}$, $Y_1 = \{y_i, i \in I_2\}$. Assume also that $M_S(t) = \{(x_i, p_i): i \in I_1\} \cup \{(z_i, q_i): i \in I_3\}$ and $\{y_i, i \in I_2\} \cap \{z_i, i \in I_3\} = \emptyset$.

By precision of a classifier-based semantics M_S on a d-query t, we mean

$$\text{Prec}(M_S, t) = [\Sigma\{p_i: i \in I_1\} + \Sigma\{(1 - q_i): i \in I_3\}] / [\text{card}(I_1) + \text{card}(I_3)]\,.$$

By recall of a classifier-based semantics M_S on a d-query t, we mean

$$\text{Rec}(M_S, t) = [\Sigma\{p_i: i \in I_1\}]/[\text{card}(I_1) + \text{card}(I_2)]\,.$$

Example 2 Assume that $N_S(t) = \{x_1, x_2, x_3, x_4\}$, $M_S(t) = \{(x_1, p_1), (x_2, p_2), (x_5, p_5), (x_6, p_6)\}$. Then:

$$\text{Prec}(M_S, t) = [p_1 + p_2 + (1 - p_5) + (1 - p_6)]/4\,,$$
$$\text{Rec}(M_S, t) = [p_1 + p_2]/4\,.$$

Cooperative Query Answering

There are cases when classical Query Answering Systems fail to return any answer to a d-query q but still a satisfactory answer can be found. For instance, let us assume that in a multi-hierarchical decision system $S = (X, A \cup D, V)$, where $D = \{d[1], d[2], \ldots, d[k]\}$, there is no single object whose description matches the query q. Assuming that a distance measure between objects in S is defined, then by generalizing q, we may identify objects in S whose descriptions are closest to the description of q. This problem is similar to the problem when the granularity of an attribute value used in a query q is finer than the granularity of the corresponding attribute used in S. By replacing such attribute values in q by more general values used in S, we may retrieve objects from S which satisfy q.

Definition 12 The distance δ_S between two attribute values $d[j_1, j_2, \ldots j_n], d[i_1, i_2, \ldots, i_m]$ in $S = (X, A \cup D, V)$, where $j_1 = i_1$, $p \geq 1$, is defined as follows:

1) if $[j_1, j_2, \ldots, j_p] = [i_1, i_2, \ldots, i_p]$ and $j_{p+1} \neq i_{p+1}$, then $\delta_S\left[d[j_1, j_2, \ldots, j_n], d[i_1, i_2, \ldots, i_m]\right] = 1/2^{p-1}$
2) if $n \leq m$ and $[j_1, j_2, \ldots, j_n] = [i_1, i_2, \ldots, i_n]$, then $\delta_S\left[d[j_1, j_2, \ldots, j_n], d[i_1, i_2, \ldots, i_m]\right] = 1/2^n$

The second condition, in the above definition, represents the average case between the best and the worth case.

Example 3 Following the above definition of the distance measure, we get:

1. $\delta_S\left[d[2, 3, 2, 4], d[2, 3, 2, 5, 1]\right] = 1/4$
2. $\delta_S\left[d[2, 3, 2, 4], d[2, 3, 2]\right] = 1/8$

Let us assume that $q = q(a[3, 1, 3, 2], b[1], c[2])$ is a d-query submitted to S. The notation $q(a[3, 1, 3, 2], b[1], c[2])$ means that q is built from $a[3, 1, 3, 2], b[1], c[2]$ which are the atomic attribute values in S. Additionally, we assume that attribute a is not only hierarchical but also it is ordered. It basically means that the difference between the values $a[3, 1, 3, 2]$ and $a[3, 1, 3, 3]$ is smaller than between the values $a[3, 1, 3, 2]$ and $a[3, 1, 3, 4]$. Also, the difference between any two elements in $\{a[3, 1, 3, 1], a[3, 1, 3, 2], a[3, 1, 3, 3], a[3, 1, 3, 4]\}$ is smaller than between $a[3, 1, 3]$ and $a[3, 1, 2]$.

Now, we outline a possible strategy which QAS can follow to solve q. Clearly, the best solution for answering q is to identify objects in S which precisely match the d-query submitted by the user. If it fails, we try to identify objects which match d-query $q(a[3, 1, 3], b[1], c[2])$. If we succeed, then we try d-queries $q(a[3, 1, 3, 1], b[1], c[2])$ and $q(a[3, 1, 3, 3], b[1], c[2])$. If we fail, then we should succeed with $q(a[3, 1, 3, 4], b[1], c[2])$. If we fail with $q(a[3, 1, 3], b[1], c[2])$, then we try $q(a[3, 1], b[1], c[2])$ and so on.

Cooperative Multi-hierarchical Query Answering Systems, Table 1
Multi-hierarchical decision system *S*

X	e	f	g	…	…	a	b	c	d
x_1	$e[1$	$f[1]$	…	…	…	$a[1]$	$b[2]$	$c[1,1]$	$d[3]$
x_2	$e[2]$	$f[1]$	…	…	…	$a[1,1]$	$b[2,1]$	$c[1,1,1]$	$d[3,1,2]$
x_3	$e[2]$	$f[1]$	…	…	…	$a[1,1,1]$	$b[2,2,1]$	$c[2,2]$	$d[1]$
x_4	$e[1]$	$f[2]$	…	…	…	$a[2]$	$b[2,2]$	$c[1,1]$	$d[1,1]$

To present this cooperative strategy in a more precise way, we use an example and start with a very simple dataset. Namely, we assume that S has four decision attributes which belong to the set $\{a, b, c, d\}$. System S contains only four objects listed in Table 1.

Now, we assume that d-query $q = a[1,2] * b[2] * c[1,1] * d[3,1,1]$ is submitted to the multi-hierarchical decision system S (see Table 1). Clearly, q fails in S.

Jointly with q, also a threshold value for a minimum support can be supplied as a part of a d-query. This threshold gives the minimal number of objects that need to be returned as an answer to q. When QAS fails to answer q, the nearest objects satisfying q have to be identified.

The algorithm for finding these objects is based on the following steps:

If QAS fails to identify a sufficient number of objects satisfying q in S, then the generalization process starts. We can generalize either attribute a or d. Since the value $d[3,1,2]$ has a lower granularity level than $a[1,1]$, then we generalize $d[3,1,2]$ getting a new query $q_1 = a[1,2] * b[2] * c[1,1] * d[3,1]$. But q_1 still fails in S. Now, we generalize $a[1,1]$ getting a new query $q_2 = a[1] * b[2] * c[1,1] * d[3,1]$. Objects x_1, x_2 are the only objects in S which support q_2.

If the user is only interested in one object satisfying the query q, then we need to identify which object in $\{x_1, x_2\}$ has a distance closer to q.

Clearly,

$$\begin{aligned}\delta_S[q, x_1] &= \delta_S\big[[a[1,2], b[2], c[1,1], d[3,1,1]],\\ &\qquad [a[1], b[2], c[1,1], d[3]]\big]\\ &= 1/4 + 0 + 0 + 1/4 = 1/2\,,\\ \delta_S[q, x_2] &= \delta_S\big[[a[1,2], b[2], c[1,1], d[3,1,1]],\\ &\qquad [a[1,1], b[2,1], c[1,1,1], d[3,1,2]]\big]\\ &= 1/4 + 1/4 + 1/8 + 1/8 = 3/4\,,\end{aligned}$$

which means x_1 is the winning object.

Note that the cooperative strategy only identifies objects satisfying d-queries and it identifies objects to be returned by QAS to the user. The confidence assigned to these objects depends on the classifier **RC**.

Future Directions

We have introduced the notion of system-based semantics and user-based semantics of queries. User-based semantics are associated with the indexing of objects by a user which is time consuming and unrealistic for very large sets of data. System-based semantics are associated with automatic indexing of objects in X which strictly depends on the support and confidence of classifiers and depends on the precision and recall of a query answering system. The quality of classifiers can be improved by a proper enlargement of the set X and the set of features describing them which differentiate the real-life objects from the same semantic domain as X in a better way. An example, for instance, is given in [16]. The quality of a query answering system can be improved by its cooperativeness. Both precision and recall of QAS is increased if no-answer queries are replaced by generalized queries which are answered by QAS on a higher granularity level than the initial level of queries submitted by users. Assuming that the system is distributed, the quality of QAS for multi-hierarchical decision system S can also be improved through collaboration among sites [14,15].

The key concept of intelligent QAS based on collaboration among sites is to generate global knowledge through knowledge sharing. Each site develops knowledge independently which is used jointly to produce global knowledge. Assume that two sites S_1 and S_2 accept the same ontology of their attributes and share their knowledge in order to solve a user query successfully. Also, assume that one of the attributes at site S_1 is confidential. The confidential data in S_1 can be hidden by replacing them with null values. However, users at S_1 may treat them as missing data and reconstruct them with the knowledge extracted from S_2 [10]. The vulnerability illustrated in this example shows that a security-aware data management is an essential component for any intelligent QAS to ensure data confidentiality.

Bibliography

1. Chmielewski MR, Grzymala-Busse JW, Peterson NW (1993) The rule induction system LERS – a version for personal computers. Found Comput Decis Sci 18(3-4):181–212
2. Chu W, Yang H, Chiang K, Minock M, Chow G, Larson C (1996) Cobase: A scalable and extensible cooperative information system. J Intell Inf Syst 6(2/3):223–259
3. Cuppens F, Demolombe R (1988) Cooperative answering: a methodolgy to provide intelligent access to databases. Pro-

ceeding of the Second International Conference on Expert Database Systems, pp 333–353

4. Gal A, Minker J (1988) Informative and cooperative answers in databases using integrity constraints. Natural Language Understanding and Logic Programming, North Holland, pp 277–300
5. Gaasterland T (1997) Cooperative answering through controlled query relaxation. IEEE Expert 12(5):48–59
6. Gaasterland T, Godfrey P, Minker J (1992) Relaxation as a platform for cooperative answering. J Intell Inf Syst 1(3):293–321
7. Giannotti F, Manco G (2002) Integrating data mining with intelligent query answering. Logics in Artificial Intelligence. Lecture Notes in Computer Science, vol 2424. Springer, Berlin, pp 517–520
8. Godfrey P (1993) Minimization in cooperative response to failing database queries. Int J Coop Inf Syst 6(2):95–149
9. Guarino N (1998) Formal ontology in information systems. IOS Press, Amsterdam
10. Im S, Ras ZW (2007) Protection of sensitive data based on reducts in a distributed knowledge discovery system. Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), Seoul. IEEE Computer Society, pp 762–766
11. Lin TY (1989) Neighborhood systems and approximation in relational databases and knowledge bases. Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems. Poster Session Program, Oak Ridge National Laboratory, ORNL/DSRD-24, pp 75–86
12. Muslea I (2004) Machine Learning for Online Query Relaxation. Proceedings of KDD-2004, Seattle. ACM, pp 246–255
13. Pawlak Z (1981) Information systems – theoretical foundations. Inf Syst J 6:205–218
14. Ras ZW, Dardzinska A (2004) Ontology based distributed autonomous knowledge systems. Inf Syst Int J 29 (1):47–58
15. Ras ZW, Dardzinska A (2006) Solving Failing Queries through Cooperation and Collaboration, Special Issue on Web Resources Access. World Wide Web J 9(2):173–186
16. Ras ZW, Zhang X, Lewis R (2007) MIRAI: Multi-hierarchical, FS-tree based Music Information Retrieval System. In: Kryszkiewicz M et al (eds) Proceedings of RSEISP 2007. LNAI, vol 4585. Springer, Berlin, pp 80–89

Coordination Dynamics

JAMES A. S. KELSO
Human Brain and Behavior Loboratory,
Center for Complex Systems and Brain Sciences,
Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary

Coordination dynamics Coordination Dynamics, defined broadly as the science of coordination, describes, explains and predicts how patterns of coordination form, adapt, persist and change in living things. In coordination dynamics the parts communicate via mutual information exchange and information is meaningful and specific to the forms coordination takes. Coordination dynamics embraces both spontaneous self-organizing tendencies and the need to guide or direct them in specific ways in a single conceptual framework. Life, brain, mind and behavior are hypothesized to be linked by virtue of sharing a common underlying coordination dynamics.

Synergies Synergies (*aka* coordinative structures) are functional groupings of structural elements (e. g. neurons, muscles, joints) that are temporarily constrained to act as a single coherent unit. They arise in many contexts on many levels of biological organization, from the genetic to the social. Synergies are the key to understanding biological coordination and as such are the significant units of coordination dynamics. The synergy hypothesis is an hypothesis about how Nature handles biological complexity.

Self-organization The 'self' in the word self-organization refers to the ability of an open system that exchanges matter, energy and information with the environment, to organize itself. Spontaneous patterns arise solely as a result of the dynamics of the system with no specific ordering influence imposed from the outside and no homunculus-like agent inside. Nonequilibrium phase transitions are the hallmark of self-organization in living things.

Collective variables Collective variables (*aka order parameters* in physics or *coordination variables* in coordination dynamics) are relational quantities that are created by the cooperation among the individual parts of a system. Yet they, in turn, govern the behavior of the individual parts. This is sometimes referred to as *circular* or *reciprocal causality*. In coordination dynamics, the identification of coordination variables depends on the level of description. What is "macro" at one level may be "meso" or "micro" at another.

Control parameters Control parameters refer to naturally occurring environmental conditions or intrinsic, endogenous factors that move the system through its repertoire of patterns and cause them to change. Experimentally, you only know for certain you have identified a control parameter if, when varied, it causes the system's behavior to change qualitatively or discontinuously, i. e., to change its functional state.

Metastability Metastability arises due to broken symmetry in the coordination dynamics where the unstable and stable fixed points (phase- and frequency-locked states) have disappeared due to tangent or saddle-node bifurcations leaving behind only remnants of the fixed points. Metastability is the simultaneous realization of two competing tendencies: the tendency of the components to couple together and the tendency of the components to express their intrinsic independent behavior. Metastability has been hailed as a new principle of organization in complex living systems, including the brain, reconciling apparent contraries such as individual and collective, part and whole, competition and cooperation, integration and segregation, and so forth.

Definition of the Subject

> *Even before man speculated about the nature and sources of his own experiences, he was probably curious about the agencies by which animal motion was affected. Life and motion are almost synonymous terms.* Franklin Fearing [48]

In his preface to the *Principia*, Isaac Newton speculated that not just the motions of the planets, the Moon and the tides could be explained by the forces of attraction and repulsion, but all other natural phenomena as well. Despite the hubris, "self-motion", Newton recognized, "was beyond our understanding" [64]. Three and a half centuries later, the problem remains: the goal-directed coordinated movements of animals are not mere mechanical consequences of the laws of physics, at least as we know them. Despite the many remarkable applications of physics to biology and entire fields devoted to them (e. g. biomechanics, biophysics, nanophysics, etc.) and despite the successes of modern molecular biology, the great unresolved problem of biology remains: *how complex living things are coordinated in space and time*. To recognize that coordination is often purposeful and goal-directed does not at all mean a return to vitalism. But it does pose the scientific challenge of extending physics to understand coordination in living things. Coordination dynamics is a response to this challenge: it is a conceptual framework and research program that deals fundamentally with *animate* (and animated) self-organizing dynamical systems (see also [181,182]). That is, it deals with animate organisms anchored to and engaged in their surrounding worlds. Table 1 compares some of the features of classical mechanics and coordination dynamics. The Table is not intended to be inclusive or to convey the idea that there have been no candidate "paradigms" between classical mechanics and coordination dynamics. For present purposes, "the complex systems" paradigm may represent the most recent break from or extension of classical mechanics. In the complex systems paradigm, self-organization would replace organization, open systems would replace closed systems, change, disorder and process would replace stasis, order and equilibrium, etc. [163]. Coordination dynamics goes a step further. In coordination dynamics it is not organization versus self-organization, order versus disorder, closed versus open systems, reductionism versus emergentism, etc., but rather both aspects that are necessary for an exhaustive account of phenomena and a deeper understanding of coordination in living things. Hence, rather than view these features in opposition, they are better viewed as complementary [100,104,110].

Coordination represents one of the most striking, most taken for granted, yet least understood features of all living things. Imagine a living system whose component parts and processes, on any level of description one choose to examine, did not interact with each other or with their surrounds. Such a collective "cell," "organ," "organism" or "society" would possess neither structure nor function. Coordination can be seen almost everywhere we look, whether in the regulatory interactions among genes that affect how an organism develops and how some diseases like cancer occur, the tumbling and twisting of the bacterial flagellum, the coordinated responses of organisms to constantly varying environmental stimuli, the coordination among nerve cells that produce basic forms of locomotion in vertebrates and invertebrates, the coordination among cell assemblies of the brain that underlies our

Coordination Dynamics, Table 1
Some complementary features of classical mechanics and coordination dynamics

Classical Mechanics	Coordination Dynamics
machines; organized	"organisms "; self-organized
inanimate	animate
motion	coordination (function); animation
matter	materially instantiated; organization
forces	information (semantic) and information exchange; couplings
fundamental dimensions (M, L, T)	collective or coordination variables
space, time	spatiotemporal
linear; smooth dynamics	essentially nonlinear; bifurcations; multi- and metastable
deterministic	fluctuations/variability play key role
decomposable;	"emergent "; synergistic;
motion of whole = sum of motion of the parts	motion of whole > sum of motion of the parts
homogeneous parts, elements, components	heterogeneous units and connectivity
micro versus macro level distinction	level independent strategy; 'one level down '
fixed laws	extensible laws; regularities
context-free	context-dependent
machine/artifactual perspective on order	organic/natural perspective on order

awareness, ability to think, remember, decide and act, the miraculous coordination between the lungs, larynx, lips and tongue that belies a child's first word, the learned coordination among fingers and brain that allows the skilled pianist to play a concerto, the congruence of motion and emotion in dance, drama and everyday life, the coordination between people – like rowers in a racing eight – working together to achieve a common goal. Everything is coordinated.

What do we mean by the word coordination? Coordination is not only spatial and temporal order. Rather, it refers to different kinds and degrees of *functional* order among interacting parts and processes in space and time. Newtonian mechanics may define limits on how biological systems are coordinated, but it says nothing about their functional organization, per se. *How are complex living things coordinated in space and time? What is the nature of the basic interactions that give rise to patterns of coordinated behavior? Why do they take the form they do*? These questions lie at the heart of understanding coordination. Given the ubiquity of coordinated behavior in living things, one might have expected its lawful basis to have been uncovered many years ago. However, it is only in the last 25 years or so, and under quite peculiar circumstances, that basic laws for a quantitative description of coordination have been found.

Introduction

A centipede was happy quite,
Until a frog in fun said:
"Pray tell which leg comes after which?"
This raised her mind to such a pitch,
She lay distracted in the ditch,
Considering how to run. Anon

Coordination Dynamics – the science of coordination – refers to the concepts, methods and tools used to describe, explain and predict how patterns of coordination form, adapt, persist, and change in living things. It is about identifying coordinated patterns in the behavior of living things and expressing how these patterns evolve and change in terms of dynamical laws or rules. The dynamics here refers to equations of motion for key *coordination variables* or *order parameters* [73] that characterize coordinated patterns of behavior on multiple levels of description. As the name implies, the dynamics deals with coordination, not (or not only) with matter in motion: *coordination dynamics* (see Table 1). Through an intimate relationship between theory, experiment, analysis and modeling, Coordination Dynamics seeks to identify the laws, principles and mechanisms underlying coordinated behavior within and between different levels of organization, explicitly addressing the connection between levels. Thus, a goal of Coordination Dynamics is to identify the nature of the functional and context-dependent coordination *within* a part of a system (e. g., the firing of cells in the heart or neurons in a part of the brain), *between* different parts of a system (e. g., parts of the brain, parts of the body, members of an audience) and *between* different kinds of system (e. g., stimuli and responses, organisms and environments, humans and robots, etc.). In coordination dynamics, the coupling between things is realized by many mechanisms, but is fundamentally informational in nature. That is, Coordination dynamics deals specifically with *meaningfully* coupled, self-organizing systems: the parts communicate via mutual information exchange and information is meaningful and specific to the forms coordination takes. As a source of biological order and pattern, self-organization has received much less attention than the teleonomic, directed nature of living things captured by terms like "program", "blueprint", "template" and so forth. Instead of treating them as opposing theories, Coordination Dynamics unites the spontaneous, self-organizing nature of coordination and the obviously directed, agent-like properties characteristic of animate nature into a single framework [100,104,110]. It does this by studying how functionally meaningful information arises from spontaneous self-organizing processes and how it in turn modifies, guides and directs them.

Coordination Dynamics is both multi- and interdisciplinary, engaging relevant aspects and subfields of psychology, philosophy, biology, neuroscience, computer science, engineering, mathematics and physics. For Coordination Dynamics, a complete understanding of coordination phenomena on *any* given level of description requires: i) specifying the individual coordinating elements and their properties; ii) identifying key parameters, boundary and task conditions that constrain coordination; and (iii) showing how interactions among coordinating elements produce or generate patterns of coordination. By demonstrating in specific cases how the nonlinear coupling among the parts produces coordinated behavior Coordination Dynamics demystifies the popular term "emergence". Even more subtly, a certain régime of Coordination Dynamics called *metastability* resolves the longstanding dichotomy between the whole and the parts by explicitly showing how the individual parts of the system may retain a certain degree of autonomy while still coordinating as a whole. To the extent that they transcend the particular mechanisms through which coordination is realized, the principles of Coordination Dynamics may be said to be "universal" and hence have the potential to describe and explain coordinated behavior in a number of fields ([75]; see also ► Movement Coordination and ► Social Coordination, from the Perspective of Coordination Dynamics). In the case of movement coordination, for example, Coordination Dynamics provides the basic laws for a quantitative description of phenomena that are observed when humans interact in a certain way with themselves, with other humans and with their environment [59,100].

History of Coordination Dynamics: Synergy and Rhythmic Order

Coordination Dynamics arose as a response to the fundamental problem of control and coordination in complex, biological systems: the problem of degrees of freedom. Consider an ordinary movement of the human body. The body itself consists of over 790 muscles and 100 joints that have co-evolved in a complex environment with a nervous system that contains $\sim 10^{12}$ neurons and neuronal connections. On the sensory side, billions of receptor elements embedded in skin, joints and muscles inform the mover about his movement. Clearly, any ordinary human activity requires the cooperation among very many structurally diverse elements – a miracle that we take for granted (e. g. [164]). How does nature compress the very high dimensional state space of such a complex system into something lower dimensional and controllable? An attractive hypothesis proposed by the Russian physiologist Bernstein (1896–1966 [17]) is that in complex living systems, the individual elements are not controlled directly but are rather organized into collectives called *synergies*. Synergies are functional groupings of structural elements (e. g. neurons, muscles, joints) that are temporarily constrained to act as a single coherent unit. Just as new states of matter form under certain conditions when a group of atoms behaves as a single particle (e. g., the Bose–Einstein condensate) so a new state of biological *function* emerges when large ensembles of different elements form a synergy. The synergy hypothesis is therefore an hypothesis about how Nature handles biological complexity. Synergies may appear in many contexts on many levels of biological organization, from the genetic to the social. Depending on context, synergies can accomplish different functions using some of the same anatomical components (e. g., those used for speaking and chewing) and the same function using different components (e. g. writing one's name with a pen attached to the big toe). Once assembled, the degrees of freedom composing a synergy take care of themselves in a relatively autonomous organization. The assembling and disassembling of synergies may be said to be "soft" demanding little energy: synergies are ready to become something else at an instant. They are the "atoms", the significant units of biological function [105,107].

The hallmark of a synergy is that the individual elements adjust to mutual fluctuations and to fluctuations in the external force field (and more generally, the synergy's environment) in order to sustain integrity of function. As a consequence, natural variations (which from the scientist's view may be seen as "errors") that occur in the individual elements of the synergy are compensated by adjustments ("covariations") in other members of the synergy in such a way as to maintain a given function stable. Retaining stability is, for a synergy, the retaining of functional integrity [105,204]. Stability, therefore, plays a key role in coordination dynamics, where the great challenge is to discover what the stability is of. Since the key variables of coordination are not known a priori in living things, they must be identified through empirical research. This, as we shall see, follows a particular strategy.

In the late 1970s and early 80s technological developments and sophisticated computer methods for analyzing complex, multidegree of freedom movements enabled stringent experimental tests of the synergy hypothesis to be carried out [112,120]. Invariably, the experiments, which ranged from postural control to speech production and complex finger and limb movements ([105,107,136,201,204] for reviews), showed: a) that a perturbation to any part of the putative synergy is rapidly compensated for by remotely linked elements in such a way as to preserve system function; b) that the same elements are used in different functions in different ways; c) that different elements may accomplish the same function; and d) that the adjustments observed could in all cases be said to be meaningful, task- and context-specific. All this evidence for the existence of synergies attests to the tremendous redundancy or degeneracy of biological systems [46].

All scientific journeys begin with a single step. The identification of synergies as significant structural–functional units of biological coordination was an important one for the development of Coordination Dynamics. Synergies simplify control by reducing the number of variables that must be independently specified: as constraints, they make control and coordination of complex, multivariable systems possible. But understanding goes far beyond identification. How are synergies formed? What principles govern their assembly? And how does one synergy change spontaneously to another as internal or external conditions change? How can distinct synergies co-exist among the same set of components? And how are individual components of the synergy engaged and disengaged as circumstances change?

Insights into these questions come from the work of a largely unheralded genius called Erich von Holst [215], a behavioral physiologist who spent his life studying coordination in a wide variety of creatures – from worms to man. Von Holst's research will not give us answers to all the questions about synergies but it will provide key insights into the essence of coordination and a stepping stone to finding the underlying principles. Using an experimental model system that allowed him to measure an elementary synergy – the to and fro motions of the fins of the swimming dogfish *Labrus* under carefully controlled water flow conditions – von Holst identified at least three basic types of coordination: *absolute* coordination, in which component parts are locked together in time (like the synchronized flashing of fireflies, a couple making love or phase synchrony between parts of the brain); partial or *relative* coordination, in which the component parts 'lock in' only transiently and then break apart as circumstances change (like a little boy walking hand in hand with his father on the beach; dad must slow down and/or son add a step so that they can stay together); and no coordination at all, in which the component parts behave quite independently (as occurs in the locomotion of millipedes and centipedes when the same little boy chops off their middle legs, or perhaps after persistent, long term practice in playing the piano or the violin). Various blends, mixes and transitions between these coordinated behaviors were also observed – always matching the exigencies of the internal and external environment.

Why might some kind of common principle exist for such diverse phenomena? The reason is that the same basic coordination phenomena seem to cut across a wide range of levels, creatures and functions. Among those observed are: patterns of coordination remain stable in time despite continuous, and often unexpected perturbations; the ease with which component parts and processes are flexibly engaged and disengaged as functional demands or environmental conditions change; the existence of multiple coordination patterns – so-called multifunctionality – that effectively satisfy the same set of circumstances; the selection of particular coordination patterns that are exquisitely tailored to suit the current needs of the organism; adaptation of coordination to changing internal and external contingencies; smooth and abrupt transitions from one coordinated pattern to another; transitions from partially to fully coordinated patterns and vice-versa; persistence of a coordinated pattern even when conditions that led to the establishment of the pattern have changed (a kind of memory), and so forth. Such phenomena appear so commonly and so consistently as to suggest the existence of an underlying lawfulness or regularity that transcends the differences between systems. Nature, as the saying goes, operates with ancient themes. Or maybe nature just is what it is.

Conceptual Foundations of Coordination Dynamics: Self-organizing Dynamical Systems

Given we accept the empirical facts about synergies and rhythmic order in the nervous system and the movements of living things, what concepts, methods and tools do we use to understand them? Were synergies simply rigid mechanical entities built by an engineer or an intelligent designer, control theory with its programs, reference levels, comparators, feedforward and feedback error correcting mechanisms and so forth, might have seemed an obvious place to look for explanatory tools. A program instructs the parts of a system what to do and when to do it. Feedback may then be used to correct errors in the outcome. But now what? The system receives an error signal: How does it know which of its many parts to correct? In a complex system composed of very many components it may take a very long time to come up with a solution, a problem computer scientists refer to as an NP-complete problem, where NP means "non-deterministic polynomial time complete". Biology with its degeneracy and redundancy has no such problem. "Error" signals from one part of a synergy are rapidly compensated by other members. So if anything, the *machine perspective* on order and regulation (Table 1) seems to compound the problem rather than solve it.

Coordination Dynamics takes its inspiration from *a natural, organic perspective*, i. e., how nature handles complexity (Table 1). It is well-known that pattern formation in open, nonequilibrium physical and chemical systems such as fluids, lasers and chemical reactions can emerge spontaneously. These patterns arise solely as a result of the dynamics of the system with no specific ordering influence imposed from the outside environment and no homunculus-like agent inside. Such "self-organized" pattern formation is a cooperative phenomenon that results from the interaction of large numbers of interacting subsystems [73,152]. It should be stressed here that there is no "self" inside the system responsible for prescribing or coding the emergent pattern. The 'self' in self-organization comes from the fact that given the ability to exchange matter, energy and information with the environment, the system organizes itself. That the organism is an open system is one of two essential criteria for life postulated by Francis Crick in *Of Molecules and Men* [33], yet it has received much less attention in biology than Crick's other criterion, the need for organisms to reproduce and pass on 'copies' of themselves to their descendants. Here already we see a dichotomy between a complex system's natural ordering tendencies and the need (at least in living systems) to guide that order in specific ways. Coordination dynamics (Table 1) reconciles this dichotomy by viewing these two fundamental aspects as complementary ([100,104,110]).

In his general theory of nonequilibrium phase transitions called "synergetics" Haken [73] showed that close to critical points where a so-called *control parameter* crosses a threshold, very complex, high-dimensional systems can be completely described by a much lower dimensional dynamics specified in terms of only a few collective variables or *order parameters*. What do these terms mean? *Control Parameters* refer to naturally occurring environmental conditions or intrinsic, endogenous factors that on first blush appear analogous to what an experimental scientist might call an independent variable. But the concept is entirely different, and the implications for experimental design far reaching [96]. The role of control parameters is to move the system through its repertoire of patterns and cause them to change. In fact, you only know for certain you have identified a control parameter if, when varied, it causes the system's behavior to change qualitatively or discontinuously, i. e., to change state. In a dynamical system, when a parameter changes smoothly, the attractor in general also changes smoothly. Sizeable changes in the input have little or no effect on the resulting output. However, when the control parameter passes through a critical point or threshold in an intrinsically nonlinear dynamical system an abrupt change in the attractor can occur. This sensitive dependence on parameters is called a *bifurcation* in mathematics, or a *nonequilibrium phase transition* in physical theories of pattern formation [73]. Indeed, control parameters are often referred to in mathematics as *bifurcation parameters*. Qualitative change does not mean that quantification is impossible. To the contrary, qualitative change is at the heart of pattern formation and, provided care is taken to evaluate system timescales (e. g., how quickly the control parameter is changed relative to the typical time of the system to react to perturbations; see [121]) quantitative predictions ensue that can be tested experimentally (see Sect. "The Theoretical Modeling Strategy of Coordination Dynamics: Symmetry and Bifurcations").

Collective variables are relational quantities that are created by the cooperation among the individual parts of a system. Yet they, in turn, govern the behavior of the individual parts. This is sometimes referred to as *circular* or *reciprocal causality*. In self-organizing systems the stranglehold of linear causality is broken. At best, simple cause-effect relations are the exception, not the rule. Depending on where the system lives in the space of its parameters, many causes can produce the same effect or the same cause can have multiple effects. One can intuit why the concept of collective or coordination variable is central to a science of coordination. The reason is that interactions in such systems are so complicated that understanding may only be possible in terms of system-specific collective or coordination variables. The latter are not necessarily "macroscopic quantities". In coordination dynamics, the identification of coordination variables depends on the level of description. What is "macro" at one level may be "meso" or "micro" at another. This strategy of folding together all aspects within the dynamics of collective or coordination variables embraces the full complexity of living things on a given level of description without proliferating arbitrary divisions (for a nice discussion see [203]).

In nonequilibrium systems, the enormous compression of degrees of freedom near critical points arises because events occur on different timescales: the faster individual elements in the system become "enslaved" to the slower, "emergent" collective variables [73]. Alternatively, one may conceive of a hierarchy of timescales for various processes involved in coordination. On a given level of the hierarchy are coordination variables subject to constraints (e. g., of the task) that act as boundary conditions on the coordination dynamics. At the next level down are component processes and events that typically operate on faster timescales. Notice for the 'tripartite scheme' of Coordination Dynamics (see pp. 66–67 in [100]) the key is to choose a level of description and understand the relation between adjacent levels, not reduce to some "fundamental" lower level (Table 1). In coordination dynamics, no level is any more or less fundamental than any other. A complete description of a phenomenon always requires three adjacent tiers: The boundary conditions and control parameters that establish the context for particular coordination phenomena to occur; the collective level and its dynamics; the component level and its dynamics including the nonlinear coupling between components.

Dynamic instability is the generic mechanism underlying self-organized pattern formation and change in all (open) systems coupled to their internal or external environments [153]. Near instability the individual elements, in order to accommodate to current conditions, must order themselves in new or different ways. The patterns that emerge at nonequilibrium phase transitions may be defined as attractive states of the collective variable dynamics. That is, the collective variable may converge in time to a certain limit set or attractor solution, a nonequilibrium steady state. Attractors can be *fixed points*, in which all initial conditions converge to some stable rest state. Attractors can also be *periodic*, exhibiting preferred rhythms or orbits on which the system settles regardless of where it starts. Or, there can be so-called *strange* attractors; strange because they exhibit *deterministic chaos*, a type of irregular behavior resembling random noise, yet often containing pockets of quite ordered behavior. Stable fixed point, limit cycle and chaotic solutions as well as a wide variety of other transient and irregular behaviors are possible in the *same* system, depending on the values of control parameters (and their time dependence). Moreover, fluctuations are always present, constantly testing whether a given pattern is stable. Fluctuations are not just noise; rather, by probing the stability of existing states they allow the system to discover new, more adaptive patterns that suit the prevailing circumstances (boundary conditions, control parameters; Table 1).

How might these conceptual tools aid our understanding of biological coordination? On first blush, it might seem a gigantic leap from the physics and mathematics of pattern formation in nonequilibrium systems to the problem of coordination in living things. Yet in science, analogy often plays a major role in bringing about conceptual breakthroughs. Although initially the analogy may seem far-fetched, great science often starts with a vague idea which, when followed by crucial experiments and mathematical theory renders the vague idea exact. A key aspect to appreciate is that cooperative phenomena in physical systems are typically independent of the particular molecular machinery or material substrate that instantiates them. This is because the elementary components are the same, i. e. homogeneous. On the other hand, in living, evolved things the component elements are often quite different. Thus any theory of coordination of living things will have to take into account the heterogeneity of its component elements. Perhaps as a consequence of inherent heterogeneity (and no doubt the advancement of technology) the tendency in biology is to focus more and more on specific processes at ever smaller and smaller scales. As a result, building huge data bases may sometimes appear to take precedence over finding scientific laws [55].

Finding Dynamical Laws of Coordination

What if biological coordination were shown to be a self-organized phenomenon? Might that be a springboard to finding laws of coordination? In the sense of T.S. Kuhn [131] such questions appear to call out for a new paradigm, special entry points where irrelevant details may be pruned away exposing the essential aspects one is trying to understand. Inspired by synergetics (and paradoxical though it may seem) the key to determining if coordination as a self-organized phenomenon is to focus on *qualitative change*, places where abrupt switches or bifurcations in coordination occur. Qualitative change is crucial because it affords a clear distinction between one coordination pattern and another, thereby enabling one to identify the key *collective variables or order parameters* that define coordination states and their coordination dynamics. If a complex system is changing smoothly and linearly it is hard to distinguish the variables that matter, so-called state variables, from the ones that don't. Qualitative change may also be used to infer relevant quantities in more naturalistic settings. In situations where many variables may be changing in uncontrolled ways, the one(s) that change(s) abruptly are likely to be the most meaningful, both for the phenomena themselves and our understanding of them [96]. Likewise, any parameter that induces qualitative behavioral change qualifies as a *control parameter*. This is the reason why stability is so important. As a control parameter crosses a critical value the previously stable pattern becomes unstable and the system switches to a different pattern that is now stable beyond the critical point. The quite general predictions of nonequilibrium phase transition theory are a strong enhancement of fluctuations (critical fluctuations) and a strong increase in the pattern's relaxation time (critical slowing down) as the transition is approached. Obviously, if nonequilibrium phase transitions are a basic mechanism of self-organization and if, as hypothesized, the forces of evolution and self-organization form the core of biological order and coordination, it should be possible to discover nonequilibrium phase transitions and their signature features in real experiments. If not, the theory that coordination in living things is due fundamentally to self-organization must go the way of all beautiful theories that are negated by the facts.

Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions

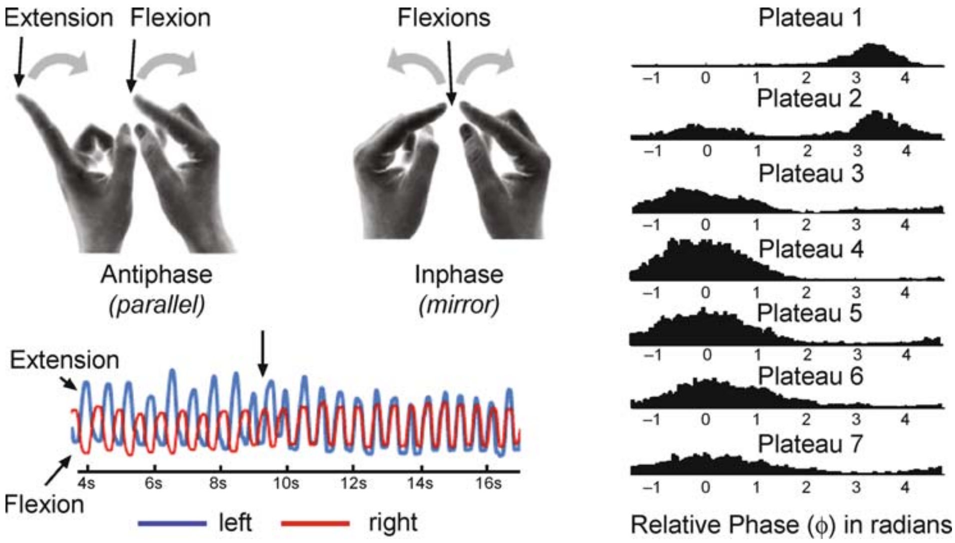
In coordination dynamics, the payoff of knowing collective variables and control parameters is high: they enable one to obtain the dynamical rules of behavior on a chosen level of description. By adopting the same strategy "one level down", the individual components and their dynamics may be studied and identified. It is the nonlinear interaction between the parts that creates coordinative patterns of the whole thereby building a bridge across levels of description (Table 1). This ability to derive phenomena from lower levels of description is at the core of what scientists usually mean by the word "understanding". In general, in complex living systems it is difficult to isolate the components and study their dynamics. The reason is that the individual components seldom exist outside the context of the functioning whole, and have to be studied as such.

If phase transitions hold the key to finding laws of coordination, where should we look for them? A central criterion for a law-based approach to coordination is reproducibility of the phenomenon in question. Although not everything is rhythmic, rhythms represent a wide variety of coordinated behaviors in a very large number of different biological systems at very many levels [24,70,71,97,119,174,212,215] and seem like an obvious entry point. One only has to look at the extensive field of so-called "central pattern generators" (CPGs) in invertebrate and vertebrate neurobiology to find remarkable similarities in the patterns that living creatures produce. Terms such as "swimming" CPG, "flight" CPG, and "locomotor" CPG reflect the reproducibility of patterns and their functional significance. Synchronization and desynchronization, frequency- and phase-locking are ubiquitous features of such patterns reflecting a high degree of neural and behavioral coordination (see Sects. "History of Coordination Dynamics: Synergy and Rhythmic Order" and "Coordination of Multiple Components: From Quadrupeds to Brains").

What then of phase transitions? And what connection exists, if any, between rhythms and phase transitions? It is well-known that quadrupeds and indeed many creatures including birds and fish exhibit characteristic gaits and may switch flexibly between them depending on circumstances. In the neurobiology literature, a key question is always "where are the switches in this thing"? [1]. Rather than assume the existence of switches, a priori, the scientific approach of coordination dynamics is to investigate the necessary and sufficient conditions that give rise to switching. Inspired by theories of self-organization in nature, coordination dynamics asks if switching may take the form of a nonequilibrium phase transition. The idea is not so far fetched as it seems. Many years ago, order-order transitions were hypothesized by Erwin Schrödinger [178] to be a crucial principle of biological organization and hypothesized to be the "new laws in the organism" [109].

Three Deceptively Simple Experiments

To investigate order-order transitions experimentally, consider an experimental paradigm introduced some years ago in which human beings are asked to move their two index fingers back and forth rhythmically [94,95]. In one condition (call it parallel, Fig. 1) they are told to alternate finger movements at a comfortable rate, one finger flexing in time as the other extends. In another separate condition (call it mirror) they are told to flex both fingers together and extend both fingers together at the same time. The key part of the experiment is that participants are instructed to increase the speed at which they perform these movements. For better experimental control a pacing metronome whose frequency can be systematically increased (say every 10 cycles called a plateau) may be used for subjects to follow. The main results are shown in Fig. 1 and described in the figure's caption.



Coordination Dynamics, Figure 1

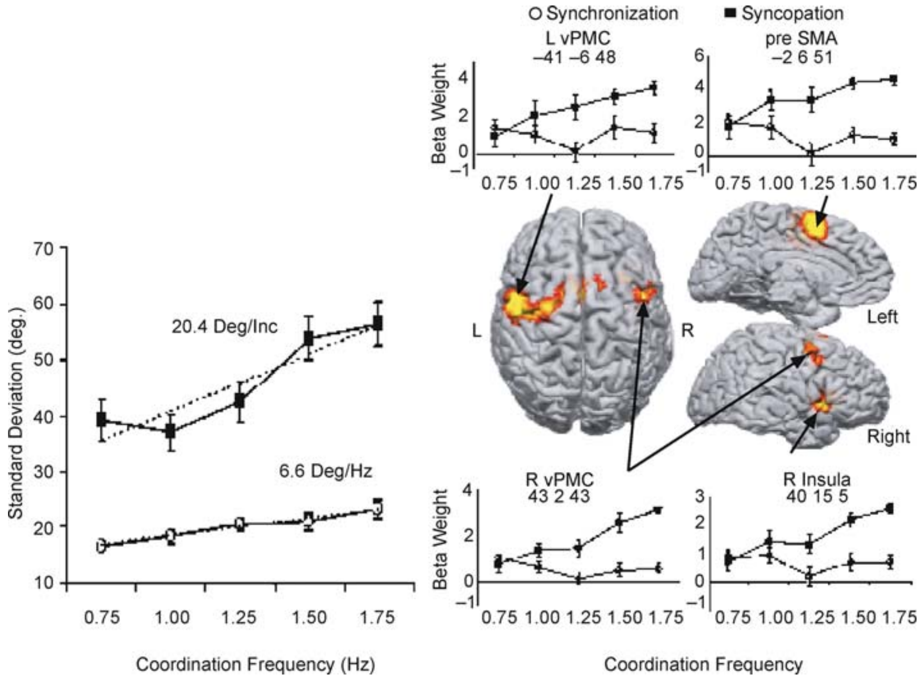
Phase transitions in bimanual movements. *Left side*. On the top are the two experimental conditions (parallel, antiphase and mirror, in phase) in the Kelso paradigm. The lower plot shows the time series of the finger movements in a representative run. As rate increases, trials initiated in the antiphase pattern spontaneously switch to the in-phase mirror pattern. The critical frequency is identified with an arrow. In contrast, trials initiated in the in-phase pattern do not switch as frequency increases (not shown). *Right side*. Distributions of relative phase between finger movements for plateaus of increasing frequency of movement. Initially the relative phase is concentrated at π radians, indicating antiphase is a stable pattern. On *plateau 2*, relative phase is concentrated around 0 and π radians, showing the bistability of antiphase and inphase coordination. For higher frequency plateaus the relative phase is concentrated at 0 radians, indicating that inphase is the only stable pattern of coordination (adapted from [7] with permission)

Experimental studies of bimanual rhythmic movement demonstrate that humans in the absence of learning procedures are able to produce two patterns of coordination at low frequency values, but only one – the symmetrical, in-phase mode – as frequency is scaled beyond a critical value. This is a remarkable result given the complexity of the nervous system, the body and their interaction with the world.

Consider another example, this time involving a human coordinating with an environmental signal [115,218]. In this experimental setup a single limb or finger is moved such that peak flexion occurs in between the beats of a pacing metronome, i. e. in a syncopated fashion. When the metronome frequency is increased, once again a critical value is reached where participants switch spontaneously to coordinating peak flexion on the beat, i. e. in a pattern of synchronization with the metronome. No such switching occurs when subjects begin in the synchronized mode of coordination. We may refer to these effects as a very basic example of *coordination between an organism and its environment*.

Now, consider the case of two people interacting with each other, an elementary form of interpersonal or *social coordination*. In this situation, each individual is instructed to oscillate a limb (the lower leg in this case) in the same or opposite direction to that of the other person [170]. In order to do the task, there must be a medium of interaction (vision, sound, touch, smell…) through which humans can couple. In this case, the two people watch each other (for details of this and other work see ► Social Coordination, from the Perspective of Coordination Dynamics). Then, either by an instruction from the experimenter or by following a metronome whose frequency is systematically increased, the social dyad speeds up their movements. When moving their legs up and down in the same direction, the two members of the dyad remain synchronized across a broad range of speeds. However, when moving their legs in the *opposite* direction (one person's leg extending at the knee while the other's is flexing), such is not the case. Once again at a certain critical rate participants change their behavior spontaneously so that their legs now move in the *same* direction.

The 'nonlinear' paradigm of coordination dynamics as illustrated in these three simple experiments has led to a wide range of investigations in many fields and a surprising variety of contexts (see Sect. "Collective Minds" for brief summary; also Books and Reviews) including detailed studies of underlying brain mechanisms using the full armamentarium of imaging technologies. Although



Coordination Dynamics, Figure 2

A pattern of generic results from combined brain and behavioral studies of coordination (adapted from [81]). On the *left* are behavioral results showing how relative phase variability is initially higher and increases more rapidly with the control parameter of frequency for syncopated antiphase (*filled squares*) than synchronized inphase coordination (*open circles*) with an external event. On the *right* are corresponding brain activations that comprise a network that is exquisitely sensitive to the stability of antiphase coordination. Areas depicted are left and right ventral premotor cortex (vPMC), pre supplementary motor area (pre-SMA) and right insula. The X Y Z Talairach coordinate of the peak voxel for each region is provided. Notice how the behavioral and brain data track each other

the findings would take us too far afield for present purposes ([19,20,191]; for reviews see [59,81]) two particular results are worth noting. First is that the basic paradigm has led to the first direct evidence of phase transitions in the human brain seen using both large scale electrode EEG [143,216] and SQuID arrays [35,61,62,108,123,124]. Second, and even more telling, is that fMRI evidence indicates that regardless of whether one is coordinating the two hands or coordinating with an external signal, a common network concerned with the *stability* of coordination is involved (see [81,156] for reviews).

The pattern of experimental findings described in this section illustrates an important conceptual distinction between coordination dynamics and other theories of self-organization [163]. In the latter, at bifurcation points or phase transitions, the system switches to a new, higher level of organization called a *dissipative structure*. Dissipative structures are so named because, compared with the "simpler" structures or patterns they replace, they require more energy to sustain them (ibid., p.xv). Not so in the order-order transitions of coordination dynamics. In coordination dynamics, the new organization that appears at bifurcation points is 'simpler' than the one it replaces and requires less energy. For example, brain electrical activity actually drops across the antiphase to inphase transition even though the system is being driven faster [124]. In Fig. 2, blood oxygen level dependent (BOLD) activity in certain brain regions is shown to increase as the stability of the antiphase pattern decreases. It is obvious that the increasing metabolic energy demands of the brain for antiphase relative to inphase coordination will diminish once the transition to in phase occurs. Thus, the key principle behind the 'simpler' self-organizing structures that emerge in coordination dynamics are based, not (or not only) upon energy per se as in the theory of dissipative structures, but on the system's *information processing demands*. Intuitively, the antiphase pattern is more difficult to coordinate as rate or frequency is increased causing the system to switch to a pattern that is easier to perform under the current conditions. Importantly, coordination dynamics replaces vague terms such as 'task difficulty' and 'task complexity' by quantitative behavioral measures of stability and quantitative brain measures of BOLD and neuroelectric activity.

The Theoretical Modeling Strategy of Coordination Dynamics: Symmetry and Bifurcations

The three pieces of experimental evidence described above cut across entirely different kinds of things and events (auditory, visual and proprioceptive sensations, finger and leg movements, people and brains, etc.). The common denominator is that all these things and processes are meaningfully coupled together in time under particular boundary conditions (task instructions, environmental context, manipulated parameters, etc.). The phenomena observed hint at an aspect that any basic law should exhibit, namely that although the patterns of coordination observed are realized by different physical structures and physicochemical processes, laws and regularities are abstract and relational.

How then do we go about identifying the actual underlying laws? More specifically, how do we explain the coordinative phenomena observed experimentally? As stressed above, in contrast to certain physical systems like the laser, in biological coordination the path from the microscopic level to collective order parameters is not known and cannot (yet?) be derived from first principles like conservation laws. In coordination dynamics we have to: a) identify the order parameters or coordination variables and their low-dimensional dynamics empirically; b) determine the key control parameters that move the system though its coordinative states; and c) relate different levels though a study of the individual subsystems and their nonlinear interaction.

Determining the dynamics of coordination variables is non-trivial. In all three experimental situations, the relative phase ϕ or phase relation between the component elements appears to qualify as a suitable order parameter or coordination variable. The reasons are as follows: ϕ characterizes the patterns of spatiotemporal order observed, in phase and anti-phase; ϕ changes far more slowly than the variables that describe the individual coordinating components (e. g., position, velocity, acceleration, electromyographic activity of contracting muscles, neuronal ensemble activity in particular brain regions, etc.); ϕ changes abruptly at the transition and is only weakly dependent on parameters outside the transition; and ϕ appears to obey a dynamics in which the patterns may be characterized as *attractors* or *attractive states* of some underlying dynamical system. Since in all cases the frequency or rate clearly drives the system through different coordination patterns without actually prescribing them, frequency qualifies as a control parameter.

Determining the coordination dynamics means mapping observed, reproducibly stable patterns onto attractors of the dynamics. A general strategy is to assume sufficiently higher order dynamics and expand the vector field of these dynamics in a Fourier series:

$$\dot{\phi} = f(\phi) = a_0 + a_1 \sin(\phi) + a_2 \sin(2\phi) + \dots \\ + b_1 \cos(\phi) + b_2 \cos(2\phi) + \dots \,. \quad (1)$$

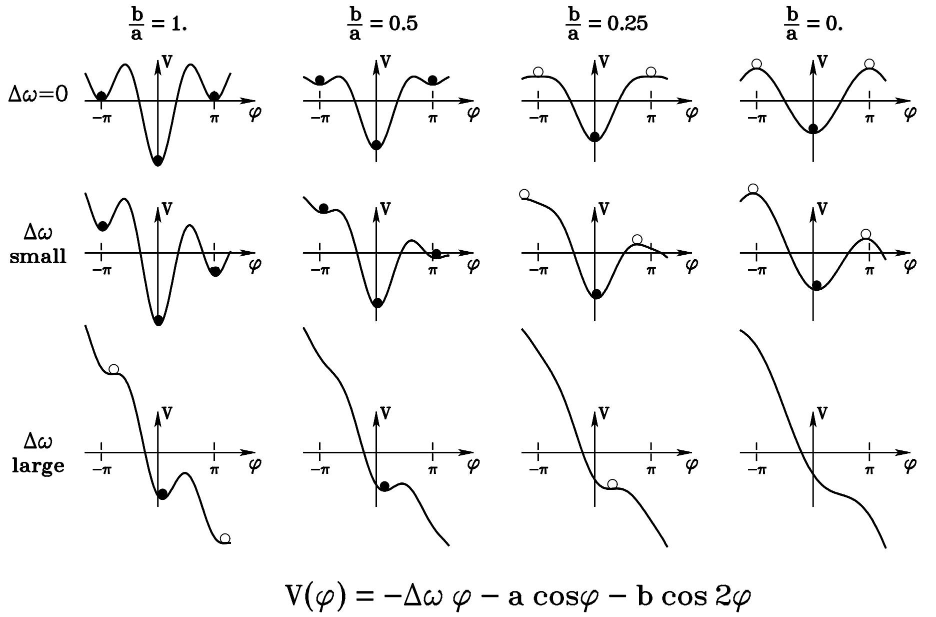
Symmetry may be used to classify patterns and restrict the functional form of the coordination dynamics. Symmetry means "no change as a result of change": pattern symmetry means a given pattern is symmetric under a group of transformations. A transformation is an operation that maps one pattern onto another, e. g. in the first experimental case, left-right transformation exchanges homologous limbs within a bimanual pattern. If all relative phases are equivalent after the transformation, then the pattern is considered invariant under this operation. Symmetry serves two purposes. First it serves as a pattern classification tool allowing for the identification of basic coordination patterns that can be captured theoretically. Given a symmetry group, one can determine all invariant patterns. Second, imposing symmetry restrictions on the dynamics itself limits possible solutions and allows one to arrive at a coordination dynamics that contains the patterns as *different* stationary states of the *same* nonlinear dynamical system. In other words basic coordination patterns correspond to attractors of the relative phase for adequate parameter values. For example, left-right symmetry of homologous limbs leads to invariance under the transformation $\phi \to \phi$ so that the simplest dynamical system that accommodates the experimental observations is:

$$\dot{\phi} = f(\phi) = -a \sin(\phi) - 2b \sin(2\phi) \quad (2)$$

where ϕ is the relative phase between the movements of the two individuals, $\dot{\phi}$ is the derivative of ϕ with respect to time, and the ratio b/a is a control parameter corresponding to the movement rate in the experiment. An equivalent formulation of Eq. (1) is

$$\dot{\phi} = -\partial V(\phi)/\partial\phi \quad \text{with } V(\phi) = -a \cos\phi - b \cos 2\phi \,. \quad (3)$$

In the literature, Eqs. (2) and (3) are the equations at the collective level of the HKB model of coordinated behavior, after Haken, Kelso and Bunz [76,106,139]. Figure 3 (top) allows one to develop an intuitive understanding of the behavior of Eqs. (2),(3), as well as to connect the key concepts



Coordination Dynamics, Figure 3

The potential, $V(\phi)$ of Eq. (3) (with $\Delta\omega = 0$) and Eq. (5) (with $\Delta\omega \neq 0$). *Black balls* symbolize stable coordinated behaviors and *white balls* correspond to unstable behavioral states (see text for details)

of stability and instability in self-organized dynamical systems to the observed experimental facts.

The dynamics can be visualized as a particle moving in a potential function $V(\phi)$. The minima of the potential are points of vanishing force, giving rise to stable solutions of the elementary coordination dynamics. As long as the speed parameter (b/a) is slow, meaning the cycle period is long, Eq. (3) has two stable fixed point attractors, coordinative states at $\phi = 0$ and $\phi = \pm\pi$ rad. Thus, two coordinated behavioral patterns coexist for exactly the same parameter values, the essentially nonlinear feature of *bistability* (Table 1). Such bi- and in general multi-stability is the dynamical signature of *multifunctionality* which can be seen at many levels in living things. As the ratio b/a is decreased, meaning that the cycle period gets shorter as the system speeds up, the formerly stable fixed point at $\phi = \pm\pi$ rad becomes unstable, and turns into a *repellor* (open circles). Any small perturbation will now kick the system into the *basin of attraction* of the stable fixed point corresponding to an in-phase coordination pattern at $\phi = 0$. Notice also that once there, the system's behavior will stay in the in-phase attractor, even if the direction of the control parameter is reversed. This is called *hysteresis*, a basic form of memory in nonlinear dynamical systems.

What about the individual components? Research has established that these take the form of self-sustaining oscillators, archetypal of all time-dependent behavior [12,14,47,92,93]. The particular functional form of the oscillator need not occupy the reader here (see ▶ Movement Coordination which uses empirical facts and symmetry arguments to restrict and thereby identify the component dynamics). More important is the nature of the nonlinear coupling that produces emergent coordination. The simplest, perhaps fundamental biophysical coupling that guarantees all the observed emergent properties of coordination: multistability, flexible switching among coordination states and primitive memory, is:

$$K_{12} = \left(\dot{X}_1 - \dot{X}_2\right)\left\{\alpha + \beta(X_1 - X_2)^2\right\} \,, \qquad (4)$$

where X_1 and X_2 are the individual components and α and β are coupling parameters. Notice that the 'next level up', the level of coordinated behavioral patterns and the dynamical rule that governs them (Eqs. (2) and (3)), can be *derived* from the level below, the individual components and their nonlinear interaction. One may call this *constructive reductionism*: by focusing on adjacent levels, under the boundary constraints of the task, the individual parts can be put together to create the behavior of the whole.

The basic self-organized dynamics, Eqs. (2) and (3) have been extended in numerous ways, only a few of which are mentioned here.

- *Critical slowing down* and *enhancement of fluctuations*. Introducing stochastic forces into Eqs. (2) and (3) ([175,210,211] see Chap. 11 in [74] and [101,121] for a thorough discussion) allows key predictions of coordination dynamics to be tested and quantitatively evaluated [113,121,172]. Critical slowing is easy to understand from Fig. 3 (top). As the minima at $\phi = \pm\pi$ become shallower and shallower, the time it takes to adjust to a small perturbation takes longer and longer. Thus, the local relaxation time is predicted to increase as the instability is approached because the restoring force (given as the gradient of the potential) becomes smaller. Likewise, the variability of ϕ is predicted to increase due to the flattening of the potential near the transition point. Both predictions have been confirmed in a wide variety of experimental systems, including recordings of the human brain ([81,100,174] for review).

- *Symmetry breaking*. Notice that Eqs. (2) and (3) are *symmetric*: the dynamical system is 2π periodic and is identical under left-right reflection (ϕ is the same as $-\phi$). This assumes that the individual components are identical, which as remarked upon earlier, is seldom the case in living things where symmetries are broken all the time.

In terms of the development of the theory, an important experimental example of symmetry breaking is the case of coordinating movement with a visual stimulus: visual stimuli and limb movement are obviously not equivalent. Thus $\phi \rightarrow \phi$ symmetry cannot be assumed. This means that symmetry partners of coordination patterns with systematic phase leads or lags do not coexist at the same parameter values. To accommodate this fact, a term $\Delta\omega$ is incorporated into the dynamics [115]:

$$\begin{aligned} &\dot{\phi} = \Delta\omega - a\sin\phi - 2b\sin 2\phi\,, \quad \text{and} \\ &V(\phi) = -\Delta\omega\phi - a\cos\phi - b\cos 2\phi \end{aligned} \tag{5}$$

for the equation of motion and the potential respectively.

Note that Eq. (5) falls out naturally from an analysis of the oscillators, ω_1 and ω_2, viz.

$$\dot{\phi} = \frac{\omega_1^2 - \omega_2^2}{2\Omega} + (\alpha + 2\beta R^2)\sin\phi - \beta R^2 \sin 2\phi \tag{6}$$

for

$$\Delta\omega = \frac{\omega_1^2 - \omega_2^2}{2\Omega} \approx \omega_1 - \omega_2 \tag{7}$$

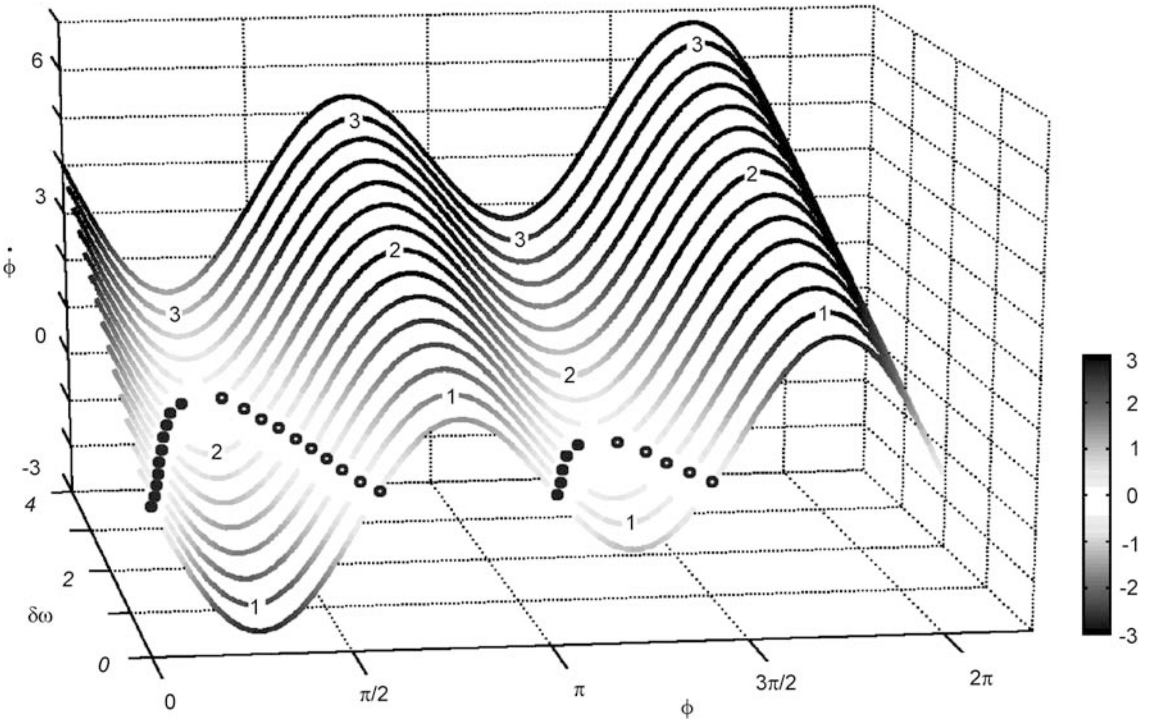
with

$$\begin{aligned} a &= -(\alpha + 2\beta R^2) \\ b &= \tfrac{1}{2}\beta R^2\,. \end{aligned} \tag{8}$$

Small values of $\Delta\omega$ shift the attractive fixed points (Fig. 3 middle) in an adaptive manner. For larger values of $\Delta\omega$ the attractors disappear entirely (Fig. 3 bottom) causing the relative phase to drift: no coordination between the components appears to be possible. Note, however, that the dynamics still retain some curvature (Fig. 3 bottom right): even though there are no attractors there is still attraction to where the attractors used to be. The reason is that the difference ($\Delta\omega$) between the individual components is sufficiently large that they do their own thing, while still retaining a *tendency* to cooperate. The introduction of the symmetry breaking term $\Delta\omega$ in Eq. (5) changes the entire coordination dynamics (layout of the fixed points, bifurcation structure) of the original HKB system in which $\Delta\omega = 0$. This is important to realize because it is the subtle interplay between the coupling ($k = b/a$) and the symmetry breaking term $\Delta\omega$ in Eq. (5) that gives rise to *metastability*.

Sometimes in the literature Eq. (5) is referred to collectively as the Haken–Kelso–Bunz equation. Though convenient, this is technically incorrect and fails to recognize both the intellectual contributions to its extension and the conceptual consequences thereof. For reasons of symmetry and simplicity, the original HKB equation did not contain the symmetry breaking term, $\Delta\omega$ [115] nor did it treat fluctuations explicitly [175] both of which are crucial for capturing the broad range of phenomena observed and testing further predictions. In particular, without $\Delta\omega$ there is: a) no fixed point shift, a sign of adaptation to changing circumstances, see Fig. 3 and 4; b) the bifurcation is a saddle node not, as in the original HKB equation, a pitchfork. These are different normal forms [98]; and c) most important of all, the original HKB equation does not and cannot exhibit metastability which is the key to understanding the complementary relationship between the synergic tendency of the elements to couple (integration) and at the same time to express their individual differences (segregation). The oscillators in the original HKB formulation were identical thereby excluding metastability. For these reasons, it seems wise to refer to Eq. (5) (with its stochastic aspect included) as the extended HKB equation.

Equation (5) is a bit strange. Even though it is an order parameter equation of motion that describes *coordinative* behavior (in words, phi dot is a function of phi), it includes also a parameter ($\Delta\omega$) that arises as a result of differences among the *individual components*. Equation (5) is thus a strange mixture of the whole and the parts, the global and the local, the cooperative and the competitive, the collective and the individual. Were the components identical, $\Delta\omega$ would be zero and we would not see component differences affecting the behavior of the whole (Fig. 3 top row). Equation (5) would simply reflect the behavior of the collective untarnished by component properties, a purely



Coordination Dynamics, Figure 4

Elementary coordination law (Eq. (5)). Surface formed by a family of flows of the coordination variable ϕ (in radians) as a function of its time derivative $\dot{\phi}$ for increasing values of $\Delta\omega$. For this example, the coupling is fixed: $a = 1$ and $b = 1$. When $\dot{\phi}$ reaches zero (*flow line becoming white*), the system ceases to change and fixed point behavior is observed. Note that the fixed points here refer to *emergent coordination states* produced by nonlinearly coupled elements. Stable and unstable fixed points at the intersection of the flow lines with the isoplane $\dot{\phi} = 0$ are represented as filled and open circles respectively. Three representative *lines labeled 1 to 3* illustrate the different régimes of the coordination dynamics. Following the *flow line 1* from left to right, two stable fixed points (*filled circles*) and two unstable fixed points (*open circles*) exist. This flow belongs to the multistable (here bistable) régime of dynamics. Following *line 2* from left to right, one pair of stable and unstable fixed points is met on the left, but notice the complete disappearance of fixed point behavior on the right side of the figure. That is, a qualitative change (bifurcation; phase transition) has occurred due to the loss of stability of the coordination state near antiphase, π rad. The flow now belongs to the monostable régime. Following *line 3* from left to right, no stable or unstable fixed points exist yet a subtle form of coordination – neither completely ordered (synchronized) nor completely disordered (desynchronized) – still remains. This is the metastable régime

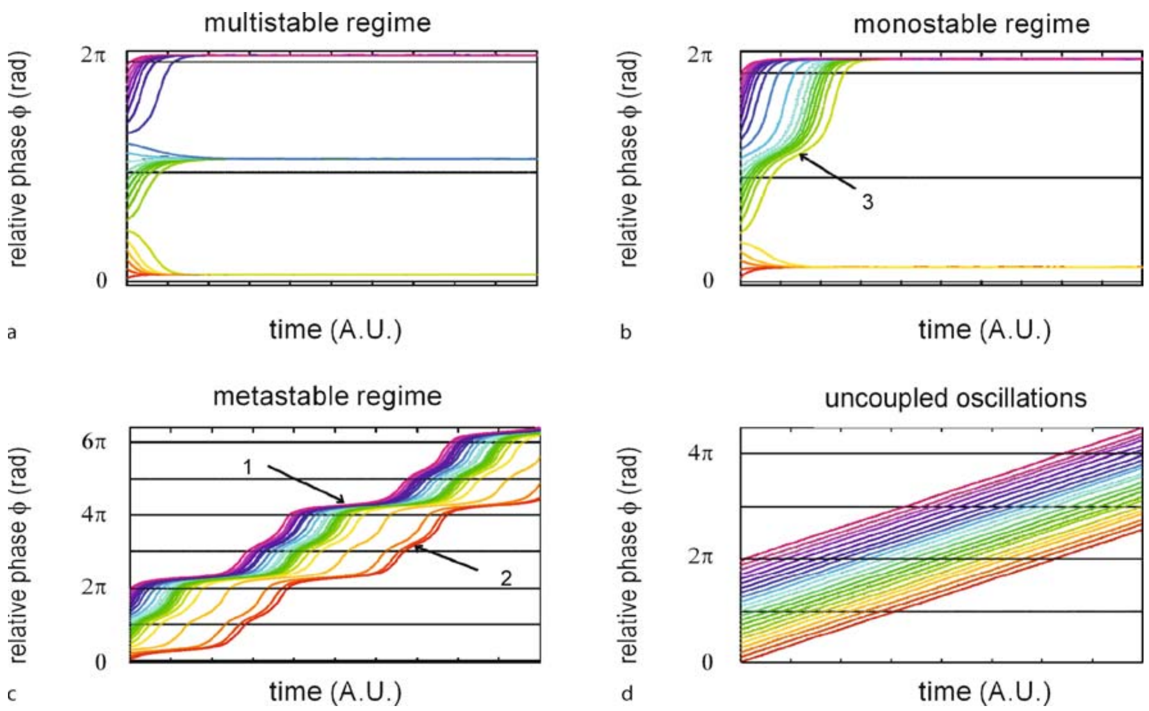
emergent interaction – the HKB equation. It is the fact that *both* the components *and* their (nonlinear) interaction appear at the same level of description that gives rise to the array of coexisting tendencies characteristic of metastability. The history of coordination (Sect. "History of Coordination Dynamics: Synergy and Rhythmic Order") may now be seen in a new light: Eq. (5) is a basic representation of a *synergy*, a low dimensional dynamic of a metastable organization in which the tendency of the parts to act together coexists with a tendency of the parts to do their own thing (see Chap. 4 in [100]). If indeed the synergy is a unit of life and mind as proposed in [105,107] then it is metastability that endows the synergy with robustness and flexibility, enabling the same parts to participate in multiple functions.

Metastable Coordination Dynamics

From States to Tendencies

Etymologically, 'metastability', comes from the latin '*meta*' (beyond) and '*stabilis*' (able to stand). In coordination dynamics, metastability corresponds to a régime near a saddle-node or tangent bifurcation in which stable coordination states no longer exist (e. g., in-phase synchronization where the relative phase between coordinating components lingers at zero), but attraction remains to where those fixed points used to be (see Fig. 3, bottom row). This gives rise to a dynamical flow consisting of *both* phase trapping and phase scattering.

To best visualize the emergence of metastability, Fig. 4 shows the flow of the dynamics for the elementary coor-



Coordination Dynamics, Figure 5

How the key coordination variable or order parameter of the elementary coordination law (Eq. (5)) behaves over time. Shown is a family of trajectories of the relative phase ϕ over time (in Arbitrary Units) arising from a range of initial conditions sampled between 0 and 2π radians, in the multistable (a), monostable (b) and metastable régimes (c) of Eq. (5). For the uncoupled case (d) the trajectories never converge indicating that the oscillations are completely independent of each other. Trajectories in the multistable régime (a) converge either to an attractor located slightly above $\phi = 0$ rad modulo 2π or to another attractor located slightly above $\phi = \pi$ rad modulo 2π. In the monostable régime (b), trajectories converge to an attractor located slightly above $\phi = 0$ rad modulo 2π. In the trajectories of relative phase for the metastable régime (c unwrapped to convey continuity), there is no longer any persisting convergence to the attractors, but rather a succession of periods of rapid drift (*escapes*) interspersed with periods inflecting toward, but not remaining on the horizontal (*dwells*). Note dwells near $\phi = 0$ rad modulo 2π in the metastable régime (e. g. dwell at about 4π rad annotated 1 in c) and nearby $\phi = \pi$ rad modulo 2π (dwell at about 3π rad annotated 2 in c) are reminiscent of the transient obtained for certain initial conditions in the monostable régime (Fig. 5b, annotation 3). The key point is that in the metastable régime the system's behavior is a blend of coupled and independent behavior

dination law (Eq. (5)) across a range of $\Delta\omega$ values with the coupling parameter, $k = b/a = 1$ fixed. Stable fixed points (attractors) are presented as filled circles and unstable fixed points (repellors) as open circles. Here, fixed points of the coordination dynamics correspond to phase- and frequency relationships between oscillatory processes.

The surface shown in Fig. 4 defines three regions under the influence of the symmetry breaking term $\Delta\omega$. In the first region present in the lower part of the surface, the system is multistable: two stable attracting fixed points (filled circles) represent possible alternative states. Which one the system settles in depends on initial conditions and the size of the basin of attraction. In an intermediate region, following the line labeled 2 from left to right, the weakest attractor near anti-phase (right side) disappears after it collides with its associated repellor somewhere near $\Delta\omega = 1.3$, but the strongest attractor (left side) is still present as well as its repellor partner. Finally in the third region in the upper part of the surface, the dynamics become metastable. Following the line labeled 3 from left to right, no fixed points exist anymore: this part of the surface no longer intersects the isoplane $\dot{\phi} = 0$ where the fixed points are located. Strictly speaking coordination *states* qua frequency- and phase-synchrony no longer exist in the metastable régime of the coordination dynamics. Metastability is thus a subtle blend of coupling and intrinsic tendencies in which behavior is neither completely ordered (synchronized) nor completely disordered (desynchronized). Both tendencies coexist.

How do individual and coordination behavior evolve in time in the metastable régime? A unique flow now exists in which the dynamics may be characterized by places where the trajectory of the coordination variable converges and pauses around the horizontal and places where the trajectory drifts or diverges from the horizontal. Let us define the former as a *dwell time*, and the latter as an *escape time*. In Fig. 5c we show two locations for the dwell times: one that lingers a long time before escaping (e. g. Fig. 5c, annotation 1) slightly above the more stable in-phase pattern near 0 rad (modulo 2π), and the other that lingers only briefly (e. g. Fig. 5c, annotation 2) slightly above π (modulo 2π). These inflections recur over and over again as long as the system self-organizes in the metastable régime, i. e. as long as it does not undergo a phase transition to a locked or unlocked state. Despite the complete absence of phase-locked attractors, the coordinating elements in the metastable régime do not behave totally independently. Rather, their interdependence takes the form of dwellings (phase gathering tendencies) nearby the remnants of the fixed points (cf. Fig. 3 bottom; Figs. 4, 5c) and may be nicely expressed by concentrations in the histogram of the relative phase (see Chap. 4 in [100]).

Recently metastability has been hailed as a "new principle" of coordination in the brain and has been embraced by a number of neuroscientists as playing a role in various cognitive functions, even consciousness itself (e. g. [44,45,57,58,111,127,187,212]). According to a recent review [49]:

> *Metastability is an entirely new conception of brain functioning where the individual parts of the brain exhibit tendencies to function autonomously at the same time as they exhibit tendencies for coordinated activity* [19,20,97,100].

For Coordination Dynamics, metastability's significance lies not in the word itself but in what it means for understanding coordination in living things. In coordination dynamics, as shown in its most elementary form (Eq. (5)), metastability is not a concept or an idea, but a direct result of the broken symmetry of a system of (nonlinearly) coupled (nonlinear) oscillators. Such a design principle for the brain seems highly plausible given that rhythms in the brain are ubiquitous, operate over a broad range of frequencies and are strongly associated with various sensory, motor and cognitive processes [10,24,100].

The Creation of Information

There is another reason for proposing metastable coordination dynamics as the essential way the brain and perhaps all complex organizations work. It concerns an analogy to how physicists understand how we know the universe we live in. According to Quantum Mechanics, out of a universe in which quantum indeterminacy rules – the wave function is spread out over all of space – nature selects an alternative. Information is thereby created. The way this is done in practice is that a device is built in which an interactive material is placed in a physically, electrically or chemically metastable state. According to the late quantum measurement theorist, H.S. Greene [69]:

> *It is the observable transition between this metastable state and a more stable state that conveys the essential information concerning a sub-microscopic event that would otherwise go undetected … The functional material of the detector must be macroscopic and in a metastable state which allows the quantal interaction to become manifest at the macroscopic level.* (see p. 173 in [69])

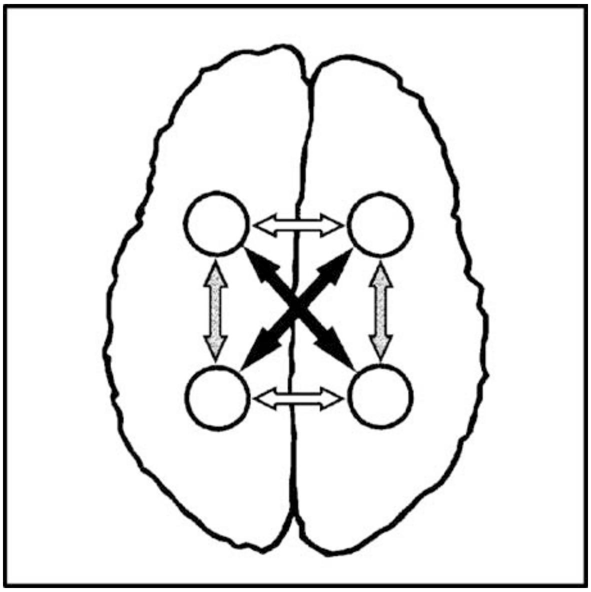
This is how some physicists view the creation of information: bit from it, as it were (in contrast to John Archibald Wheeler's 'it from bit'). Quantum Mechanics thus implies the creation of new information in the process of measurement and observation. Likewise, we have seen in the human brain that information (as a marginally coupled, phase-locked state) is created and destroyed in the metastable régime of the coordination dynamics, where tendencies for apartness and togetherness, individual and collective, segregation and integration, phase synchrony and phase scattering *coexist*. New information is created because the system operates in a special régime where the slightest nudge will put it into a new coordinated state. In this way, the (essentially nonlinear) coordination dynamics creates new, informationally meaningful coordination states that can be stabilized over time. The *stability of information over time* is guaranteed by the coupling between component parts and processes and may constitute a dynamic kind of (non-hereditary) memory. It does not seem a big step then to say that once created, this information can then guide, modify and direct the system's dynamics. As we shall see in Sect. "Modifying Coordination: Meaningful Information" studies of intentional change, environmental change, learning and so forth have demonstrated both empirically and theoretically that an intentional goal – as memorized information – acts in the same information space as the coordination dynamics ([114,173]; see also [141]).

Coordination of Multiple Components: From Quadrupeds to Brains

Phase- and frequency synchronization, the coupling among oscillatory processes, are an example, par excellence, of self-organization in natural systems [76,119,132,161,219]. Think of the famous clocks on the wall, no two exactly the same in frequency, but somehow ticking perfectly in time with each other. The weakest of coupling, whether through vibration in the wall or displacement of the air, enables the clocks to be mutually coordinated without any coordinator at all. Similar phenomena have been seen in the brain and have been hypothesized to play a key role in "binding": oscillations in the brain appear to be coupled or "bound" together into a coherent network when people attend to a stimulus, perceive, think and act (e. g., [34,43,67,184,214]). For example, synchronization in the so-called gamma band (approx. 30–50 Hz) has been proposed as a neural correlate of consciousness. The Journal *Science* announced the synchronization effects observed in monkey cortex as "The Mind Revealed" (see also [34]).

The brain, it seems, has latched on to phase synchrony as a principle of self-organization. Though the connection is seldom made, phase- and frequency- synchronization is typical of central pattern generators (CPGs), neural circuits in vertebrates and invertebrates that generate timing sequences without feedback from the periphery or the help of reflexes (see Sect. "Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions"; [70,71]). Indeed, it is the temporal order observed that allows us to talk about 'pattern generators' in the first place. Though the specific mechanisms are obviously different between the visual cortex of the monkey and the stomatogastric ganglion of the lobster, the dynamic patterns are the same, hinting at the source of an underlying principle [97,100,174]. But what form might the coordination dynamics of the brain take?

Obviously when it comes to the brain there are, in principle, very many regions to coordinate. In practice, however, only a restricted set of regions appear to be functionally connected during particular tasks (see, e. g. Fig. 2). The idea, then, is that one could use the Central Pattern Generator (CPG) design for quadrupedal locomotion [32,66,176] as a basic model of interaction among cortical pattern generators. This is not as far fetched as it seems. It is likelier than not that the precursors to the structure and function of the cerebral cortex are self-contained circuits in the spinal cord and brainstem that generate intrinsic patterns of rhythmic activity [70,71,220]. Such CPGs typically work by transforming tonic driving inputs into detailed spatiotemporal patterns of (usually oscillatory) activity. Several of the properties of CPGs are conserved throughout evolution rendering them a likely candidate for the basic building blocks of the brain [72]. The hypothesis proposed here is that cortical pattern generators may underlie the coordination that is needed for everything the brain is purported to do – think, feel, remember, act, socialize, etc.



Coordination Dynamics, Figure 6
A schematic of brain coordination dynamics among four brain regions. Each *circle* represents an area of the brain capable of intrinsic oscillation and the *arrows* correspond to connections among brain areas giving rise to cortical pattern generation (see text)

Following the footsteps of basic coordination dynamics, in the quadruped analogy each "limb" corresponds to a neural region capable of intrinsic oscillatory activity and the patterns emerge from (broken) symmetries and changes in coupling between neural regions. For example, Fig. 6 shows a cartoon of the neuroanatomical connections underlying the anterior-posterior coordination of the hemispheres of the brain.

The relevant variables are four phase variables, ϕ_{ij} (with $i \in \{\text{right, left hemisphere}\}$, $j \in \{\text{anterior, posterior}\}$ characterizing the oscillatory behavior of each brain area with respect to its timing. Much research on coordination dynamics shows that the *relative phase* is a key coordination variable or order parameter although it is quite possible that amplitudes and frequencies are important variables too [4,57,111]. For the sake of simplicity, we stick to the case of interareal cortical coordination between 4 brain regions, where a set of 3 relative phases suffices to characterize any pattern uniquely.

As already illustrated, a key notion is to use symmetry to classify patterns and restrict the functional form of the coordination dynamics. Here, pattern symmetry means a given cortical pattern is symmetric under a group of transformations. As we have noted, a transformation is an operation that maps one pattern onto another, e. g. the anterior-posterior (a-p) transformation exchanges anterior and posterior regions within a cortical pattern. If all relative phases are equivalent after the transformation, then the pattern is considered invariant under this operation.

Symmetry serves as a pattern classification tool allowing for the identification of basic cortical patterns that can be captured theoretically. Given a symmetry group, one can determine all invariant patterns. For example, certain idealized cortical patterns are invariant under the symmetry group generated by the following operations: exchange of anterior and posterior, exchange of left and right, and inversion of all phases (inversion of time). A good way to illustrate these patterns is with phase pictograms.

Imposing symmetry restrictions on the dynamics itself limits possible solutions and allows one to arrive at a coordination dynamics that contains the patterns as *different* stationary states of the *same* nonlinear dynamical system. In other words basic cortical patterns correspond to attractive states of the relative phase for adequate parameter values:

$$\begin{aligned}\dot{\phi}_{ij} = \sum_{n=1}^{\infty}\{&A_n \sin(n(\phi_{ij} - \phi_{\hat{i}j}))\\ &\text{Homologous contralateral coupling (white arrows)}\\ &+ C_n \sin(n(\phi_{ij} - \phi_{i\hat{j}}))\\ &\text{Ipsilateral coupling (gray arrows)}\\ &+ E_n \sin(n(\phi_{ij} - \phi_{\hat{i}\hat{j}}))\}\\ &\text{Nonhomologous contralateral coupling (black arrows)}\end{aligned} \tag{9}$$

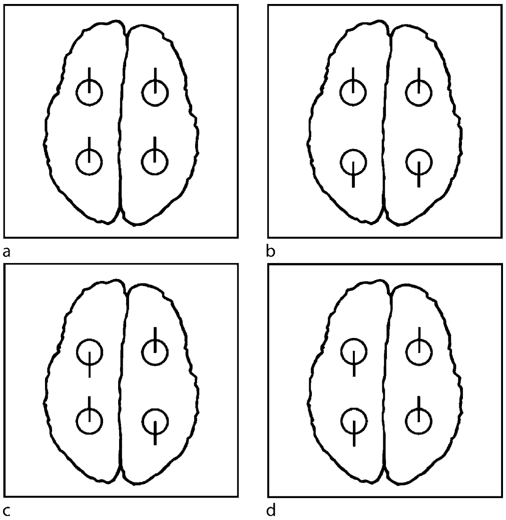
where A_n, C_n and E_n are parameters and a hat over an index means that the opposite value is taken, e. g. if $i = \text{right}$, then $\hat{i} = \text{left}$.

To analyze the solutions to the phase dynamics, for the sake of simplicity the coordination dynamics may be restricted to second order. Higher orders generate parameter régimes where many patterns may coexist; first order removes the possibility that some patterns may coexist. Also, diagonal coupling (black arrows in Fig. 6) may be neglected by setting $E_n = 0$. Notice that an effective diagonal coupling still exists because two couplings are sufficient to stabilize any pattern of activity among 4 cortical areas. Moreover, it is easier to generalize a system with contralateral and ipsilateral couplings to systems in which more areas are involved.

In sum, following exactly the basic theoretical modeling strategy (Eq. (1)) the dynamical system takes the following form:

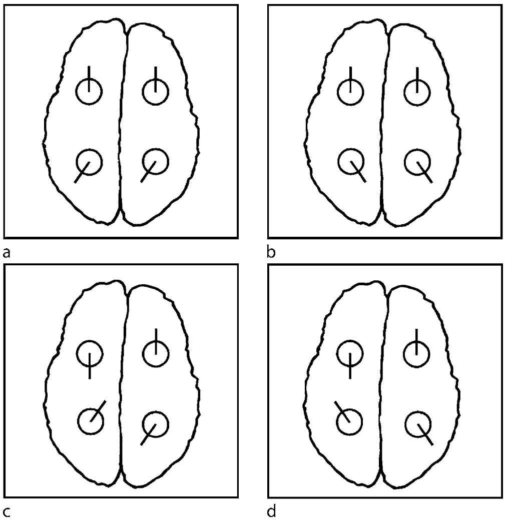
$$\begin{aligned}\dot{\phi}_{\mathrm{rp}} &= A_1[\sin(\phi_{\mathrm{rp}} - \phi_{\mathrm{lp}}) + \sin(\phi_{\mathrm{la}})]\\ &\quad + A_2[\sin(2(\phi_{\mathrm{rp}} - \phi_{\mathrm{lp}})) + \sin(2\phi_{\mathrm{la}})]\\ &\quad + 2C_1 \sin(\phi_{\mathrm{rp}}) + 2C_2 \sin(2\phi_{\mathrm{lp}})\\ \dot{\phi}_{\mathrm{lp}} &= A_1[\sin(\phi_{\mathrm{lp}} - \phi_{\mathrm{rp}}) + \sin(\phi_{\mathrm{la}})]\\ &\quad + A_2[\sin(2(\phi_{\mathrm{lp}} - \phi_{\mathrm{rp}})) + \sin(2\phi_{\mathrm{la}})]\\ &\quad + C_1[\sin(\phi_{\mathrm{lp}} - \phi_{\mathrm{la}}) + \sin(\phi_{\mathrm{rp}})]\\ &\quad + C_2[\sin(2\phi_{\mathrm{lp}} - \phi_{\mathrm{la}})) + \sin(2\phi_{\mathrm{rp}})]\\ \dot{\phi}_{\mathrm{la}} &= 2A_1 \sin(\phi_{\mathrm{la}}) - \phi_{\mathrm{lp}}) + 2A_2 \sin(2\phi_{\mathrm{la}})\\ &\quad + C_1[\sin(\phi_{\mathrm{la}} - \phi_{\mathrm{lp}}) + \sin(\phi_{\mathrm{rp}})]\\ &\quad + C_2[\sin(2(\phi_{\mathrm{la}} - \phi_{\mathrm{lp}})) + \sin(2\phi_{\mathrm{rp}})]\ .\end{aligned} \tag{10}$$

Solving $\dot{\phi}_{ij} = 0$ yields stationary solutions that correspond to idealized cortical "gaits". Trot, pace, gallop and jump patterns may be identified as multistable or monostable solutions in various parameter régimes. Patterns of lower symmetry can also be captured. Obviously the foregoing analysis is intended for illustrative purposes only. The examples provided in Fig. 7 and 8 are only a few of very many possible cortical patterns that can be obtained by further symmetry groups. It is important to empha-



Coordination Dynamics, Figure 7
Brain phase pictograms. Each brain area is represented as a *circle*. The spatial arrangement of the circles viewed from looking down on the top of the head represents the brain's hypothesized anterior-posterior and left-right functional organization. Phase is represented by the angle the stick makes on each circle, with the reference phase being zero for the right frontal region. If all phases are rotated by the same amount in the same direction, the cortical coordination pattern remains the same. The patterns in a, b, c and d are those idealized cortical patterns that remain invariant under anterior-posterior, left-right and time inversion operations. Notice that the relative phase between any two brain regions is either inphase or antiphase

size that not all patterns in a given symmetry group are observable. Which ones are actually observed is dictated by the coordination dynamics, in particular a given pattern's *stability* which, as we have seen, can be measured (cf. Sect. "Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions"). As in the simpler cases of coordination dynamics described above, *switching* among cortical patterns is due, not to switches per se but to instabilities – phase transitions or bifurcations in the phase dynamics. Neuromodulators are candidate control parameters capable of sculpting cortical patterns by leading the system through instabilities [130]. Moreover, when the oscillatory frequencies in the anterior and posterior regions of the brain are slightly different, a kind of partial coordination among cortical regions may occur. This is exactly the *metastable* coordination dynamics of the brain as described in the previous section. Here again, the key point is that the rules of the game appear to be run by principles of coordination dynamics and symmetry. As always, experiments are now needed to test this hypothesis. EEG measures of cross-frequency phase synchrony of the human brain may reflect a start in this direction (e. g., [84,158]). More direct attempts are underway in our laboratory [199].



Coordination Dynamics, Figure 8
Representative brain phase pictograms corresponding to cortical patterns of lower symmetry. Here the anterior-posterior symmetry is dropped and the cortical patterns that remain form two one parameter families. One family consists of in-phase ordering *within* anterior and posterior areas and any fixed phase relation *between* anterior and posterior regions (a,b). The other family (c,d) consists of anti-phase relations within frontal and anterior regions and any fixed phase relations between them

"Collective Minds"

The basic coordination dynamics for two and four nonlinearly interacting components (Eqs. (2)–(10)) can readily be elaborated as a model of emergent coordinated behavior or "group cohesion" among very many anatomically different components (see, e. g. [5]). Self-organized behavioral patterns such as singing in a group or making a "wave" during a football game are common, yet unstudied examples. By virtue of information exchange nearest neighbors adjust their motions to each other generating, and being influenced by, their social environment. Recently, Néda and colleagues [148,149] have examined a simpler group activity: applause in theater and opera audiences in Romania and Hungary. After an exceptional performance, initially thunderous incoherent clapping gives way to slower, synchronized clapping. Measurements indicate that the clapping period suddenly doubles at the onset of the synchronized phase, and slowly decreases as synchronization is lost. This pattern is a cultural phenomenon in many parts of Europe: a collective request for an encore. Increasing frequency (decreasing period) is a measure of the urgency of the message, and culminates in the transition back to noise when the performers reappear. These results are readily explained by a model of a group of globally coupled nonlinear oscillators [132]:

$$\frac{\mathrm{d}\phi_k}{\mathrm{d}t} = \omega_k + \frac{K}{N}\sum_{j=1}^{N}\sin(\phi_j - \phi_k) \tag{11}$$

in which a critical coupling parameter, K_c determines the different modes of clapping behavior. K is a function of the dispersion (D) of clapping frequencies:

$$K_c = \sqrt{\frac{2}{\pi^3}}\mathrm{D}\,. \tag{12}$$

During fast clapping, synchronization is not possible due to the large dispersion of clapping frequencies. Slower, synchronized clapping at double the period arises when small dispersion appears. Period doubling rhythmic applause tends not to occur in big open-air concerts where the informational coupling among the audience is small. K can also be societally imposed. In Eastern European communities during communist times, synchronization was seldom destroyed because enthusiasm was often low for the "great leader's" speech. For people in the West, the cultural information content of different clapping patterns may be quite different. Regardless, the mathematical descriptions for coordinated behavior – of social dyads and the psychology of large groups – are remarkably similar.

Modifying Coordination: Meaningful Information

Unlike the behavior of inanimate things, the self-organizing dynamics of animate behavior is based on information (Table 1), though not in the standard sense of data communicated across a channel [180]. In coordination dynamics, *collective or coordination variables are context-dependent and intrinsically meaningful*. Context-dependence does not imply subjectivity and lack of reproducibility. Nor does it mean that every new context requires a new collective variable or order parameter. As we have seen already, for example, within- and between-person coordinated behaviors are described by the same self-organizing coordination dynamics. One of the consequences of identifying the latter is that in order to *modify or change* the system's behavior, new information (in the form say, of an environmental input, a task to be learned, or an intention to change behavior) is expressed in terms of parameters acting on system-relevant collective dynamics. On the one hand, the benefit of identifying collective variables is that they embrace the full complexity of the system and hence provide the relevant information about *what* to modify. On the other, the collective variable dynamics – prior to the introduction of any new information – influences how that information can be used. The upshot is that information is not lying out there as mere data: information is meaningful to the extent that it modifies, and is modified by, the collective variable dynamics.

A minimum mathematical form for the full coordination dynamics which encompasses both spontaneous self-organizing tendencies and specific parametric influences is

$$\dot{\phi} = f(\phi) + f\,\mathrm{inf}(\phi) \qquad (13)$$

where the first term is the typical so-called "intrinsic dynamics" e. g., of Eq. (5) or Eq. (10) and the second term represents 'informational forcing', i. e., a perturbation of the vector field of the dynamics attracting the system toward a required coordination pattern. It is important to emphasize that the plus sign in Eq. (13) is for operational purposes only, affording the measurement of the complementary contributions to the coordination dynamics of both spontaneous and directed (parametric) influences. The conceptual advantage of Eq. (13) is that information acts in the same space as the collective variables that define the intrinsic coordination patterns, i. e., those patterns that characterize spontaneous coordination tendencies. Thus, information is not arbitrary with respect to the dynamics. A corollary of this formulation is that information has no meaning outside its influence on the intrinsic dynamics. They are cut, as Sheets-Johnstone [182] remarks, from the same dynamic cloth.

Intentional Dynamics

Self-organizing processes, in the manner of Haken's synergetics, provide a theoretical foundation for all forms of coordination. However, we do not want to throw the baby out with the bathwater. Coordinated behavior often has a goal-directedness to it as well. We humans, for example, have no doubt whatsoever that it is us, and us alone, that direct the motions of our own bodies. Where do agency and directedness come from? A clue comes from considering the elementary spontaneous movements we are born with which consist of a large repertoire of spontaneous (thus self-organized) movements – making a fist, kicking, sucking, etc. etc. Only at some point does the child realize – through his own movements and the kinaesthetic sensations they give rise to – that these movements are his own. If one attaches the string of a mobile to his foot, he comes to realize that it is *his* kicking movements that are causing the mobile to move in ways that *he* likes. The pre-existing repertoire enables activities to happen before we make them happen. Evolutionarily constrained self-organizing coordination tendencies ('intrinsic dynamics') thus appear to lie at the origins of conscious agency. They are, in the words of the philosopher Maxine Sheets-Johnstone, "the mother of all cognition", presaging every conscious mind that ever said "I". From spontaneous self-organized behavior emerges the self – "I am" "I do" and from there a huge range of potentialities ('I can do'). "I-ness" arises from spontaneity, and it is this "I" that directs human action. As Sheets-Johnstone [182] cogently remarks, we literally discover ourselves in movement. In our spontaneity of movement, we discover arms that extend, mouths that open, knees that flex and so forth. We make sense of ourselves as living things.

Following these insights, consider briefly how Coordination Dynamics addresses the role of intentional information in bringing about behavioral change. How is the process of intentionally switching among patterns of coordination to be understood? According to the theory, the relative stability of the intrinsic patterns plays a role in determining how easily the system can switch in and out of coordination states. As defined previously, "intrinsic dynamics" expresses the fact that the system (which may include the brain) – prior to any specific input – already possesses a repertoire of behavioral patterns that are unique to each individual. Theoretically, intention parametrizes the intrinsic dynamics in two ways: (i) by destabilizing an ongoing pattern and stabilizing a target pattern [114,137,171]; and (ii) by stabilizing an intrinsically unstable pattern that under the current circumstances might otherwise become unstable and switch (see

Chap. 5 in [100], [114]). Measurement of switching time shows that intention both acts upon and is constrained by the intrinsic dynamics of coordination. First, the system switches far faster from less stable to more stable patterns (as measured by variability) than vice-versa. Second, the data show that it is possible to intentionally stabilize intrinsically unstable patterns under conditions in which they would normally switch. Both results are in excellent agreement with theory [114,171,173].

The neural basis of the interaction between intrinsic brain dynamic activity and *intentional* pattern selection and switching is just beginning to be studied and the results look very promising [39]. *Spontaneous* switching between patterns is known to be associated with increased activity in prefrontal, premotor and parietal regions [3,37,146], a network that is compatible with the stability dependent circuits described in Fig. 2 (see also Jantzen & Kelso, 2007). Increased activity reported in specific brain regions appears to reflect the loss of pattern stability that precedes spontaneous pattern switching. New results from our laboratory show that there is greater activity in the basal ganglia (BG) – a region known to be crucial for starting and controlling voluntary movements [68] – when moving from a more to a less stable pattern [39]. The heightened level of activity in BG may be related to the stability of the original pattern, the stabilization of the selected pattern switched into or both. Regardless, this intriguing result suggests that the basal ganglia play a key role in parametrizing the coordination dynamics.

Stimulus (Parametric) Stabilization and Change

Not only internally generated information is able to stabilize and destabilize coordination states under suitable circumstances: coupling sound, vision or touch conditions to specific aspects of an individual movement have been shown, not only to modify the movement but to *globally* stabilize coordination [25,50,122,133]. Thus the role of 'stimuli' in Coordination Dynamics is much more than to trigger preset motor commands or provide feedback to the motor system. To account for these kinds of effects, Jirsa and colleagues [89] introduced the notion of *parametric stabilization*: coupling specific sensory input parametrically to a set of limit cycle oscillators (see also [4,91]):

$$\begin{aligned}\ddot{x}_1 + f(x_1, \dot{x}_1)\dot{x}_1 + \omega^2 x_1 &= g(x_1, \dot{x}_1, x_2, \dot{x}_2) + \varepsilon(t)x_1\\ \ddot{x}_2 + f(x_2, \dot{x}_2)\dot{x}_2 + \omega^2 x_2 &= g(x_2, \dot{x}_2, x_1, \dot{x}_1) + \varepsilon(t)x_2\end{aligned} \tag{14}$$

where *f* is a nonlinear oscillator function, *g* represents the HKB coupling (Eq. (4)), ω is the eigenfrequency of the oscillator and $\varepsilon(t)$ represents sensory information. Here again in Eq. (14) we see a key aspect of coordination dynamics, namely that perception and action, sensory information and the dynamics of movement are inextricably linked. Notice the linkage in this case is of a parametric, multiplicative nature which is necessary to account for both the local changes to component trajectories produced by sensory information (called 'anchoring') and the global stabilization effects on the coordination dynamics. Fink et al. [50] for example, were able to show that such localized and specific sensory information was capable of shifting (and thereby delaying) the critical point at which phase transitions occurred.

A Brief Survey of Applications and Elaborations of Coordination Dynamics

The foregoing discussion pertains to just two of the many kinds of adaptive modification of coordination dynamics that have been investigated in the literature. Here only a flavor can be provided. The sample includes, but is by no means limited to: the processes underlying the ability of biological systems to stabilize intrinsically unstable systems [26,54,199]; the initiation (including 'false starts') and coordination of discrete, discontinuous behaviors [52,80,86,112,188,189] including neurally-based comparisons with those of a continuous, rhythmic nature [166,185]; the spontaneous recruitment and annihilation of biomechanical degrees of freedom to accomplish task and environmental conditions [21,23,51,118]; the coordination dynamics of trajectory formation [22,38] and cursive handwriting [6]; the important role that perception [145] and attention [2,27,147,190,193,194] play in modulating coordinative stability; how practice and learning alter the entire coordination repertoire by reshaping the landscape of the coordination dynamics using competitive and cooperative mechanisms [53,140,151, 177,222,223]; the stabilization and consolidation of new memorized states of coordination and the dynamics of the forgetting process [128,129]; how handedness amplifies asymmetries in the coordination dynamics [2,201], and so forth. The same concepts and methods have been applied to problems ranging from maintaining posture and stabilizing postural sway [8,9,42,83] to understanding how concurrent cognitive tasks modulate coordination dynamics [159,183].

Theoretical research at the neural level has progressed from phenomenological modeling at behavioral (e. g. [60,76,89,115,118,175,201]) and brain levels [88,205] to neurobiologically-grounded accounts of both unimanual [56,63,125] and bimanual coordination [87] that are based on known cellular and neural ensemble properties of the cerebral cortex. Recent work [85] has extended this neural theory to include the heterogeneous connectivity between neural ensembles in the cortex. Once general laws of coordination at behavioral and brain levels have been identified, it has proved possible to derive them from a deeper theory founded on neuroanatomical and neurophysiological facts, thereby causally connecting different levels of description [117] for review). The neural theory, in turn, poses a number of challenges to experiment, such as how synaptic and cellular properties are influenced by learning, arousal and attention [103].

Remarkable applications of coordination dynamics have occurred in expected directions (though none the less remarkable for all that) including many physical activities and sports such as the relation between respiration and locomotion [36], juggling [13,79], gymnastics [142], running [41], tennis [157], swimming [179], boxing [135], skiing [154], golf [125] and even riding horses [134] to name only a few, as well as in entirely unexpected directions, such as modeling coordination of infant breathing as a way to understand the effects of premature birth [65], studies of coordination dynamics in children with Developmental Coordination Disorder [213] and the introduction of coordination dynamics therapy to treat a wide variety of CNS disorders and diseases (e. g. [167,168,169]; see also [207]). Principles of coordination dynamics have been shown to apply to perceptual grouping as nicely illustrated by the classic bistable properties of reversible figures such as the Necker cube (e. g. [100]), pattern recognition [77], the visual perception of spatiotemporal inphase and antiphase moving stimuli [18,40,78,221] and speech categorization [28,202]. In many cases the foregoing research findings have expanded, if not overturned, conventional explanations of phenomena that have seldom considered dynamics.

Increase in research activity using the concepts and methods of coordination dynamics has been such that the term has taken on a life of its own in different fields. Thus, it is commonplace in the literature to hear the words 'cognitive', 'brain', 'neural', 'social', 'behavioral', 'developmental' 'multimodal', 'postural', etc., qualify and precede the words *coordination dynamics*. The dynamical approach is currently center stage in a number of fields, for example, dynamical neuroscience (e. g., [90,160]), dynamical cognitive science (e. g., [16,162,186,209]), behavioral [217] and task [165] dynamics, dynamical social psychology (e. g., [15,155,206]), dynamical systems accounts of development (e. g., [135,149,191,192,204]; see also [195]) and its implications are under careful consideration in philosophical circles (e. g. [11,29,30,31,198]).

Future Directions and Conclusions: The Complementary Nature of Coordination Dynamics

General laws and principles of biological coordination – to the extent they exist – are, by definition, abstract and mathematical. Yet, these laws are always conditioned by and realized by specific mechanisms and contexts. Over the last twenty-five years, often using the field of animate movement as an entry point it has been shown that the same coordination dynamics applies to functional coordination in a wide variety of situations. Although the basic laws for a quantitative description of the phenomena observed when human beings (and human brains) move, interact with the environment and with each other are the same, the anatomical, mechanical and physiological mechanisms realizing these dynamics are obviously not. Laws and mechanisms are complementary aspects of coordination dynamics.

Current research and theory views coordination as arising from the mutual interplay of constraints on multiple levels of description – ranging from the intrinsic properties and modes of interaction among cells and cellular ensembles in brain circuitry to biomechanical influences at the behavioral level all the way to cognitive and task constraints. Coordination dynamics is not only a theoretical framework, but also a research program that explicitly attempts to incorporate and connect known constraints at multiple levels of description. For instance, coordination dynamics successfully identified and later quantified the form of the nonlinear coupling among interacting components. In showing that the stability and change of coordination is due to nonlinear interactions among individual components coordination dynamics removes some of the mysticism behind the contemporary terms "emergence" and "self-organization". At the same time, coordination dynamics expands and modifies the concept of self-organization in non-living systems by introducing new concepts to account for the fact that coordination is not only characterized by self-organization but also by directed or supervised forms of coordination. The two origins or cornerstones of coordination dynamics may be reconciled by showing how meaningful information originates from self-organizing processes and may in turn modify them.

In studies of coordinated movement, the field to which coordination dynamics owes its origins, it has proven useful to try to isolate the role of various constraints and how they are mediated by the central nervous system. On the one hand, this strategy has helped identify different factors that serve to stabilize coordination under conditions in which it may otherwise become unstable and susceptible to change. On the other hand, a focus on isolating particular constraints can lead, albeit unwittingly, to dichotomies (e. g., coordination principles versus neuromuscular-skeletal mechanisms of implementation) that may not be so useful. In reality it seems safe to conclude that a coalition of constraints – acting on multiple levels – impinges upon the stability of coordination depending on task and environmental context and the mover's intent. For example, the multilevel theory offered by Kelso [100] connects task goals (Level 1) to constraints on nonlinear oscillators (Level 3), the interactions among which determine the coordinative patterns observed (Level 2). Thus, rather than pose "abstract laws of coordination dynamics" against "neuromuscular-skeletal determinants of coordination", more important is to understand how the balance between identified constraints plays itself out in the course of any coordinated activity. Situations in which constraints are placed in competition with each other often prove to be highly revealing [122].

Throughout this article, every effort has been made to articulate the key notions of coordination dynamics, both conceptual and technical, and to present them in close proximity in order to help both the novice and the expert reader. The behavioral simplicity of the basic coordination patterns studied in the laboratory is deceptive; their understanding, however, requires recent advances in physics and mathematics. The theoretical concepts and methods of coordination dynamics are likely to play an ever greater role in the social, behavioral, economic, cognitive and neurosciences, especially as the interactions among disciplines continues to grow. Up to now, the use of nonlinear dynamics is still quite restricted, and often metaphorical. One reason is that the tools are difficult to learn, and require a degree of mathematical sophistication. Their implementation in real systems is nontrivial, requiring a different approach to experimentation and observation. Another reason is that the dynamical perspective is often cast in opposition to more conventional theoretical approaches, instead of as an aid or complement to understanding. The former tends to emphasize decentralization, collective decision-making and cooperative behavior among many interacting elements. The latter tends to focus on individual psychological processes such as intention, perception, attention, memory and so forth. Yet there is increasing evidence that intending, perceiving, attending, deciding, emoting and remembering have a dynamics as well. The language of dynamics serves to bridge individual and group processes. In each case, dynamics must be filled with content, with key variables and parameters obtained for the systems under study. A beauty about coordination dynamics is that the coordination variables or order parameters are semantic, relational quantities that "enfold" different aspects together thereby reducing often arbitrary divisions. Every system is different, but what we learn about one may aid in understanding another. What may be most important of all is to see animated living things in the light of a theory – coordination dynamics – that embraces both spontaneous self-organizing and directed processes, the complementary nature.

Acknowledgments

The writing of this article was supported by NIMH Innovations Award MH 42900, NINDS Grant NS48220, the US Office of Naval Research and the Pierre de Fermet chair of the Midi-Pyrénées Région for which the author is grateful.

Bibliography

Primary Literature

1. Abbott LF (2006) Where are the switches on this thing? In: van Hemmen JL, Sejnowski TJ (eds) 23 Problems in Systems Neuroscience. Oxford University Press, Oxford, pp 423–431
2. Amazeen EL, Amazeen PG, Treffner PJ, Turvey MT (1997) Attention and handedness in bimanual coordination dynamics. J Exp Psychol Hum Perc Perf 23:1552–1560
3. Aramaki Y, Honda M, Okada T, Sadato N (2006) Neural correlates of the spontaneous phase transition during bimanual coordination. Cereb Cortex 16:1338–1348
4. Assisi CG, Jirsa VK, Kelso JAS (2005) Dynamics of multifrequency coordination using parametric driving: Theory and Experiment. Biol Cybern 93:6–21
5. Assisi CG, Jirsa VK, Kelso JAS (2005) Synchrony and clustering in heterogeneous networks with global coupling and parameter dispersion. Phys Rev Lett 94:018106
6. Athènes S, Sallagoïty I, Zanone PG, Albaret JM (2004) Evaluating the coordination dynamics of handwriting. Hum Mov Sci 23:621–641
7. Banerjee A (2007) Neural information processing underlying rhythmic bimanual coordination: theory, method and experiment. Ph D Thesis, Complex Systems and Brain Sciences, Florida Atlantic University
8. Bardy BG (2004) Postural coordination dynamics in standing humans. In: Jirsa VK, Kelso JAS (eds) Coordination Dynamics: Issues and Trends. Springer, Berlin
9. Bardy BG, Oullier O, Bootsma RJ, Stoffregen TA (2002) The dynamics of human postural transitions. J Exp Psychol Hum Perc Perf 28:499–514
10. Başar E (2004) Memory and Brain Dynamics: Oscillations Integrating Attention, Perception, Learning, and Memory. CRC Press, Boca Raton
11. Bechtel W (1998) Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science. Cogn Sci 22:295–318
12. Beek PJ, Beek WJ (1988) Tools for constructing dynamical models of rhythmic movement. Hum Mov Sci 7:301–342
13. Beek PJ, Turvey MT (1992) Temporal Patterning in Cascade Juggling. J Exp Psychol Hum Perc Perf 18:934–947

14. Beek PJ, Rikkert WEI, van Wieringen PCW (1996) Limit cycle properties of rhythmic forearm movements. J Exp Psychol Hum Perc Perf 22:1077–1093
15. Beek PJ, Verschoor F, Kelso JAS (1997) Requirements for the emergence of a dynamical social psychology. Psychol Inq 8:100–104
16. Beer RD (1999) Dynamical approaches to cognitive science. Trends Cogn Sci 4:91–99
17. Bernstein N (1967) The coordination and regulation of movement. Pergammon, Oxford
18. Bingham GP, Schmidt RC, Zaal FT (1999) Visual perception of the relative phasing of human limb movements. Perc Psychophys 61:246–258
19. Bressler SL (2003) Cortical coordination dynamics and the disorganization syndrome I schizophrenia. Neuropsychopharmacology 28:S35-S39
20. Bressler SL, Kelso JAS (2001) Cortical coordination dynamics and cognition. Trends Cogn Sci 5:26–36
21. Buchanan JJ, Kelso JAS (1999) To switch or not to switch: Recruitment of degrees of freedom stabilizes biological coordination. J Motor Behav 31:126–144
22. Buchanan JJ, Kelso JAS, de Guzman GC (1997) The self-organization of trajectory formation: I Experimental evidence. Biol Cybern 76:257–273
23. Buchanan JJ, Kelso JAS, de Guzman GC, Ding M (1997) The spontaneous recruitment and annihilation of degrees of freedom in rhythmic hand movements. Hum Mov Sci 16:1–32
24. Buzsáki G (2006) Rhythms of the Brain. Oxford University Press, Oxford
25. Byblow WD, Carson RG, Goodman D (1994) Expressions of asymmetries and anchoring in bimanual coordination. Hum Mov Sci 13:3–28
26. Cabrera JL, Milton JG (2004) Stick balancing: On-off intermittency and survival times. Nonlinear Sci 11:305–317
27. Carson RG, Chua R, Byblow WD, Poon P, Smethurst CS (1999) Changes in posture alter the attentional demands of voluntary movement. Proc R Soc Lond B 266:853–857
28. Case P, Tuller B, Ding M, Kelso JAS (1995) Evaluation of a dynamical model of speech perception. Perc Psychophys 57:977–988
29. Chemero A (2001) Dynamical explanation and mental representation. Trends Cogn Sci 5:140–141
30. Chemero A, Silberstein M (2008) After the philosophy of mind. Philos Sci (in press)
31. Clark A (1997) Being there. MIT Press, Cambridge
32. Collins JJ, Stewart IN (1993) Coupled nonlinear oscillators and the symmetries of animal gaits. J Nonlinear Sci 3:349–392
33. Crick FHC (1966) Of molecules and men. University of Washington Press, Seattle
34. Crick FHC (1994) The astonishing hypothesis. Scribner, New York
35. Daffertshofer A, Peper CE, Beek PJ (2000) Spectral analysis of event-related encephalographic signals. Phys Lett A 266:290–302
36. Daffertshofer A, Huys R, Beek PJ (2004) Dynamical coupling between locomotion and respiration. Biol Cybern 90: 157–164
37. Debaere F, Swinnen SP, Beatse E, Sunaert S, van HP, Duysens J (2001) Brain areas involved in interlimb coordination: a distributed network. Neuroimage 14:947–958
38. DeGuzman GC, Kelso JAS, Buchanan JJ (1997) The self-organization of trajectory formation: II Theoretical model. Biol Cybern 76:275–284
39. DeLuca C, Jantzen KJ, Bertollo M, Comani S, Kelso JAS (2008) The role of basal ganglia in the intentional switching between coordination patterns of different stability. Paper presented at 18th Annual Meeting of the Neural Control of Movement, Naples, Florida 29 April-4 May 2008
40. De Rugy RA, Oullier O, Temprado JJ (2008) Stability of rhythmic visuo-motor tracking does not depend on relative velocity. Exp Brain Res 184:269–273
41. Diedrich FJ, Warren WH Jr (1995) Why change gaits? Dynamics of the walk-run transition. J Exp Psychol Hum Perc Perf 21:183–202
42. Dijkstra TMH, Schoner G, Giese MA, Gielen CCAM (1994) Frequency dependence of the action-perception cycle for postural control in a moving visual environment: Relative phase dynamics. Biol Cybern 71:489–501
43. Eckhorn R, Bauer R, Jordan W, Borsch M, Kruse W, Munk M, Reitboeck HJ (1988) Coherent oscillations: a mechanism of feature linking in the visual cortex. Multiple electrode correlation analyses in the cat. Biol Cybern 60:121–130
44. Edelman GM (2004) Naturalizing consciousness: a theoretical framework. Proc Nat Acad Sci USA 100:520–524
45. Edelman G, Tononi G (2000) A Universe of Consciousness. Basic Books, New York
46. Edelman GE, Gally JA (2001) Degeneracy and complexity in biological systems. Proc Nat Acad Sci USA 98:13763–13768
47. Eisenhammer T, Hubler A, Packard N, Kelso JAS (1991) Modeling experimental time series with ordinary differential equations. Biol Cybern 65(2):107–112
48. Fearing F (1930/1970) Reflex Action: a study in the history of physiological psychology. MIT Press, Cambridge
49. Fingelkurts AnA, Fingelkurts AlA (2004) Making complexity simpler: multivariability and metastability in the brain. Int J Neurosci 114:843–862
50. Fink PW, Kelso JAS, Jirsa VK, Foo P (2000) Local and global stabilization of coordination by sensory information. Exp Brain Res 134:9–20
51. Fink PW, Kelso JAS, DeGuzman GC (2000) Recruitment of degrees of freedom stabilizes coordination. J Exp Psychol Hum Perc Perf 26:671–692
52. Fink PW, Kelso JAS, Jirsa VK (in press) Perturbation-induced false starts as a test of the Jirsa–Kelso Excitator Model. J Motor Behav (in press)
53. Fontaine RB, Lee TD, Swinnen SP (1997) Learning a new bimanual coordination pattern: Reciprocal influences of intrinsic and to-be-learned patterns. Can J Exp Psychol 51(1):1–9
54. Foo P, Kelso JAS, deGuzman GC (2000) Functional stabilization of unstable fixed points: Human pole balancing using time-to-balance information. J Exp Psychol Hum Perc Perf 26:1281–1297
55. Fox Keller E (2007) A clash of two cultures. Nature 445:603
56. Frank TD, Daffertshofer A, Peper CE, Beek PJ, Haken H (2000) Towards a comprehensive theory of brain activity: Coupled oscillator systems under external forces. Physica D 144:62–86
57. Freeman WJ, Holmes MD (2005) Metastability, instability, and state transition in neocortex. Neural Netw 18:497–504
58. Friston KJ (1997) Transients, metastability, and neuronal dynamics. Neuroimage 5:164–171
59. Fuchs A, Jirsa VK (eds) (2008) Coordination: Neural, behavioral and social dynamics. Springer, Heidelberg

60. Fuchs A, Kelso JAS (1994) A theoretical note on models of interlimb coordination. J Exp Psychol Hum Perc Perf 20:1088–1097
61. Fuchs A, Kelso JAS, Haken H (1992) Phase transitions in the human brain: Spatial mode dynamics. Int J Bifurc Chaos 2:917–939
62. Fuchs A, Deecke L, Kelso JAS (2000) Phase transitions in human brain revealed by large SQuID arrays: Response to Daffertshofer, Peper and Beek. Phys Lett A 266:303–308
63. Fuchs A, Jirsa VK, Kelso JAS (2000) Theory of the relation between human brain activity (MEG) and hand movements. NeuroImage 11:359–369
64. Gleick J (2003) Isaac Newton. Pantheon, New York
65. Goldfield EC, Schmidt RC, Fitzpatrick P (1999) Coordination dynamics of abdomen and chest during infant breathing: A comparison of full-term and preterm infants at 38 weeks postconceptional age. Ecol Psychol 11:209–233
66. Golubitsky M, Stewart I, Buono P-L, Collins JJ (1999) Symmetry in locomotor central pattern generators and animal gaits. Nature 401:693–695
67. Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchroniation which reflects global stimulus properties. Nature 338:334–337
68. Graybiel AM, Saka E (2003) The basal ganglia and the control of action. In: Gazzaniga MS (ed) The New Cognitive Neurosciences, 3rd edn. MIT Press, Cambridge, pp 495–510
69. Greene HS (2000) Information Theory and Quantum Physics. Springer, Berlin
70. Grillner S (1985) Neurobiological bases of rhythmic motor acts in vertebrates. Science 228:143–149
71. Grillner S (2003) The motor infrastructure: from ion channels to neuronal networks. Nat Rev Neurosci 4:573–586
72. Grillner S (2008) Personal communication
73. Haken H (1977/1983) Synergetics: Nonequilibrium phase transitions and self-organization in physics, chemistry and biology. Springer, Berlin
74. Haken H (1988) Information and self-organization. Springer, Berlin
75. Haken H (1996) Principles of Brain Functioning. Springer, Berlin
76. Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transitions in human hand movements. Biol Cybern 51:347–356
77. Haken H, Kelso JAS, Fuchs A, Pandya A (1990) Dynamic pattern recognition of coordinated biological motion. Neural Netw 3:395–401
78. Hock HS, Kelso JAS, Schöner G (1993) Bistability, hysteresis, and phase transitions in the perceptual organization of apparent motion. J Exp Psychol Hum Perc Perf 19:63–80
79. Huys R, Daffertshofer A, Beek PJ (2004) Multiple time scales and subsystem embedding in the learning of juggling. Hum Mov Sci 23:315–336
80. Huys R, Studenka BE, Rheaume NL, Zelaznik HN, Jirsa VK (in press) Distinct timing mechanisms produce discrete and continuous movements. Public Library of Science (in press)
81. Jantzen KJ, Kelso JAS (2007) Neural coordination dynamics of human sensorimotor behavior: A Review. In: Jirsa VK, McIntosh R (eds) Handbook of Brain Connectivity. Springer, Heidelberg, pp 421–461
82. Jensen O, Colgin LL (2007) Cross-frequency coupling between neuronal oscillations. Trends Cogn Sci 11:267–269
83. Jeka JJ, Schoner G, Dijkstra TMH, Ribeiro P, Lackner JR (1997) Coupling of fingertip somatosensory information to head and body sway. Exp Brain Res 113:475–483
84. Jirsa VK, Haken H (1997) A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics. Physica D 99:503–526
85. Jirsa VK, Kelso JAS (2000) Spatiotemporal pattern formation in neural systems with heterogeneous connection topologies. Phys Rev E 62:8462–8465
86. Jirsa VK, Kelso JAS (2005) The excitator as a minimal model for the coordination dynamics of discrete and rhythmic movements. J Motor Behav 37:35–51
87. Jirsa VK, Fuchs A, Kelso JAS (1998) Connecting cortical and behavioral dynamics: bimanual coordination. Neural Comput 10:2019–2045
88. Jirsa VK, Friedrich R, Haken H, Kelso JAS (1994) A theoretical model of phase transitions in the human brain. Biol Cybern 71:27–35
89. Jirsa VK, Fink PW, Foo P, Kelso JAS (2000) Parametric stabilization of biological coordination: A theoretical model. J Biol Phys 26:85–112
90. Jirsa VK, McIntosch AR (eds) (2007) Handbook of brain connectivity. Springer, Heidelberg
91. Kay BA, Warren WH Jr (2001) Coupling of posture and gait: mode locking and parametric excitation. Biol Cybern 85:89–106
92. Kay BA, Kelso JAS, Saltzman EL, Schöner G (1987) The space-time behavior of single and bimanual rhythmical movements: Data and a limit cycle model. J Exp Psychol Hum Perc Perf 13:178–192
93. Kay BA, Saltzman EL, Kelso JAS (1991) Steady-state and perturbed rhythmical movements: Dynamical modeling using a variety of analytic tools. J Exp Psychol Hum Perc Perf 17:183–197
94. Kelso JAS (1981) On the oscillatory basis of movement. Bull Psychon Soc 18:63
95. Kelso JAS (1984) Phase transitions and critical behavior in human bimanual coordination. Am J Physiol Regul Integr Comp 15:R1000-R1004
96. Kelso JAS (1990) Phase transitions: Foundations of behavior. In: Haken H, Stadler M (eds) Synergetics of cognition. Springer, Berlin, pp 249–268
97. Kelso JAS (1991) Behavioral and neural pattern generation: The concept of Neurobehavioral Dynamical System (NBDS). In: Koepchen HP, Huopaniemi T (eds) Cardiorespiratory and Motor Coordination. Springer, Berlin
98. Kelso JAS (1994) Elementary coordination dynamics. In: Swinnen S, Heuer H, Massion J, Casaer P (eds) Interlimb Coordination: Neural Dynamical and Cognitive Constraints, pp 301–318. Academic Press, San Diego
99. Kelso JAS (1994) The informational character of self-organized coordination dynamics. Hum Mov Sci 13:393–413
100. Kelso JAS (1995) Dynamic Patterns: The Self-organization of Brain and Behavior. MIT Press, Cambridge. [Paperback edition 1997:4th Printing]
101. Kelso JAS (2000) Fluctuations in the coordination dynamics of brain and behavior. In: Arhem P, Blomberg C, Liljenstrom H (eds) Disorder versus order in brain function: Essays in Theoretical Biology. World Scientific, Singapore

102. Kelso JAS (2000) Principles of dynamic pattern formation and change for a science of human behavior. In: Bergman LR, Cairns RB, Nilsson L-G, Nystedt L (eds) Developmental science and the holistic approach. Erlbaum, Mahwah, pp 63–83
103. Kelso JAS (2000) The self-organized dynamics of human skill learning. Dynamical Neuroscience VIII, New Orleans (available from NIMH, Bethesda, Maryland)
104. Kelso JAS (2002) The complementary nature of coordination dynamics: Self-organization and the origins of agency. J Nonlinear Phenom Complex Syst 5:364–371
105. Kelso JAS (2007) Synergies. Scholarpedia (Computational Neuroscience/Dynamical Systems)
106. Kelso JAS (2007) The Haken–Kelso–Bunz Model. Scholarpedia (Computational Neuroscience/Dynamical Systems)
107. Kelso JAS (2008) Synergies: Atoms of brain and behavior. In: Sternad D (ed) A multidisciplinary approach to motor control. Springer, Heidelberg
108. Kelso JAS, Fuchs A (1995) Self-organizing dynamics of the human brain: Critical instabilities and Sil'nikov chaos. Chaos 5(1):64–69
109. Kelso JAS, Haken H (1995) New laws to be expected in the organism: Synergetics of brain and behavior. In: Murphy M, O'Neill L (eds) What is Life? The Next 50 Years. Cambridge University Press, Cambridge
110. Kelso JAS, Engstrøm DA (2006) The Complementary Nature. MIT Press, Cambridge
111. Kelso JAS, Tognoli E (2007) Toward a complementary neuroscience: Metastable coordination dynamics of the brain. In: Kozma R, Perlovsky L (eds) Neurodynamics of Higher-level Cognition and Consciousness. Springer, Heidelberg, pp 39–60
112. Kelso JAS, Southard D, Goodman D (1979) On the nature of human interlimb coordination. Science 203:1029–1031
113. Kelso JAS, Scholz JP, Schöner G (1986) Nonequilibrium phase transitions in coordinated biological motion: Critical fluctuations. Phys Lett A 118:279–284
114. Kelso JAS, Scholz JP, Schöner G (1988) Dynamics governs switching among patterns of coordination in biological movement. Phys Lett A 134:8–12
115. Kelso JAS, DelColle JD, Schöner G (1990) Action-Perception as a pattern formation process. In: Jeannerod M (ed) Attention and Performance XIII. Erlbaum, Hillsdale, pp 139–169
116. Kelso JAS, Buchanan JJ, Wallace SA (1991) Order parameters for the neural organization of single, multijoint limb movement patterns. Exp Brain Res 85:432–444
117. Kelso JAS, Fuchs A, Jirsa VK (1999) Traversing scales of brain and behavioral organization. I-III. In: Uhl C (ed) Analysis of neurophysiological brain functioning. Springer, Berlin, pp 73–125
118. Kelso JAS, Buchanan JJ, DeGuzman GC, Ding M (1993) Spontaneous recruitment and annihilation of degrees of freedom in biological coordination. Phys Lett A 179:364–368
119. Kelso JAS, Holt KG, Rubin P, Kugler PN (1981) Patterns of human interlimb coordination emerge from the properties of non-linear oscillatory processes: Theory and data. J Motor Behav 13:226–261
120. Kelso JAS, Tuller B, Bateson EV, Fowler CA (1984) Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. J Exp Psychol Hum Perc Perf 10:812–832
121. Kelso JAS, Schöner G, Scholz JP, Haken H (1987) Phase-locked modes, phase transitions and component oscillators in coordinated biological motion. Phys Scr 35:79–87
122. Kelso JAS, Fink P, DeLaplain CR, Carson RG (2001) Haptic information stabilizes and destabilizes coordination dynamics. Proc R Soc B 268:1207–1213
123. Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T (1991) Cooperative and critical phenomena in the human brain revealed by multiple SQUIDS. In: Duke D, Pritchard W (eds) Measuring Chaos in the Human Brain. World Scientific, New Jersey, pp 97–112
124. Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T (1992) A phase transition in human brain and behavior. Phys Lett A 169:134–144
125. Kelso JAS, Fuchs A, Holroyd T, Lancaster R, Cheyne D, Weinberg H (1998) Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392:814–818
126. Knight CA (2004) Neuromotor issues in the learning and control of golf skill. Res Q Exerc Sport 75:9–15
127. Koch C (2005) The Quest for Consciousness. Roberts and Co, Englewood
128. Kostrubiec V, Zanone PG (2002) Memory dynamics: Distance between the new task and existing behavioral patterns affects learning and interference in bimanual coordination. Neurosci Lett 331:193–197
129. Kostrubiec V, Tallet J, Zanone P-G (2006) How a new behavioral pattern is stabilized with learning determines its persistence and flexibility in memory. Exp Brain Res 170:238–244
130. Kryukov VI (1991) An attention model based on principle of dominanta. In: Holden AV, Kryukov VI (eds) Neurocomputers and Attention. 1. Neurobiology, synchronization and chaos. Manchester University Press, Manchester, pp 319–352
131. Kuhn TS (1962) The Structure of Scientific Revolutions. University of Chicago Press, Chicago
132. Kuramoto Y (1984) Chemical oscillations, waves, and turbulences. Springer, Berlin
133. Lagarde J, Kelso JAS (2006) Binding of movement, sound and touch: Multimodal coordination dynamics. Exp Brain Res 173:673–688
134. Lagarde J, Peham C, Licka T, Kelso JAS (2005) Coordination dynamics of the horse-rider system. J Motor Behav 37:418–424
135. Lagarde J, Deguzman GC, Oullier O, Kelso JAS (2006) Interpersonal interactions during boxing: Data and model. J Sport Exerc Psychol 28:S108
136. Latash ML, Anson JG (2006) Synergies in health and disease: Relations to adaptive changes in motor coordination. Phys Ther 86:1151–1160
137. Lee TD, Blandin Y, Proteau L (1996) Effects of task instructions and oscillation frequency on bimanual coordination. Psychol Res 59:100–106
138. Lewis MD (2000) The promise of dynamic systems approaches for an integrated account of human development. Child Dev 71:36–43
139. Liese T, Cohen A (2007) Nonlinear oscillators at our fingertips. Am Math Mon 114:14–28
140. Magne C, Kelso JAS (2008) A dynamical framework for human skill learning. In: Guadagnoli M, Etnyre B (eds) Brain, Behavior and Movement. Elsevier, North Holland
141. Mainzer K (1994) Thinking in Complexity. Springer, Berlin
142. Marin L, Bardy BG, Bootsma RJ (1999) Level of gymnastic skill

as an intrinsic constraint on postural coordination. J Sports Sci 17:615–626

143. Mayville JM, Bressler SL, Fuchs A, Kelso JAS (1999) Spatiotemporal reorganization of electrical activity in the human brain associated with a timing transition in rhythmic auditory-motor coordination. Exp Brain Res 127:371–381
144. Mayville JM, Fuchs A, Ding M, Cheyne D, Deecke L, Kelso JAS (2001) Event-related changes in neuromagnetic activity associated with syncopation and synchronization timing tasks. Hum Brain Mapp 14:65–80
145. Mechsner F, Kerzel D, Knoblich G, Prinz W (2001) Perceptual basis of bimanual coordination. Nature 414:69–73
146. Meyer-Lindenberg A, Ziemann U, Hajak G, Cohen L, Berman KF (2002) Transitions between dynamical states of differing stability in the human brain. Proc Nat Acad Sci USA 99:10948–10953
147. Monno A, Chardenon A, Temprado JJ, Zanone PG, Laurent M (2000) Effects of attention on phase transitions between bimanual coordination patterns: A behavioral and cost analysis in humans. Neurosci Lett 283:93–96
148. Néda Z, Ravasz E, Vicsek T, Brechet Y, Barabasi AL (2000) Physics of the rhythmic applause. Phys Rev E 61:6987–6992
149. Néda Z, Ravasz E, Vicsek T, Brechet Y, Barabasi AL (2000) Self-organization in the concert hall: The dynamics of rhythmic applause. Nature 403
150. Newell KM, Molenaar PCM (1998) Applications of nonlinear dynamics to developmental process modelling. Erlbaum, Mahwah
151. Newell KM, Liu Y-T, Mayer-Kress G (2008) Landscapes beyond the HKB model. In: Fuchs A, Jirsa VK (eds) (2008) Coordination: Neural, behavioral and social dynamics. Springer, Heidelberg
152. Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems. Wiley, New York
153. Nicolis G, Prigogine I (1993) Exploring complexity. WH Freeman, San Francisco
154. Nourrit D, Delignieres D, Caillou N, Deschamps T, Lauriot B (2003) On discontinuities in motor learning: a longitudinal study of complex skill acquisition on a ski-simulator. J Motor Behav 35:151–170
155. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford Publications, New York
156. Oullier O, Jantzen KJ (2008) Neural indices of behavioral instability in coordination dynamics. In: Fuchs A, Jirsa VK (eds) Coordination: Neural, Behavioral and Social Dynamics. Springer, Heidelberg, pp 205–227
157. Palut Y, Zanone P-G (2005) A dynamical analysis of tennis: concepts and data. J Sports Sci 23:1021–1032
158. Palva S, Palva JM (2007) New vistas for α-frequency band oscillations. Trends Neurosci 30:150–158
159. Pellecchia GL, Shockley K, Turvey MT (2005) Concurrent cognitive task modulates coordination dynamics. Cogn Sci 29:531–557
160. Perlovsky L, Kozma R (eds) (2007) Neurodynamics of Higher-level Cognition and Consciousness. Springer, Heidelberg
161. Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization: A universal concept in nonlinear science. Cambridge University Press, Cambridge
162. Port RF, van Gelder T (eds) (1995) Mind as motion: Explorations in the dynamics of cognition. MIT Press, Cambridge
163. Prigogine I, Stengers I (1984) Order out of chaos: man's new dialogue with nature. Bantam Books, London
164. Rosenbaum DA (2005) The Cinderella of psychology: The neglect of motor control in the science of mental life and behavior. Am Psychol 60:308–317
165. Saltzman EL, Kelso JAS (1987) Skilled actions: A task dynamic approach. Psychol Rev 94:84–106
166. Schaal S, Sternad D, Osu H, Kawato M (2004) Rhythmic arm movement is not discrete. Nat Neurosci 7:1136–1143
167. Schalow G (2002) Recovery from spinal cord injury achieved by 3 months of coordination dynamic therapy. Electromyogr Clin Neurophysiol 42:367–376
168. Schalow G (2005) Phase and frequency coordination between neuron firing as an integrative mechanism of human CNS self-organization. Electromyogr Clin Neurophysiol 45:369–83
169. Schalow G, Jaiqma P (2005) Cerebral palsy improvement achieved by coordination dynamics therapy. Electromyogr Clin Neurophysiol 45:433–445
170. Schmidt RC, Carello C, Turvey MT (1990) Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. J Exp Psychol Hum Perc Perf 16:227–247
171. Scholz JP, Kelso JAS (1990) Intentional switching between patterns of bimanual coordination is dependent on the intrinsic dynamics of the patterns. J Motor Behav 22:98–124
172. Scholz JP, Kelso JAS, Schöner G (1987) Nonequilibrium phase transitions in coordinated biological motion: Critical slowing down and switching time. Phys Lett A 8:390–394
173. Schöner G, Kelso JAS (1988) A dynamic pattern theory of behavioral change. J Theor Biol 135:501–524
174. Schöner G, Kelso JAS (1988) Dynamic pattern generation in behavioral and neural systems. Science 239:1513–1520. Reprinted in: Kelner KL, Koshland DE Jr (eds) Molecules to Models: Advances in Neuroscience, pp 311–325
175. Schöner G, Haken H, Kelso JAS (1986) A stochastic theory of phase transitions in human hand movement. Biol Cybern 53:247–257
176. Schöner G, Jiang W-Y, Kelso JAS (1990) A synergetic theory of quadrupedal gaits and gait transitions. J Theor Biol 142:359–391
177. Schöner G, Zanone PG, Kelso JAS (1992) Learning as change of coordination dynamics: Theory and experiment. J Motor Behav 24:29–48
178. Schrödinger E (1944) What is Life? The physical aspect of the living cell. Camridge University Press, Cambridge
179. Seifert L, Chollet D, Bardy BG (2004) Effect of swimming velocity on arm coordination in the front crawl: a dynamic analysis. J Sports Sci 22:651–660
180. Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Chicago
181. Sheets-Johnstone M (1999) The primacy of movement. John Benjamin, Amsterdam
182. Sheets-Johnstone M (2004) Preserving integrity against colonization. Phenomenol Cogn Sci 3:249–261
183. Shockley K, Turvey MT (2006) Dual-task influences on retrieval from semantic memory and coordination dynamics. Psychon Bull Rev 13:985–990
184. Singer W (1999) Neural synchrony: a versatile code for the definition of relations. Neuron 24:49–65
185. Spencer RM, Zelaznik H, Diedrichson J, Ivry RB (2003) Disrupted timing of discontinous but not continuous movements by cerebellar lesions. Science 300:1437–1439

186. Spivey MJ (2007) The continuity of mind. Oxford University Press, New York
187. Sporns O (2004) Complex neural dynamics. In: Jirsa VK, Kelso JAS (eds) Coordination Dynamics: Issues and trends. Springer, Berlin, pp 197–215
188. Sternad D, Dean WJ (2005) Rhythmic and discrete elements in multijoint coordination. Brain Res 989:151–172
189. Sternad D, de Rugy A, Pataky T, Dean WJ (2002) Interaction of discrete and rhythmic movements over a wide range of periods. Exp Brain Res 147:162–174
190. Summers JJ, Maeder S, Hiraga CY, Alexander JRM (2008) Coordination dynamics and attentional costs of continuous and discontinuous bimanual circle drawing movements. Hum Mov Sci (in press)
191. Swinnen SP (2002) Intermanual coordination: From behavioural principles to neural-network interactions. Nat Rev Neurosci 3:350–361
192. Swinnen S, Heuer H, Massion J, Casaer P (eds) (1994) Interlimb Coordination: Neural, Dynamical and Cognitive Constraints. Academic Press, San Diego
193. Temprado JJ, Zanone PG, Monno A, Laurent M (1999) Attentional load associated with performing and stabilizing preferred bimanual patterns. J Exp Psychol Hum Perc Perf 25:1579–1594
194. Temprado JJ, Monno A, Zanone PG, Kelso JAS (2002) Attentional demands reflect learning-induced alterations of bimanual coordination dynamics. Eur J Neurosci 16:1–6
195. Thelen E, Smith LB (eds) (1994) Dynamic Approach to Development. MIT Press, Cambridge
196. Thelen E, Kelso JAS, Fogel A (1987) Self-organizing systems and infant motor development. Dev Rev 7:39–65
197. Thelen E, Skala KD, Kelso JAS (1987) The dynamic nature of early coordination: Evidence from bilateral leg movements in young infants. Dev Psychol 23:179–186
198. Thompson E (2007) Mind in life. Harvard, Cambridge
199. Tognoli E, Kelso JAS (2008) Brain coordination dynamics: true and false faces of phase synchrony and metastability. Ms under review and available from the authors
200. Treffner PJ, Kelso JAS (1999) Dynamic encounters: Long memory during functional stabilization. Ecol Psychol 11:103–137
201. Treffner PJ, Turvey MT (1996) Symmetry, broken symmetry and handedness in bimanual coordination dynamics. Exp Brain Res 107:463–478
202. Tuller B, Case P, Ding M, Kelso JAS (1994) The nonlinear dynamics of speech categorization. J Exp Psychol Hum Perc Perf 20:1–16
203. Turvey MT (2004) Impredicativity, dynamics and the perception-action divide. In: Jirsa VK, Kelso JAS (eds) (2004) Coordination Dynamics: Issues and Trends. Springer, Berlin
204. Turvey MT (2007) Action and perception at the level of synergies. Hum Mov Sci 26:657–697
205. Uhl C, Friedrich R, Haken H (1995) Analysis of spatiotemporal signals of complex systems. Phys Rev E 51:3890–3900
206. Vallacher RR, Nowak A (1997) The emergence of dynamical social psychology. Psychol Inq 8:73–99
207. van Den Berg C, Beek PJ, Wagenaar RC, Van Wieringen PC (2000) Coordination disorders in patients with Parkinson's disease. Exp Brain Res 134:174–186
208. van Geert P (1994) Dynamic systems of development: Change between complexity and chaos. Prentice Hall, New York
209. van Gelder TJ (1998) The dynamical hypothesis in cognitive science. Behav Brain Sci 21:1–14
210. van Mourik AM (2006) Structure from randomness. A window into coordination. Ph D Thesis, Free University of Amsterdam
211. van Mourik AM, Dafferthofer A, Beek PJ (2006) Determinsitic and stochastic features of rhythmic human movement. Biol Cybern 94:233–244
212. Varela FJ, Lachaux J-P, Rodriguez E, Martinerie J (2001) The brainweb: Phase synchronization and large-scale integration. Nat Rev Neurosci 2:229–239
213. Volman MJM (1997) Rhythmic coordination dynamics in children with and without a developmental coordination disorder. Ph D Thesis, University of Groningen
214. von der Malsburg C (1981) The correlation theory of brain function. Internal Report, pp 81–82. MPI Biophysical Chemistry, Goettingen
215. von Holst E (1938/73) The behavioral physiology of man and animals. In: Martin R (ed) The collected papers of Erich von Holst. University of Miami Press, Coral Gables
216. Wallenstein GV, Kelso JAS, Bressler SL (1995) Phase transitions in spatiotemporal patterns of brain activity and behavior. Physica D 84:626–634
217. Warren WH (2006) The dynamics of perception and action. Psychol Rev 113:358–389
218. Wimmers RH, Beek PJ, van Wieringen PCW (1992) Phase transitions in rhythmic tracking movements: a case of unilateral coupling. Hum Mov Sci 11:217–226
219. Winfree AT (2002) On emerging coherence. Science 298:2336–2337
220. Yuste R, MacLean JN, Smith J, Lansner A (2005) The cortex as a central pattern generator. Nat Rev Neurosci 6:477–483
221. Zaal FT, Bingham GP, Schmidt RC (2000) Visual perception of mean relative phase and phase variability. J Exp Psychol Hum Perc Perf 26:1209–1220
222. Zanone PG, Kelso JAS (1997) The coordination dynamics of learning and transfer: Collective and component levels. J Exp Psychol Hum Perc Perf 23:1454–1480
223. Zanone PG, Kelso JAS (1992) The evolution of behavioral attractors with learning: Nonequilibrium phase transitions. J Exp Psychol Hum Perc Perf 18/2:403–421

Books and Reviews

Jirsa VK, Kelso JAS (eds) (2004) Coordination Dynamics: Issues and Trends. Springer, Berlin

Kugler PN, Kelso JAS, Turvey MT (1980) Coordinative structures as dissipative structures I Theoretical lines of convergence. In: Stelmach GE, Requin J (eds) Tutorials in motor behavior. North Holland, Amsterdam

Murphy M, O'Neill L (eds) (1995) What is Life? The Next 50 Years. Cambridge University Press, Cambridge

Port RF, van Gelder T (eds) (1995) Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge

Strogatz SH (1994) Nonlinear dynamics and chaos. Addison-Wesley, Reading

Tschacher W, Dauwalder JP (eds) (2003) The Dynamical Systems Approach to Cognition: Concepts and Empirical Paradigms Based on Self-organization, Embodiment and Coordination Dynamics. World Scientific, Singapore

Turvey MT (1990) Coordination. Am Psychol 45:938–953

Corporate and Municipal Bond Market Microstructure in the U.S.

MICHAEL S. PIWOWAR
Securities Litigation and Consulting Group, Inc.,
Fairfax, USA

Article Outline

Glossary

ABS Automated Bond System. The original automated limit-order market for bonds operated by the NYSE that executed orders according to strict price/time priority. ABS was replaced by the NYSE Bonds Platform in 2007.

Agency trade A bond transaction executed by a broker-dealer on behalf of another party. A broker-dealers is compensated by a commission on an agency trade.

Broker A firm that acts as an intermediary by executing agency trades.

Broker-dealer A firm that engages in both agency trades and principal trades.

Broker's broker A broker-dealer that exclusively executes agency trades of municipal bonds with other broker-dealers. Broker's brokers do not execute principal trades and they do not trade directly with public investors.

Commission A form of compensation that a customer pays a broker-dealer for executing an agency trade. Broker-dealers must explicitly disclose the commission to the customer as a separate item on the customer's trade confirmation.

Dealer A firm that engages in principal trades for its own account.

FINRA Financial Industry Regulatory Authority. The self-regulatory organization (SR0) created in July 2007 from the consolidation of NASD and the member regulation, enforcement and arbitration functions of the NYSE. FINRA rules are approved by the SEC and enforced by themselves.

FIPS Fixed Income Pricing Service. The electronic system operated by the National Association of Securities Dealers (NASD) from 1994 through 2002 to collect and disseminate real-time quotations and hourly trade reports for a subset of high-yield corporate bonds. FIPS was retired in July 2002 with the implementation of TRACE.

Market maker A specific designation made by a regulatory authority for a broker-dealer that holds itself out to trade securities by publishing regular or continuous quotations to buy (bid) or sell (offer). Currently, there are no broker-dealers regulated as market makers in the US corporate or municipal bond markets.

Mark-up and mark-down A form of compensation that a customer pays a broker-dealer for executing a principal trade. Customers pay a mark-up when they buy a bond from a broker-dealer; they pay a mark-down when they sell a bond to a broker-dealer. Unlike commissions, mark-ups and mark-downs do not need to be disclosed on customer trade confirmations.

MSRB Municipal Securities Rulemaking Board. The self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their municipal bond transactions. MSRB rules are approved by the SEC and enforced by FINRA (formerly NASD).

NASD Formerly known as the National Association of Securities Dealers. The self-regulatory organization (SRO) charged with, among other things, primary rulemaking authority over broker-dealers in connection with their corporate bond transactions. In July 2007, NASD and the member regulation, enforcement and arbitration functions of the NYSE consolidated to form FINRA.

NYSE New York Stock Exchange. Operates the NYSE Bonds Platform (formerly ABS) trading system for exchange-listed corporate bonds.

OTC securities Over the-counter securities. Securities that are not traded on an organized exchange.

Principal trade A bond transaction executed by a broker-dealer for its proprietary account. The broker-dealer is compensated by a mark-up or mark-down on a principal trade.

Riskless principal trade A principal trade in which a broker-dealer purchases a bond to satisfy a previously received order to buy, or a broker-dealer sells a bond to satisfy a previously received order to sell. The transaction is riskless to the broker-dealer because the firm does not bear any inventory (price) risk.

RTTRS (or TRS) (Real-Time) Transaction Reporting System. MSRB's municipal bond transaction reporting and dissemination system.

Serial offering A bond issuance in which several different bonds are offered with different, often consecutive, maturities. Municipal bonds are typically issued in serial offerings.

SRO Self-regulatory Organization. A non-governmental industry association that has statutory authority to regulate members through the promulgation and enforcement of rules and regulations governing business practices. The SEC oversees SRO activities and approves SRO rules.

SEC US Securities and Exchange Commission. The primary governmental overseer and regulator of US securities markets, including the corporate and municipal bond markets. Broker-dealers and SROs are overseen by the SEC's Division of Trading and Markets (formerly Division of Market Regulation).

TRACE (formerly NASD TRACE) Transaction Reporting and Compliance Engine. FINRA's corporate bond transaction reporting and dissemination system.

Definition of the Subject

The subject of this article is the microstructure of the US corporate and municipal bond markets. ▶ Treasury Market, Microstructure of the U.S. provide a complementary discussion of the microstructure of the US Treasury bond market.

Market microstructure is broadly defined as the study of the economics of markets and trading. Market microstructure research covers a wide range of interrelated topics including market structure and design issues (e. g., trading systems and rules); price formation and price discovery; strategic trading behavior; market quality, liquidity, and trading costs (explicit and implicit); information, disclosure, and transparency; and consequences of regulatory policy (intended and unintended).

While much has been written on the microstructure of equity markets since the mid-1980s, the bond markets have only recently started receiving attention from academic researchers. The development of research in both markets can largely be attributed to the availability of quality intraday trade, quote, and/or order data ("tick" data) to empirical researchers.

The seminal theoretical work in market microstructure was conducted contemporaneously with the early equity market microstructure research, and much of the underlying economics is general enough to be appropriate for the bond markets. As a result, the significant contributions of bond market research so far have been almost exclusively empirical in nature. The last study featured in this article by Green, Hollifield, and Schurhoff [23] is a notable exception.

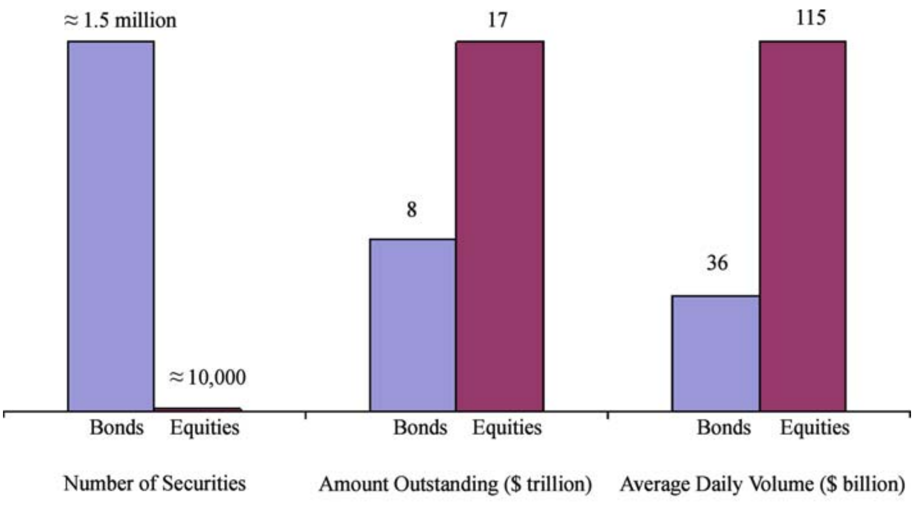
Conversely, the empirical methods developed specifically for the structure and design of equity markets are not well-suited for the bond markets. Accordingly, many of the important contributions of bond market microstructure research stem from not only the results and conclusions, but also from the development of new empirical methods. This article will provide details on some of these methods as well as discuss the important results, conclusions, and policy implications.

But, before moving on to a detailed discussion of bond market microstructure research, an important question needs to be answered. Why should we care about the bond markets? We should care because the bond markets provide an important source of capital for issuers and an important source of securities for investors. In other words, the bond markets are large. How large are they? The answer to this question depends on one's definition of size.

Figure 1 shows that an astonishingly large number, approximately 1.5 million, corporate and municipal bonds are outstanding. The vast majority of these are municipal bonds, which are typically issued in serial offerings consisting of a set of up to 20 (or more) bonds issued at the same time with different maturities. Thus, the number of bonds dwarfs the number of equities.

In terms of total dollar amounts outstanding, Fig. 1 shows that US corporate and municipal bond markets combined are roughly half the size of the US equity markets. The average daily trading volume in these bond markets is about $36 billion, which is about 1/3 of the average daily trading volume of $115 billion in the equity markets. While the discussion of the microstructure of the Treasury bond markets is left to ▶ Treasury Market, Microstructure of the U.S., it is worth noting that total US bond market trading volume (corporate, municipal, and Treasury) exceeds US equity market trading volume. Thus, no matter what measure is used, it is apparent that the bond markets offer important sources of capital for issuers and securities for investors.

The remainder of this article proceeds as follows. Section "Introduction" provides a historical overview of the US corporate and municipal bond markets. Section "Early Corporate and Municipal Bond Market Microstructure Research" through "The Links Between Bond Market Microstructure Research and Other Finance and Economics Research" review the significant contributions to the bond



Sources: SIFMA, Board of Governors of the Federal Reserve System, NYSE, AMEX, Nasdaq, OTCBB, NASD

Corporate and Municipal Bond Market Microstructure in the U.S., Figure 1
Comparison US municipal corporate and bond markets with US equity markets

market microstructure literature. Section "Early Corporate and Municipal Bond Market Microstructure Research" reviews the early corporate and municipal bond market microstructure research. Section "Fixed Income Pricing Service (FIPS) Research" reviews the research enabled by the National Association of Securities Dealer's (NASD's) Fixed Income Pricing Service that began in May 1994. Section "Municipal Bond Market Research" reviews the municipal bond market research enabled by the Municipal Securities Rulemaking Board's (MSRB's) transaction data. Section "Transaction Reporting and Compliance Engine (TRACE) Research" reviews the research enabled by the NASD's Transaction Reporting and Compliance Engine (TRACE) system that began in July 2002. Section "The Links Between Bond Market Microstructure Research and Other Finance and Economics Research" provides examples of how bond market microstructure research is linked to other areas of finance and economics research.

Introduction

Today, virtually all US corporate and municipal bond trading occurs in over the-counter (OTC) dealer markets with transparent prices. But, that was not always the case. In the early 20th century there were active and transparent markets for both corporate bonds and municipal bonds on the New York Stock Exchange (NYSE). Then, bond trading began migrating to opaque OTC dealer markets. In the late 20th century, post-trade transparency was added to the both the corporate and municipal bond OTC markets.

What factors are responsible for the evolution of the bond markets over the past century? What caused the migration of trading in the early 20th century? How (and why) was post-trade transparency added to the bond markets in the late 20th century? The brief history of US corporate and municipal bond markets below provides answers to these questions.

The Early 20th Century

Biais and Green [9] provide a fascinating historical overview of the US corporate and municipal bond markets. Early 20th century NYSE bond trading took place among the "bond crowd". Bond trading originally took place in the same trading room as stock trading, with the bond crowd organizing around three trading booths in the "bond corner" of the Exchange. In 1928, the NYSE opened a separate trading room, the "bond room", in response to increases in trading volumes. Trading in the bond room was separated into four different crowds. US corporate and municipal bonds were traded in either the "active" crowd or the "inactive" crowd. The inactive crowd was also known as the "cabinet" crowd because bond orders were written on slips of paper and filed in the bond cabinets. Foreign bonds and Government securities each had their own bond crowds. A small number of active bonds were traded on the floor in an open outcry market.

NYSE bond trading was "order-driven". The exchange collected, posted, and matched public customer orders.

Public investors paid commissions to brokers to facilitate their NYSE bond trades. All NYSE bond brokers could observe the book of available orders and the recent trades, and inform their customers about them. Thus, NYSE bond trading enjoyed a very high level of "pre-trade transparency" and "post-trade transparency". Pre-trade transparency refers to the dissemination of information about trading interests. Pre-trade information can include price (bid and ask) and depth quotations, as well as limit order prices and sizes. Post-trade transparency refers to the dissemination of information about past trades. While post-trade information includes not only prices, such as trade execution times and volumes, post-trade transparency in the bond markets is sometimes referred to as simply "price transparency". Madhavan [38] and Harris [24] provide excellent discussions all the different dimensions of transparency as well as the related market microstructure literature.

In the late 1920s, municipal bond trading migrated to the over the-counter (OTC) market. Corporate bond trading migrated to the OTC market in the 1940s. Biais and Green [9] examine a number of potential explanations for the decline in municipal and corporate bond trading on the NYSE. They find that the decline of exchange trading in bonds was not due to a decline in the supply of bonds outstanding or a decline in listings in response to costly rules and regulations promulgated by the newly created SEC.

Biais and Green [9] find that the migration of bond trading from the NYSE to the OTC markets coincided with changes in the investor base. In the late 1920s, retail investor interest in municipal bonds waned, as they became more attracted to the higher returns on equities. As retail interest in municipal bonds waned, institutions became the dominant investor in the market. During the 1940s, a similar shift in the relative importance of retail investors and institutional investors occurred in the corporate bond market. Biais and Green [9] conclude that the migration of bond trading from the NYSE to the OTC markets was an evolution in response to the changing investor base.

Biais and Green [9] provide evidence that institutions fared better in OTC bond markets and argue that the dealers were happy to accommodate this new class of dominant investors. Because liquidity was no longer concentrated on a centralized transparent exchange, retail investors were effectively forced into trading with dealers in these decentralized opaque OTC markets. Not surprisingly, retail investors faredjt worse in these markets. Both municipal and corporate bond transaction costs increased significantly for retail investors.

The Late 20th Century and Early 21st Century

While the most significant change in the bond markets in the early 20th century was a migration of trading from the exchange to OTC, the most significant change in the late 20th century was the introduction of price transparency. Unlike trading migration, bond market transparency was not caused by market forces. Rather, transparency was added to the bond markets by direct regulatory intervention.

The Municipal Securities Rulemaking Board (MSRB) introduced price transparency to the municipal bond market. The MSRB was created by Congress in 1975 as the self-regulatory organization (SRO) charged with primary rule-making authority over broker-dealers in connection with their municipal bond transactions.

The MSRB began publicly disseminating municipal bond price information in January 1995. "Interdealer Daily Reports" provided statistics on total interdealer market activity reported for the previous day, as well as information about price and volume for each security that was "frequently traded" on that day. The MSRB defined frequently traded securities to be securities with four or more interdealer transactions on a particular day. The Interdealer Daily Report included the total par value traded, the daily high and low price, and the average price of trades having a par value between $100,000 and $1 million for each frequently traded issue. Transaction price information on securities with three or fewer interdealer transactions on a particular day ("infrequently traded" securities) was not disseminated.

In August 1998, the MSRB began producing "Combined Daily Reports". The Combined Daily Reports merged information from customer and interdealer transactions and provided daily high, low, and average prices for frequently traded securities of municipal securities on a one-day delayed basis. The frequently traded threshold was four transactions per day, taking into account both customer and interdealer transactions.

In January 2000, the MSRB began publicly disseminating transaction details on individual trades in frequently traded securities through "Daily Transaction Reports". Trade information on infrequently traded securities was still not disseminated until October 2000, when the MSRB began producing "Monthly Comprehensive Reports". These reports provided information on a one-month delayed basis for all transactions from the previous month, including infrequently traded issues.

By June 2003, the MSRB was publicly disseminating transaction details for all trades in all securities (frequently traded and infrequently traded) on a one-day lag basis through "T+1 Daily Reports". In January 2005, the MSRB began disseminating prices on a real-time basis through its Real-Time Transaction Reporting System (RTTRS or TRS).

The National Association of Securities Dealers (NASD) introduced price transparency to the corporate bond market. NASD (now FINRA) is the self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their corporate bond transactions. Regulatory concern for price transparency spiked in the late 1980s and early 1990s.

At that time, the high-yield corporate bond market faced unprecedented instability, highlighted by insider trading scandals and the ultimate collapse of the dominant dealer and underwriter Drexel Burnham Lambert. Concern over future market instability, along with the recognition of a need for better monitoring, led to regulatory intervention that provided a small degree of transparency in this market segment. The Fixed Income Pricing Service (FIPS) began in 1994. FIPS was the result of the SEC encouraging NASD to develop an electronic reporting and dissemination facility for non-convertible high-yield corporate bonds.

But, FIPS only provided partial transparency for this particular segment of the corporate bond market. While every trade in FIPS-eligible bonds was reported to FIPS, only summary information on a small subset (50 bonds) of the most active bonds was disseminated to the public. Alexander, Edwards, and Ferri [2] point out that some members of the SEC staff at that time feared that adding price transparency to less active bonds could possibly harm the market. FIPS added both pre-trade transparency and post-trade transparency to the market by disseminating quotations and hourly trade summaries, respectively. The hourly trade summaries contained high and low prices as well as total trading volume.

Many bond market participants and some SEC staff felt that FIPS added a sufficient amount of transparency to the corporate bond market. SEC Chairman Arthur Levitt disagreed. In 1998, he gave a speech entitled *The Importance of Transparency in America's Debt Market* in which he famously quipped "The sad truth is that investors in the corporate bond market do not enjoy the same access to information as a car buyer or a homebuyer or, dare I say, a fruit buyer". To address the lack of price transparency in the corporate bond market, he called on NASD to take several related actions. He called on NASD to adopt rules requiring dealers to report all corporate bond transactions; to develop a system to receive all corporate bond transaction information; to create a database of the transactions, and in conjunction, create a surveillance program to better detect fraud in corporate bonds; and, to disseminate the bond transaction prices to the public in order to help them make better investment decisions.

NASD responded by developing the Transaction Reporting and Compliance Engine (TRACE) system, which began operation in July 2002. Corporate bond dealers were required to report all transaction in TRACE-eligible securities. TRACE-eligible securities included investment grade and high-yield debt, convertible and non-convertible debt, and publicly issued debt and privately issued (Rule 144A) debt. While all TRACE-eligible transactions were reported to TRACE from the beginning of its operation, the dissemination of the trade information was phased-in over time. The phase-in approach was adopted by NASD, and approved by the SEC, because of industry concerns that adding transparency to the bond market would somehow harm liquidity.

The TRACE phase-in approach began with the dissemination of trade information on the largest, highest-rated bonds first. Price transparency was introduced to smaller and lower-rated bonds over time. By February 2005, prices were transparent on effectively 99% of trades, and by the end of that year, pricing information on all TRACE-eligible trades was being disseminated on a real-time basis to the public.

By the beginning of the 21st century, investors (and market microstructure researchers) were able to access an unprecedented amount of information about the OTC municipal and corporate bond markets from the post-trade transparency brought by TRS and TRACE, respectively. It is worth noting that bond trading never completely migrated to the OTC markets. The NYSE continues to list and trade some bonds. The NYSE developed the Automated Bond System (ABS), a computerized limit-order market for bonds, in an effort to encourage the migration of trading back to the exchange. The displayed public limit orders on ABS provided pre-trade transparency for some bonds.

However, the vast majority of bonds are not listed on the NYSE ABS or any other exchange, so all of the trading in these bonds occurs in the OTC markets. Moreover, for many of the bonds that are listed on the NYSE ABS, a majority of their trades still occur over the-counter. Therefore, the early 21st century bond markets can be characterized as dealer markets with a high degree of post-trade transparency, but with virtually no pre-trade transparency. It remains to be seen whether market forces, regulatory initiatives, or some combination of the two will eventually lead to the introduction of some form of pre-trade transparency, the emergence of bond market-makers, and/or a migration of trading back to an order-driven market in the US corporate and municipal bond markets.

Early Corporate and Municipal Bond Market Microstructure Research

Early bond market microstructure researchers were forced to rely on data sources that covered only certain segments of the market because comprehensive bond market data simply did not exist. Bond dealers kept their own trading records because there was no central reporting facility. Bond dealers were not required to publicly disseminate their trades. Researchers, as well as investors and regulators, were able to see snapshots of parts of the bond markets, but no one was able to see the complete picture.

But, even with the data limitations, early bond market researchers found creative ways to tease out useful information. Their initial findings shed the first light on the opaque bond markets. For example, Schultz [42] provides indirect evidence that higher bond market trading costs may be attributable to the lack of transparency. Variants of their original empirical methods continue to be used by more recent researchers.

Any encyclopedic article on corporate bond market microstructure research would be incomplete if it did not mention the efforts of the Fixed Income Research Program (FIRP), and more importantly its founder Professor Arthur Warga, in promoting early bond market research. Art Warga's influence on the development of the bond market microstructure literature extends beyond his own research. He collected, consolidated, cleaned, and organized various fragmented sources of bond market data to create the Fixed Income Securities Database (FISD), which he made accessible to academic and regulatory researchers. Many within the market microstructure research community informally refer to the FISD as simply the "Warga database".

Warga (1991)

Warga [44] uses an econometric model to investigate bond pricing discrepancies that arise when researchers (and commercial bond pricing services) use data from the two different sources that were generally available in 1990. The one source was exchange data in the form of actual transaction prices from the NYSE Automated Bond System (ABS). The other source was OTC dealer data in the form trader-quoted prices.

Warga [44] denotes the unobserved, true value of bond i as P_i^* and the unobserved, true bid-ask spread as BA_i. He assumes that P_i^* is the midpoint of BA_i. He also assumes that the prices/quotes observed in both markets are unbiased in the sense that they deviate from the true unobserved prices/quotes by a random error term. Then, for month-end NYSE transaction prices (P_{NY}):

$$P_{NY_i} = P_i^* + u_i \ ,$$

and, for month-end Lehman Brothers bid quotes (P_B):

$$P_{B_i} = P_i^* - \frac{1}{2} BA_i + \zeta_i \ .$$

Combining these two equations and letting $\varepsilon_\iota = \zeta_\iota + \mu_\iota$ yields:

$$P_{B_i} - P_{NY_i} = -\frac{1}{2} BA_i + \varepsilon_i \ .$$

Squaring both sides results in:

$$\left(P_{B_i} - P_{NY_i}\right)^2 = \frac{1}{4} (BA_i)^2 - (BA_i)\,\varepsilon_i + \varepsilon_i^2 \ .$$

Assuming the random error terms are orthogonal to prices/quotes, the expected squared price discrepancies is:

$$E\left[\left(P_{B_i} - P_{NY_i}\right)^2\right] = \frac{1}{4} (BA_i)^2 + \sigma_{\varepsilon i}^2 \ ,$$

where σ_ε^2 equals the variance of the discrepancy.

Warga [44] regresses the squared price discrepancies on six observable liquidity-related variables – bond rating, duration, NYSE dollar trading volume, bond age, issue amount outstanding, and the time of trade of the last trade price on the NYSE – with the following equation:

$$\begin{aligned}\left(P_{B_i} - P_{NY_i}\right)^2 &= \alpha_0 + \alpha_1 MOODYS + \alpha_2 DURTN \\ &+ \alpha_3 OUTSTD + \alpha_4 DVOL + \alpha_5 AGE + \alpha_6 TIME + \omega \ .\end{aligned}$$

Warga [44] finds that squared discrepancies are larger for bonds with lower credit ratings, higher duration, smaller issue sizes, lower trading volume, and trade prices that occurred earlier in the day. While he finds that these variables are capable of explaining some of the observed variation in price discrepancies, he also concludes that commingling exchange and dealer bond pricing data does not induce any biases.

Hong and Warga (2000)

Hong and Warga [29] use a benchmark method to estimate average daily effective spreads with newly available trade data. They obtain exchange market data from the NYSE ABS and OTC dealer market data from the Capital Access International (CAI) database. CAI obtains trading data on insurance companies, mutual funds, and pension funds from various regulatory filings.

They calculate the effective spread for a given bond on a given day as the dollar-volume-weighted average price transacted at the ask minus the dollar-volume-weighted average price transacted at the bid:

$$\sum_{i=1}^{N} P_i^{\mathrm{A}} W_i^{\mathrm{A}} - \sum_{j=1}^{M} P_j^{\mathrm{B}} W_j^{\mathrm{B}} ,$$

where P_i^{A} is the price of transaction i occurring at the ask, W_i^{A} is the dollar-value weight of transaction i, and N is the number of transactions occurring at the ask for a given bond on a given day. Similarly, P_j^{B} is the price of transaction j occurring at the bid, W_j^{B} is the dollar-value weight of transaction j, and M is the number of transactions occurring at the bid for a given bond on a given day.

Hong and Warga [29] find that the average daily effective spreads for corporate bond transactions occurring on the NYSE ABS that involve at least 10 bonds is about $0.21 for investment grade bonds and about $0.19 for high-yield bonds. For corporate bond trades occurring in the OTC dealer market, they find that the average daily effective spreads is about $0.13 for investment grade bonds and about $0.19 for high-yield bonds. They note that these spread estimates are smaller than previous estimates based on data from an earlier period, which is consistent with evidence that corporate bond spreads may have declined over time.

Hong and Warga [29] also find that OTC dealer market spreads exhibit much larger dispersion than NYSE ABS spreads. The standard deviations of daily effective spreads for the dealer market are two to three times larger than those for the exchange market. This result suggests that investors, particularly uninformed retail investors, could benefit from more transparency in the OTC markets.

Schultz (2001)

Schultz [42] also uses CAI institutional trade data. He develops an econometric model similar to Warga [44] to estimate average round trip corporate bond trading costs from institutional trade data and estimated contemporaneous bid quotes.

Schultz [42] estimates daily corporate bond bid quotes from the month-end bid quotes available from the Warga database. He develops a threestep estimation procedure that uses daily Treasury bond bid quotes, based on the observation that most of the day-to-day changes in investment grade corporate bond prices are explained by changes in the level of risk-free interest rates.

The first step is to calculate a predicted month-end quote for each corporate bond by taking its previous month-end quote and multiplying it by the change in the price of Treasury bonds of similar duration. The second step is to subtract the predicted month-end quote from the actual month-end quote. This calculation yields the monthly pricing error from predicting that the change in the corporate bond prices is exactly the same as the change in Treasury bond prices. The monthly pricing error is converted to an average daily pricing error by dividing it by the number of trading days in the month. The third step is to estimate the bid quote for a particular within-month trade date by starting with the previous end-of-month quote and adding on the average daily pricing error times the number trading days since the previous month end.

Schultz [42] finds that his bid quote estimates are accurate for investment grade bonds, but not for high-yield bonds. This is not surprising, since changes in high-yield prices are more often due to changes in firm-specific factor than changes in Treasury bond prices. Therefore, he does not attempt to estimate trading costs for high-yield bonds with this methodology.

For investment grade bonds, Schultz [42] estimates round-trip transactions costs by regressing the difference between the CAI trade prices and his estimate of the contemporaneous bid quote on a dummy variable that takes the value of 1 for buys and 0 for sells:

$$\Delta_i = \alpha_0 + \alpha_1 D_i^{\mathrm{Buy}} + \varepsilon_i ,$$

where Δ_i is the price of trade i minus estimated bid price and D_i^{Buy} equals one if trade i is a buy and zero otherwise. The coefficient α_i is an estimate of the average round-trip transaction costs. His estimate of the average round-trip transaction costs across all trades is about $0.27 per $100 of par value.

Schultz [42] also examines the determinants of corporate bond trading costs with the following regression:

$$\begin{aligned}\Delta_i &= \alpha_0 + \alpha_1 D_i^{\mathrm{Buy}} + \alpha_2 S_i + \alpha_3 D_i^{\mathrm{Inst}} + \alpha_4 D_i^{\mathrm{Deal}} \\ &\quad + \alpha_5 D_i^{\mathrm{Inst}} D_i^{\mathrm{Deal}} + \alpha_6 D_i^{\mathrm{Inst}} S_i + \alpha_7 D_i^{\mathrm{Deal}} S_i + \varepsilon_i ,\end{aligned}$$

where S_i is the signed (positive for buys, negative for sells) natural logarithm of the dollar trade size, D_i^{Inst} is a dummy variable that takes a value of one for buys and negative one for sells by one of the 20 most active institutions and a value of zero otherwise, and D_i^{Deal} is a dummy variable that takes a value of one for buys and negative one for sells if the trade involves one of the 12 active dealers and a value of zero otherwise. For the interactive term for active institutions and dealers, $D_i^{\text{Inst}} D_i^{\text{Deal}}$, the product of the dummies is positive for buys and negative for sells when the trade involves both an active institution and an active dealer and a value of zero otherwise.

Schultz [42] finds that institutional corporate bond trading costs decline with trade size. He does not find any evidence that trading costs are related to credit rating. But, this is not surprising given the fact that his analysis is limited to the four investment grade rating classes (Aaa, Aa, A, Baa).

Schultz [42] finds that trading costs are lower when a large bond dealer is used. In other words, small bond dealers charge more than large ones. He also finds that inactive institutions pay more than twice as much as active institutions to trade the same bonds. Schultz [42] attributes this result to the lack of transparency in the corporate bond market during his sample period. In an opaque market, obtaining price information is costly, so only active institutions will find it worthwhile to bear them.

Chakravarty and Sarkar (2003)

Chakravarty and Sarkar [12] use CAI data and benchmark methods to calculate "traded bid ask-spreads" over one-day, two-day, and five-day windows in the corporate, municipal, and Treasury bond markets. Similar to Hong and Warga [29], they define the traded bid-ask spread per day as the difference between its mean daily selling price and its mean daily buying price.

To check the sensitivity of their estimates to the requirement of one buy trade and one sell trade for each bond day, Chakravarty and Sarkar [12] calculate spreads over non-overlapping two-day and five-day windows. Their two-day traded bid-ask spread is calculated as the difference between the two-day means of the selling prices and the buying prices.

Chakravarty and Sarkar [12] find that the mean traded bid-ask spread per day per \$100 par value is \$0.21 for corporate bonds, \$0.23 for municipal bonds, and \$0.08 for Treasury bonds. In all three markets, they find that spreads increase with longer time-to-maturity, lower credit ratings, and lower trading volume. These results suggest that spreads are positively related to interest rate risk and credit risk, and negatively related to trading activity. For corporate bonds, Chakravarty and Sarkar [12] find that spreads increase with age.

Chakravarty and Sarkar [12] pool observations across all three bond markets for cross-market comparisons. After controlling for credit risk, Chakravarty and Sarkar [12] find no significant difference in the spreads of corporate bonds and Treasury bonds, but they find that municipal bonds have higher spreads.

Fixed Income Pricing Service (FIPS) Research

FIPS provided new price and volume data that allowed market microstructure researchers to conduct studies that were not previously possible. Alexander, Edwards, and Ferri [1] use FIPS volume data to test various hypotheses about bond liquidity. Alexander, Edwards, and Ferri [2] use FIPS returns and equity returns to tease out new evidence on agency conflicts between stockholders and bondholders. Hotchkiss and Ronen [31] use FIPS returns and equity returns to examine the relative informational efficiency of the corporate bond market. Somewhat surprisingly, they find that that informational efficiency of the corporate bond market is similar to the stock market.

Alexander, Edwards, and Ferri (2000a)

Alexander, Edwards, and Ferri [1] examine the determinants of trading volume of FIPS high-yield bonds using a pooled time-series cross-sectional approach. They use the following linear specification:

$$\begin{aligned} \textit{Trading Volume}_{it} &= \beta_0 + \beta_1 \text{Ln}\,(\textit{Size}_{it}) + \beta_2 \textit{Age}_{it} \\ &\quad + \beta_3 \textit{Private Equity}_{it} + \beta_4 \textit{Credit Rating}_{it} \\ &\quad + \beta_5 \textit{Duration}_{it} + \beta_6 \textit{Price Variability} + \varepsilon_{it}\,, \end{aligned}$$

where the dependent variable, *Trading Volume*, is measured in three different ways for each bond *i* in each month *t*. The three trading volume measures are the natural log of the average daily number of trades (Ln*(Trades)*), the natural log of the average daily number of bonds traded (Ln*(Bonds)*), and average daily turnover (*Turnover*). Ln*(Size)* is the natural logarithm of the issue's par value outstanding, *Age* is a dummy variable equal to one if the issue has been outstanding for less than two years, *Private Equity* is a dummy variable equal to one if the issue has no public equity outstanding in any part of the month, *Credit Rating* is a dummy variable equal to one if the issue is rated below B- by Standard & Poor's at any point during the month, *Duration* is the bond's modified duration, and *Price Variability* is the monthly average absolute value of the daily percentage change in volume-weighted price.

They find consistent results for all three measures of trading volume. Larger issues and younger issues are more heavily traded. They point out that the age result extends earlier empirical results that found that the liquidity of Treasury securities drops off a few months after issuance.

Alexander, Edwards, and Ferri [1] also find that bonds of firms without public equity trade more frequently than bonds of firms with public equity. This last finding is inconsistent with a disclosure hypothesis that predicts that more relaxed disclosure rules for firms without public equity will lead lower liquidity, as measured by trading volume. However, it is consistent with the competing substitution hypothesis that predicts that high-yield bonds of private firms will attract trading volume that otherwise would have occurred in the equity.

Alexander, Edwards, and Ferri (2000b)

Alexander, Edwards, and Ferri [2] use equity data and FIPS bond data to investigate the relationship between a firm's stock return and its bond return. They examine the long-term co-movement between a firm's bond returns and its stock returns. They also examine stock and bond returns around events typically associated with agency conflicts to see whether their co-movements provide evidence of agency conflicts.

To examine the long-term co-movement between a firm's bond returns and its stock returns, Alexander, Edwards, and Ferri [2] use three regression approaches. The first approach is a time-series regression model:

$$\begin{aligned} RB_{it} = \beta_0 + \beta_1 XRS_{it} + \beta_2 XRS_{it-1} + \beta_3 RBIND_{it} \\ + \beta_4 RBIND_{it-1} + \varepsilon_{it} \,, \end{aligned}$$

where RB_{it} is the bond return for firm i on day t, XRS is the current (t) and lagged ($t-1$) excess stock return, and $RBIND$ is the current (t) and lagged ($t-1$) high-yield bond index return, and ε_{it} is the residual bond return for firm i on day t. The second approach is a pooled time-series cross-sectional model that uses the regression equation above and follows the pooling technique of Greene (1993). The third approach is a cross-sectional regression model that follows the approach of Fama and MacBeth. For each sample day, the excess bond returns are regressed on the current and lagged excess stock returns:

$$XRB_{it} = \beta_0 + \beta_1 XRS_{it} + \beta_2 XRS_{it-1} + \varepsilon_{it} \,.$$

The estimates of β_1 in each of the three regressions show whether the stock and bond returns tend to co-move together (positive), in the opposite direction (negative), or not at all (insignificant). Alexander, Edwards, and Ferri [2] find that all three regressions produce similar results. The β_1 estimates are positive and statistically significant, indicating that excess bond returns are positively correlated with excess stock returns. But, Alexander, Edwards, and Ferri [2] point out the that the magnitudes of the coefficients suggest that the correlation is economically small.

To examine the behavior of stock and bond returns around events typically associated with agency conflicts, Alexander, Edwards, and Ferri [2] look at cumulative excess stock and bond returns around announcements of corporate events that are typically associated with wealth transfers from bondholders to stockholders, or vice versa. Events include debt issuances and redemptions, stock issuances and repurchases, dividend changes, new credit agreements, and others. They use Wilcoxon rank-sum tests to determine whether the means of the cumulative excess bond returns around potentially wealth-transferring events are significantly different from the returns otherwise. They find that the means are significantly different and that the mean cumulative excess bond returns around the wealth-transferring events is negative, while the returns are at other times are positive.

Thus, Alexander, Edwards, and Ferri [2] show that wealth-transferring corporate events (from bondholders to stockholders, or vice versa) can cause a firm's bond returns to diverge from its typical positive (weak) co-movement with its stock returns. In addition, they point out that this result is a likely factor in the weak long-term time-series correlations observed between stock and bond returns.

Hotchkiss and Ronen (2002)

Hotchkiss and Ronen [31] use FIPS data for 55 high-yield bonds to examine the informational efficiency of the corporate bond market relative to the market for the underlying stock. They find that stocks do not lead bonds in reflecting firm-specific information. They also find that pricing errors for bonds are no worse than for the underlying stocks, even on an intraday level.

Hotchkiss and Ronen [31] use a vector autoregression (VAR) approach applied to daily and hourly returns with a return-generating process that includes an interest rate risk factor and an equity market (systematic) risk factor:

$$\begin{aligned} RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{L}} RL_{t-i} \\ + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \varepsilon_t \,, \end{aligned}$$

where RB_t is the FIPS bond portfolio return, RL_t is the Lehman Intermediate Government Bond Index return, and RM_t is the S&P 500 Index return. The number of lags for the bond, interest rate, and stock returns are $nb = 3$, $ni = 0$, and $ns = 4$, respectively. Hotchkiss and Ronen [31] include lagged bond returns (RB_{t-i}) to consider autocorrelation-adjusted bond returns. They also consider a specification that replaces the Lehman index return with the default risk-free return, RD_t, as interest rate factor:

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{D}} RD_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \varepsilon_t \,.$$

Finally, they add the underlying stock (firm-specific) return, RS_t :

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{D}} RD_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{S}} RS_{t-i} + \varepsilon_t \,.$$

With these three regressions, Hotchkiss and Ronen [31] find that high-yield bond returns exhibit a very strong interest rate risk component and that this component is significantly greater for higher-rated bonds. They also find that high-yield bond returns exhibit a very strong systematic risk component and that this component is slightly weaker for higher-rated bonds.

To test whether stock returns lead bond returns, Hotchkiss and Ronen [31] conduct Granger causality tests at the daily and hourly levels. They estimate the VAR for the variable set $z_t = [RB_t, RS_t]'$ using the specification:

$$z_t = B_1 z_{t-j} + B_2 z_{t-j} + \mu_t \,,$$

where RB_t is the bond return and RS_t is the stock return, for day (hour) t, B_i are conformable matrices, and μ_t is a disturbance vector. To test whether stock returns Granger cause bond returns they estimate the following bivariate VAR model using ordinary least squares (OLS):

$$RB_t = c_1 + \sum_{i=1}^{j} a_i RB_{t-i} + \sum_{i=1}^{j} b_i RS_{t-i} + \nu_{1,t} \,,$$

where c is a constant, as and bs are coefficients, ν_t is the disturbance vector, and j is the lag length. The null hypothesis is that stock returns do not Granger cause bond returns, or that $H_0 = \left[b_i\right] = 0$, for all i. Tests of whether bond returns Granger cause stock returns are conducted in a similar way. F-tests indicate that lagged stock returns are not significant in explaining bond returns. Thus, stocks do not lead bonds in reflecting firm-specific information. The Granger causality test results also indicate that lagged bond returns are not significant in explaining stock returns.

Hotchkiss and Ronen's [31] interpretation of the Granger causality test results is that the contemporaneous correlations between stock returns and bond returns are best described as a joint reactions to common factors. This motivates an additional investigation of the comparative reaction of stocks and bonds to firm-specific information. To conduct this investigation, they examine how quickly firm-specific information contained in earnings announcements are incorporated into bond prices relative to stock prices. First, they compare reported earnings to the median of analysts' earnings forecasts and calculate the log forecast error:

$$FE_i = \ln\left(A_i / F_i\right) \,,$$

where FE_i is the log forecast error for firm i, A_i is the announced earnings per share, and F_i is the forecast earnings per share. Next, they run the following regressions to examine whether earnings information is reflected in bond returns or stock returns:

$$RB_{[-1,t]} = \alpha_0 + \alpha_1 * FE + \alpha_2 * RM_{[-1,t]} + \varepsilon$$
$$RS_{[-1,t]} = \alpha_0 + \alpha_1 * FE + \alpha_2 * RM_{[-1,t]} + \varepsilon \,,$$

where RB and RS are the bond and stock returns, respectively, for the period starting at day (hour) -1 prior to the announcement and ending at day $+7$ (hour $+14$) after the announcement, and RM is the market (S&P 500 Index) return. Both the daily and hourly regression results indicate that all information is quickly impounded into both bond prices and stock prices.

Finally, Hotchkiss and Ronen [31] compare the market quality for the high-yield FIPS bonds to the underlying stocks by examining whether price errors of different magnitudes are associated with the different markets. The estimate the following market quality measure:

$$MQ_i = 1 - 2 * \left(\sigma_{si}^2 / \sigma_{Ri}^2\right) \,,$$

where σ_{si}^2 is the variance of the pricing error described in Hasbrouck [27] and σ_{Ri}^2 is the variance of the return. The intuitive interpretation of this measure is the proportion of the total return variance that is due to fundamental variance. In general, they find that the market quality measure for bonds is no worse than for the underlying stocks.

Municipal Bond Market Research

With the MSRB's introduction of central reporting and price transparency to the municipal bond market, microstructure researchers were able to make use of a comprehensive source quality transaction-level municipal bond data for the first time. This new data provided researchers the opportunity to develop new methods, examine existing microstructure issues in greater detail, and identify new avenues of research.

Two prominent municipal bond studies, Harris and Piwowar [26] and Green, Hollifield, and Schurhoff [22], develop and use very different methods to examine various economic aspects of trading in the municipal bond market. These two studies provide independent sets of similar and robust results that support two important conclusions related to retail investors. The first is that municipal bonds are expensive for retail investors to trade. Harris and Piwowar [26] and Green, Hollifield, and Schurhoff [22] both find that, unlike in equity markets, municipal bond trading costs decrease with trade size.

The second is that "complexity" is costly for retail investors. Harris and Piwowar [26] find that "instrument complexity" makes municipal bonds more expensive to trade. Instrument complexity is measured in terms of attached features, such as calls, puts, sinking funds, credit enhancement, nonstandard interest payment frequencies, and nonstandard interest accrual methods. Green, Hollifield, and Schurhoff [22] find that "deal complexity" also increases trading costs. Bond dealers charge higher markups on more difficult trades.

Harris and Piwowar (2006)

Harris and Piwowar [26] estimate municipal bond trading costs using an econometric model. They denote the unobserved "true value" of the bond at the time t asV_t and assume that the price of a trade, P_t, is equal to V_t plus or minus a price concession that depends on whether the trade is buyer-initiated or seller-initiated. The absolute customer transaction cost, $c(S_t)$, is estimated as the effective half-spread, measured as a percentage of the price.

I^{D}_t is an indicator variable that takes a value of 1 if the trade was an interdealer trade or 0 if the trade was a customer trade. Q_t is an indicator variable that takes a value of 1 if the customer was a buyer, −1 if the customer was a seller, or 0 if it was an interdealer trade. This results in:

$$\begin{aligned} P_t &= V_t + Q_t P_t c(S_t) + I^{\mathrm{D}}_t P_t \delta_t \\ &= V_t \left(1 + \frac{Q_t P_t c(S_t) + I^{\mathrm{D}}_t P_t \delta_t}{V_t}\right) . \end{aligned}$$

The continuously compounded bond price and "true value" returns between trades t and s, r^{P}_{ts} and r^{V}_{ts} respectively, are found by taking logs of both sides, making two small approximations, and subtracting the same expression for trade s:

$$r^{\mathrm{P}}_{ts} = r^{\mathrm{V}}_{ts} + Q_t c(S_t) - Q_s c\,(S_s) + I^{\mathrm{D}}_t \delta_t - I^{\mathrm{D}}_s \delta_s \,.$$

The "true value" return r^{V}_{ts} is represented with a factor model by decomposing it into the linear sum of a time drift, a short-term municipal bond factor return, a long-term municipal bond factor return, and a bond-specific valuation factor, ε_{ts}:

$$\begin{aligned} r^{\mathrm{V}}_{ts} &= \mathit{Days}_{ts}\,(5\% - \mathit{CouponRate}) \\ &\qquad + \beta_{\mathrm{Avg}} \mathit{SLAvg}^{+}_{ts} \beta_{\mathrm{Dif}} \mathit{SLDif}^{+}_{ts}\, \varepsilon_{ts} \,, \end{aligned}$$

where Days_{ts} counts the number of calendar days between trades t and s, *CouponRate* is the bond coupon rate. SLAvg_{ts} and SLDif_{ts} are the average and difference, respectively, of continuously compounded short- and long-duration factor returns between trades t and s. The first term models the continuously compounded bond price return that traders expect when interest rates are constant and the bond's coupon interest rate differs from a notional five percent bond, the median coupon rate in their sample. The two index returns model municipal bond value changes due to changes in interest rates and tax-exempt yield spreads. Harris and Piwowar [26] use repeat sales methods to estimate these indices. They assume that the bond-specific valuation factor ε_{ts} has mean zero and variance given by

$$\sigma^2_{\varepsilon_{ts}} = N^{\mathrm{Sessions}}_{ts} \sigma^2_{\mathrm{Sessions}}$$

where $N^{\mathrm{Sessions}}_{ts}$ is the total number of full and partial trading sessions between trades t and s.

To model customer transaction costs, Harris and Piwowar [26] consider several alternative functional forms that are flexible enough to model very high average trading costs for small trade sizes and very low average trading costs for large trade sizes. Harris and Piwowar [26] choose following parsimonious expression:

$$c(S_t) = c_0 + c_1 \frac{1}{S_t} + c_2 \log S_t + \kappa_t \,,$$

where the first three terms specify the cost function that represents average trade costs and κ_t represents the unexplained variation in the observed customer trading costs. The constant term allows total transaction costs to grow in proportion to size. The second term captures fixed costs per trade and the third term allows the costs per bond to vary by size.

The Harris and Piwowar [26] time-series estimation model is obtained by combining the last four equations:

$$r^{\mathrm{P}}_{ts} - Days_{ts}\left(5\% - CouponRate\right) = c_0\left(Q_t - Q_s\right) + c_1\left(Q_t\frac{1}{S_t} - Q_s\frac{1}{S_s}\right) + c_2\left(Q_t\log S_t - Q_s\log S_s\right) + \beta_{\mathrm{SLAvg}}SLAvg^{+}_{ts}\beta_{\mathrm{SLDif}}SLDif^{+}_{ts}\eta_{ts}\,,$$

where the expression for the regression term, η_{ts}, is given by:

$$\eta_{ts} = \varepsilon_{ts} + Q_t\kappa_t - Q_s\kappa_s + I^{\mathrm{D}}_t\delta_t - I^{\mathrm{D}}_s\delta_s\,.$$

The mean of the error term is zero and its variance is:

$$\sigma^2_{ts} = N^{\mathrm{Sessions}}_{ts}\sigma^2_{\mathrm{Sessions}} + D_{ts}\sigma^2_{\delta} + (2 - D_{ts})\,\sigma^2_{\kappa}\,,$$

where D_{ts} represents the number (0, 1, or 2) of interdealer trades involved in trades t and s. For each bond, Harris and Piwowar [26] separately estimate their time-series transaction cost estimation model using an iterated least squares method, with the weight given by the inverse of the estimates of σ^2_{ts}. For a wide range of trade sizes, they calculate weighted cross-sectional mean cost estimates across all municipal bonds. Each bond's weight is given by the inverse of its estimation error variance at that trade size.

Harris and Piwowar [26] find that retail-size municipal bond trades are substantially more expensive than similar-sized equity trades. Average effective spreads in municipal bonds are almost 2% for representative retail-size trades ($20,000). They point out that this is the equivalent of almost 4 months of total annual return for a bond with a 6% yield-to-maturity.

Harris and Piwowar [26] also find that retail-size municipal bond trades are more expensive than institutional-size trades. Unlike in equities, municipal bond transaction costs decrease with trade size. Harris and Piwowar [26] also find that, unlike in equities, municipal bond transaction costs do not depend on trade frequency. They attribute these results to the lack of price transparency in the municipal bond market during their sample period.

To investigate how estimated transaction costs vary across municipal bonds, Harris and Piwowar [26] conduct cross-sectional weighted least squares regressions for various trade sizes. The dependent variable is the estimated average transaction costs in a given municipal bond at a given trade size. The weight for each bond observation is given by the inverse of the estimation error variance of its cost estimate. The independent variables include measures of credit quality, age, and instrument complexity.

Harris and Piwowar [26] show that bond trading costs increase with credit risk, time to maturity, and time since issuance. They also find that trading costs increase with instrument complexity, and that retail investors are more adversely affected by instrument complexity than institutional investors. They conjecture that investors and issuers might benefit if simpler bonds were issued.

Green, Hollifield, and Schurhoff (2007a)

Green, Hollifield, and Schurhoff [22] focus on trades that can reasonably be assumed to represent two sides of a single intermediated transaction, and employ a structural model to decompose the cost faced by a customer into a portion that represents the cost the dealer incurs and a portion attributable to the dealer's market power. They formulate and estimate a simple structural bargaining model that allows them to estimate measures of dealer bargaining power and relate it to characteristics of the trades.

Green, Hollifield, and Schurhoff [22] use a theoretical model to seek evidence that the high costs of trading are due to dealer market power and to find out how the exercise of market power depends on the characteristics of the trade. They develop a simple theoretical model of the interaction between dealers and their customers in which the expected profits to the dealer reflect both the dealer's costs and his bargaining power relative to the customer. Both of these, in turn, can be parametrized as functions of observable variables, and estimated as a Stochastic Frontier Model. The dealer's cost is the stochastic frontier, which represents the expected mark-up the customer would obtain if dealers were always driven by their reservation values, as they would be if the provision of dealer services were perfectly competitive. The observed markup, expressed in excess returns over a municipal bond index, can be written as:

$$\frac{p_i - p^*_i}{p^*_i} - R_{\mathrm{index},i} = \left[\frac{c\left(X_i,\theta\right)}{p^*_i} - E\left(R_{\mathrm{index},i}\middle|\,X_i\right)\right] + \varepsilon_i + \xi_i\,,$$

where p_i is the dealer's selling price, p^*_i is the dealer's purchase price, and $R_{\mathrm{index},i}$ is the municipal bond market index return.

The first term on the right-hand side of the equation represents the dealer's costs in excess of the expected municipal bond index return, where X_i is a set of conditioning variables observable to the buyer and seller and θ is a set of parameters to be estimated. They refer to this term as the cost of intermediation.

The second and third terms capture how the observed markup can differ from the dealer's cost of intermediation.

The second term, ε_i, is a symmetric, normally-distributed error term:

$$\varepsilon_i \equiv \frac{e_i}{p_i^*} - \eta_i \,,$$

reflecting a zero-mean forecast error:

$$\eta_i = R_{\text{index},i} - E\left(R_{\text{index},i} \middle| X_i\right) \,.$$

The third term, ξ_i, is a one-sided, exponentially-distributed error term:

$$\xi \equiv \frac{\rho_i \left[E\left(p_i \middle| X_i\right) - c\left(X_i, \theta\right) - \nu_i\right]}{p_i^*} \,,$$

reflecting the distribution of sellers' reservation values (ν_i) and dealer bargaining power.

Green, Hollifield, and Schurhoff [22] estimate restricted and unrestricted versions of the following regression model via maximum likelihood:

$$\frac{p_i - p_i^*}{p_i^*} - R_{\text{index},i} = \theta_0 + \sum_{l=1}^{L} \theta_l X_{il} + \varepsilon_i + \xi_i \,,$$

with $l = 1, \ldots, L$ conditioning variables. The residual ε_i is normally distributed with standard deviation $b_0 \Pi_{k=1}^{K} \mathrm{e}^{b_{ik} Z_k}$, with Z_{ik} for $k = 1, \ldots, K$ conditioning variables. The residual ξ_i is exponentially distributed with mean and standard deviation $a_0 \Pi_{k=1}^{K} \mathrm{e}^{a_{ik} Z_k}$. In the "market power" version of their model, all of the parameters are unrestricted. In the restricted ("no market power") model, all of the parameters on the one-sided error are constrained to zero: $a_0 = a_1 = \ldots = a_k = 0$.

The data used by Green, Hollifield, and Schurhoff [22] includes both customer trades and interdealer trades. But, because their data does not identify the specific broker-dealer associated with a given trade, they must infer their trades and profits indirectly by studying pairs of trades that appear to be direct exchanges of bonds through a single dealer. They assume that a customer buy transaction of a given size of a given bond that occurs within a very short time of customer sell transaction of the same size in the same bond are most likely related. The reasonableness of this assumption is confirmed by Harris and Piwowar [26], whose data contains dealer identities.

Green, Hollifield, and Schurhoff [22] find that municipal bond dealers earn lower average markups on larger trades, even though larger trades lead the dealers to bear more risk of losses. Their results suggest that municipal bond dealers exercise substantial market power, particularly in retail-sized transactions. Their measures of market power decrease in trade size and increase in variables that indicate the complexity of the trade for the dealer.

Transaction Reporting and Compliance Engine (TRACE) Research

TRACE not only brought unprecedented transparency to corporate bond market investors, it also provided an unprecedented opportunity for market microstructure researchers to examine new issues. Chief among them was the "natural experiment" of adding price transparency to an opaque market. Three prominent studies (collectively, "the TRACE studies") that examined the introduction of price transparency to the corporate bond market were Edwards et al. [17], Bessembinder, Maxwell, and Venkataraman [7], and Goldstein, Hotchkiss, and Sirri [20].

These TRACE studies were very complementary in terms of their contributions to the market microstructure literature. To understand the full impact of this collective research, it is important to remember that they were written at a time when many market participants and some regulators were concerned that public dissemination of bond pricing data might have an adverse impact on liquidity. Using different experimental designs and empirical methods, the TRACE studies produced similar results, conclusions, and implications for regulatory policymakers. Overall, the results in all three TRACE studies show that public investors benefit significantly from the introduction of price transparency.

Edwards et al. [17] estimate transaction costs for all corporate bonds that trade at least nine times between January 2003 and January 2005. Their TRACE data set includes all reported OTC trades in corporate bonds, whether transparent or not. Consistent with the results of Harris and Piwowar [26] for the municipal bond market, Edwards et al. [17] find that corporate bonds are expensive for retail investors to trade and that corporate bond transaction costs decrease significantly with trade size. They find that effective spreads in corporate bonds average 1.24% of the price of representative retail-sized trades ($20,000). They point out that this is the equivalent of over 2 months of the total annual return for a bond with a 6% yield to maturity, or 52 cents per share for a $40 stock. In cross-sectional tests, Edwards et al. [17] find that transaction costs are lower for highly rated bonds, recently issued bonds, and bonds close to maturity.

Edwards et al. [17] find that costs are lower for bonds with transparent trade prices, and they drop when the TRACE system starts to publicly disseminate their prices. Their results suggest that introduction of price transparency results in a drop in customer trading costs of at least 5 basis points (bps). In 2003, public investors traded approximately $2 trillion in bonds for which prices were not disseminated. If the prices for these bonds had been

TRACE-transparent, a quick back-of the-envelope calculation shows investors could have saved a minimum of $1 billion that year. Edwards et al. [17] point out that the $1 billion figure represents a lower bound for two reasons. First, because many unsophisticated investors were unaware that prices became available, and because learning how to use the price data takes, time, the long-run benefits are undoubtedly much greater. Second, they do not capture the initial reduction in trading costs at the initiation of TRACE. Bessembinder, Maxwell, and Venkataraman [7] find that sophisticated institutional investors benefited from an immediate reduction in trading costs of about $1 billion.

Bessembinder, Maxwell, and Venkataraman [7] estimate their trade execution costs for a sample of institutional (insurance company) trades in corporate bonds before and after the initiation of public transaction reporting for some bonds through the TRACE system in July 2002. They find that the average reduction in one-way trading costs or bonds eligible for TRACE transaction reporting is about 5 to 8 bps. This translates to a reduction in trade execution costs of about 50%. Moreover, they find a 20% reduction for bonds not eligible for TRACE reporting. Bessembinder, Maxwell, and Venkataraman [7] interpret their results as suggesting that better pricing information regarding some bonds also improves valuation and execution cost monitoring for related bonds. They find no evidence that market quality deteriorated in other dimensions.

Bessembinder, Maxwell, and Venkataraman [7] also find that larger trading cost reductions for less liquid and lower-rated bonds, and for larger trades. They estimate that their results equate to annual trading cost reductions of roughly $1 billion per year for the entire corporate bond market, reinforcing that market design can have first-order effects, even for relatively sophisticated institutional customers.

Goldstein, Hotchkiss, and Sirri [20] design and construct a controlled experiment to examine the impact of introducing price transparency on liquidity for BBB-rated corporate bonds. They selected the 120 BBB-rated bonds for which the NASD began disseminating trade data on April 14, 2003. They simultaneously selected a control sample of non-disseminated bonds.

Goldstein, Hotchkiss, and Sirri [20] find that DRT spreads decrease for most BBB-rated corporate bonds whose prices become transparent, and that this effect is strongest for intermediate trade sizes. The only caveat to this result is that they do not find any significant transparency effect for the most thinly-traded bonds. Overall, Goldstein, Hotchkiss, and Sirri [20] conclude that their finds indicate that the introduction of post-trade price transparency has a neutral or positive effect on market liquidity.

The similar results and conclusions in the three complementary TRACE studies collectively generate important policy implications. Foremost, policymakers should take comfort in the fact that there are few, if any, instances in the combined results that show any harm to investors from introducing price transparency to securities markets. To the contrary, the results show that both retail and institutional investors benefit from price transparency. The empirical results from the TRACE studies support the well-founded economic theoretical arguments that transparency should lower transaction costs, especially for smaller trades.

Speeches and testimony by US bond market regulators, such as those listed in the bibliography, show that these studies critically informed the debate over adding price transparency to the US bond markets. Moreover, they continue to provide important lessons for policy makers in bond markets outside of the United States. The bibliography also contains a partial listing of international reports, discussion papers, and conference proceedings that prominently cite the TRACE studies.

Edwards, Harris, and Piwowar (2007)

Edwards et al. [17] apply the Harris and Piwowar [26] econometric approach to corporate bonds. They also extend the approach by allowing liquidity to be time varying. This extension allows them to examine how the introduction of price transparency affects corporate bond transaction costs.

They model the unobserved value return r^{V}_{ts} by decomposing it into the linear sum of a time drift, an average bond index return, differences between index returns for long and short term bonds and for high and low quality bonds, and a bond-specific valuation factor, ε_{ts}.

$$\begin{aligned} r^{V}_{ts} = {} & Days_{ts}\,(DriftRate) + \beta_1 AveIndexRet_{ts} \\ & + \beta_2 DurationDif_{ts} + \beta_3 CreditDif_{ts} + \varepsilon_{ts}\,, \end{aligned}$$

where $Days_{ts}$ counts the number of calendar days between trades t and s, $DriftRate$ is the bond coupon rate subtracted from five percent, $AveIndexRet_{ts}$ is the index return for the average bond between trades t and s and $DurationDif_{ts}$ and $CreditDif_{ts}$ are the corresponding differences between index returns for long and short term bonds and high and low credit risk bonds. The first term accounts for the continuously compounded bond price return that traders expect when interest rates are constant and the bond's coupon interest rate differs from five percent. The three factor returns account for bond value changes due to shifts in interest rates and credit spreads. Edwards et al. [17] estimate the bond indices using repeat sale regression methods with terms that account for bond transaction costs. Finally, the bond-specific valuation factor ε_{ts} has mean zero and variance given by

$$\sigma^2_{\varepsilon_{ts}} = N^{\text{Sessions}}_{ts} \sigma^2_{\text{Sessions}} \,,$$

where N^{Sessions}_{ts} is the number of trading sessions and fractions of trading sessions between trades t and s.

Edwards et al. [17] model customer transaction costs using the following additive expression:

$$c(S_t) = c_0 + c_1 \frac{1}{S_t} + c_2 \log S_t + c_3 S_t + c_4 S_t^2 + \kappa_t \,,$$

where κ_t represents variation in the actual customer transaction cost that is unexplained by the average transaction cost function. This variation may be random or due to an inability of the average transaction cost function to represent average trade costs for all trade sizes. They assume κ_t has zero mean and variance given by σ^2_κ.

The first three terms of the cost function are the same as in Harris and Piwowar [26], where the constant term allows total transaction costs to grow in proportion to size, the second term characterizes any fixed costs per trade, and the third term allows for costs per bond to vary by trade size. The two additional terms allow more flexibility in the costs to vary by size. Because corporate bonds trade more frequently than municipal bonds, Edwards et al. [17] did not need to be as concerned about degrees of freedom as Harris and Piwowar [26].

Combining the last three equations produces the Edwards et al. [17] version of the Harris and Piwowar [26] transaction cost estimation model:

$$\begin{aligned} r^{\text{P}}_{ts} - \mathit{Days}_{ts}\left(\mathit{DriftRate}\right) &= c_0 \left(Q_t - Q_s\right) \\ &+ c_1 \left(Q_t \frac{1}{S_t} - Q_s \frac{1}{S_s}\right) + c_2 \left(Q_t \log S_t - Q_s \log S_s\right) \\ &+ c_3 \left(Q_t S_t - Q_s S_s\right) + c_4 \left(Q_t S_t^2 - Q_s S_s^2\right) \\ &+ \beta_1 \mathit{AveIndexRet}_{ts} + \beta_2 \mathit{DurationDif}_{ts} + \beta_3 \mathit{CreditDif}_{ts} \\ &+ \eta_{ts} \,, \end{aligned}$$

where the left hand side is simply the continuously compounded bond return expressed as the equivalent rate on a notional five percent coupon bond. Edwards et al. [17] estimate their time-series model in the same way as Harris and Piwowar [26].

Edwards et al. [17] extend Harris and Piwowar [26] by introducing a pooled time-series regression model that they use to estimate average transaction costs for each day for a class of bonds. With this model, they are able to estimate the daily average transaction costs for bonds that became transparent in 2003, and compare these estimates to those for comparable bonds that were either TRACE-transparent throughout 2003 or never TRACE-transparent in 2003.

The pooled time-series regression model that Edwards et al. [17] use to estimate daily transaction costs differs in two respects from the time-series regression model that they use to estimate average transaction costs for a given bond. First, they specify separate average transaction cost functions, $c_T(S_t)$, for each day T in the sample. Second, to minimize the total number of parameters to be estimated, they use the three-parameter average cost function:

$$c_T(S_t) = c_{0T} + c_{1T} \frac{1}{S_t} + c_{2T} \log S_t + \kappa_t \,.$$

For a given bond, the change in value between bond trades is modeled as:

$$\log V_t - \log V_s = f_s r_S + \sum_{J=S+1}^{T-1} r_J + f_t r_T + e_{st} \,,$$

where S is the day on which trade s took place and T is the day on which a subsequent trade t took place, r_J is the common index return (to be estimated) for day J and f_s and f_t, respectively, are the fractions of the S and T trading days overlapped by the period spanned by transactions s and t. This portion of the specification is the same as appears in many paired trade regression index estimation procedures. With these changes, the regression model is

$$\begin{aligned} r^{\text{P}}_{ts} - \mathit{Days}_{ts}\left(5\% - \mathit{CouponRate}\right) &= c_{0T} Q_t - c_{0S} Q_s \\ + c_{1T} Q_t \frac{1}{S_t} - c_{1S} Q_s \frac{1}{S_s} &+ c_{2T} Q_t \log S_t - c_{2S} Q_s \log S_s \\ &+ f_s r_S + \sum_{J=S+1}^{T-1} r_J + f_t r_T + \eta_{ts} \,. \end{aligned}$$

They use iterated weighted least squares methods, where the weights are equal to the predicted values of the regression of the squared residuals on the independent variables appearing in the residual variation expression. Edwards et al. [17] estimate the model using a three-month wide sliding window that moves forward one month at a time. The time series of coefficient estimates are assembled from the center months of each of the sliding regressions. They compute transaction costs for various transaction sizes by evaluating the estimated transaction cost functions at the given transaction sizes. Using the estimated variance-covariance matrix of the estimators, they also compute daily standard errors of the various daily transaction cost estimates.

Bessembinder, Maxwell, and Venkataraman (2006)

Bessembinder, Maxwell, and Venkataraman [7] develop and estimate an indicator variable regression model:

$$\Delta P = a + wX_t + \gamma S Q_t^* + \alpha S \Delta Q + \omega_t \,,$$

where ΔP is the change in the price of the bond from time $t-1$ to time t, a is the regression intercept, w is the fraction of public information that is observable in the data with realizations X_t, γS is the informational component of the spread, αS is the non-informational component of the spread (where $\alpha = (1-\gamma)$), Q_t^* is the market maker's estimate of bond value due to surprises in order flow, ΔQ is the change in indicator variable Q (which takes a value of 1 if the trade is a customer buy and -1 if it is a customer sell) from time $t-1$ to time t, and ω_t is the regression error term.

Bessembinder, Maxwell, and Venkataraman [7] develop this regression model in the following way. $E_t(V)$ is the market-maker's estimate of the bond's unobserved true value (V) at time t conditional on whether the observed trade is a customer buy or a customer sell. Transaction prices are given by:

$$P_t = E_t(V) + \alpha S Q_t \,.$$

The market maker's estimate of bond value at time t, $E_t(V)$, is her estimate from the prior period, $E_{t-1}(V)$, updated to reflect surprises in order flow, $Q_t - E_{t-1}(Q_t)$, and public information revealed since the prior period, η_τ Substitution yields:

$$E_t(V) = E_{t-1}(V) + \gamma S Q_t^* + \eta_t \,,$$

where $Q_t^* = Q_t - E_{t-1}(Q_t)$. To allow for the possibility that bond market order flow is positively autocorrelated, Bessembinder, Maxwell, and Venkataraman [7] assume that it follows a simple AR1 process, so that $E_{t-1}(Q_t) = \rho(Q_{t-1})$. The change in the price of the bond from time $t-1$ to time t is:

$$P_t - P_{t-1} = \gamma S Q_t^* + \alpha S Q_t - \alpha S Q_{t-1} + \eta_t \,.$$

Substituting ΔP for P_t – P_{t-1} and ΔQ for Q_t – Q_{t-1}, this expression can be rewritten as:

$$\Delta P = \gamma S Q_t^* + \alpha S \Delta Q + \eta_t \,.$$

To incorporate observable public information that affects bond value, they assume that a fraction w of public information becomes observable in the data with realizations X_t, while the remaining portion $(1-w)$ is due to unobservable innovations U_t that represent statistical noise. Substitution yields their regression model:

$$\Delta P = wX_t + \gamma S Q_t^* + \alpha S \Delta Q + \omega_t \,,$$

where $\omega = (1-w)\,U_t$. Bessembinder, Maxwell, and Venkataraman [7] show that their model is equivalent to the Madhavan et al. [39] model. Moreover, in the special case of no autocorrelation in order flow ($\rho = 0$), their model is equivalent to Huang and Stoll [33], Schultz [42] and Warga [44].

Goldstein, Hotchkiss, and Sirri (2007)

Goldstein, Hotchkiss, and Sirri [20] use two different methods to estimate transaction costs for a sample of BBB-rated bonds. Their first method involves identifying “dealer round-trip” (DRT) transaction chains. These transaction chains involve a dealer purchasing a bond from a customer and then selling that same bond to another customer within a specified period of time. DRT spreads are calculated as the difference between the customer buy price at the end of the transaction chain and the customer sell price at the beginning of the chain. Their DRT method is similar to the methods used in the municipal bond studies of Green et al. [22] and Biais and Green [9], except their data contains individual dealer identifiers. Goldstein, Hotchkiss, and Sirri [20] estimate DRT spreads for transaction chains that occur with one-day, five-day, and unlimited time intervals. They estimate DRT spreads for various trade size groups.

Their second method is a regression method similar to Warga [44] and Schultz [42]. For each trade size group, Goldstein, Hotchkiss, and Sirri [20] estimate spreads by regressing the difference between the transaction price for a customer (T) and an estimated bid price (B) on a dummy variable that equals one for customer buys and zero for customer sells:

$$[T - B]_i = \alpha_0 + \alpha_1 D_i^{\text{Buy}} + \varepsilon_i \,,$$

where estimated bid prices are obtained from Reuters dealer bid price estimates from the end of the day prior to the transaction. Reuters estimates are based on daily quotes obtained directly from individual dealers.

Goldstein, Hotchkiss, and Sirri [20] estimate a second regression to consider the effect of dissemination while controlling for other bond characteristics impacting the spread:

$$
\begin{aligned}
[T-B]_i &= \alpha_0 + \alpha_1 D_i^{\text{Buy}} + \alpha_2 D_i^{\text{DisseminatedBond}} \\
&+ \alpha_3 D_i^{\text{Post-disseminationPeriod}} \\
&+ \alpha_4 D_i^{\text{DisseminationBond}^{*}\,\text{Post-disseminationPeriod}} \\
&\qquad + \alpha_5 X_5 + \cdots + \alpha_{10} X_{10} + \varepsilon_i \ ,
\end{aligned}
$$

where additional dummies are included for disseminated bonds, transactions that occur in the post-dissemination period, and the interaction of these two dummies for transaction in disseminated bonds that occur in the post-dissemination period. As in Schultz [42] the additional dummies are expressed a $+1$ for buys and -1 for sells. Variables $X_5, \ldots, X_{10}$ are six bond characteristics related to spreads: trade size, time-to-maturity, age, issue amount, average daily volume over the prior 30 days, and days since last trade.

The Links Between Bond Market Microstructure Research and Other Finance and Economics Research

The discussion of bond market research has thus far been presented solely within the framework of the market microstructure literature. However, some of the most interesting bond market research is connected to other areas of finance and economics. The instrument complexity results of Harris and Piwowar [26], for example, provide evidence to support Carlin's [11] formal model of strategic price complexity in security design for retail markets. Additionally, Chen, Lesmond, and Wei [13] provide an example of how bond market microstructure research is linked to asset pricing models in finance. Green, Hollifield, and Schurhoff [23] develop a theoretical model that is analogous to the costly consumer search models in the broader economics literature.

Chen, Lesmond, and Wei (2007)

Beginning with Amihud and Mendelson [3], market microstructure research has consistently shown that a liquidity premium exists in equity markets. Recently, bond market microstructure researchers have begun investigating whether a liquidity premium also exists in bond markets. One such paper is Chen, Lesmond, and Wei [13]. Their investigation of the link between bond liquidity and corporate yield spreads provides important implications for the default risk literature.

Chen, Lesmond, and Wei [13] investigate whether liquidity is priced in corporate yield spreads. They use several approaches, including a regression approach that is an extension to the Lesmond, Ogden, and Trzcinka [36] (LOT) approach developed for equities. The LOT approach assumes that a zero return day (or a non-trading day) is observed when the true price changes by less than the transaction costs. Because marginal informed investors will only trade on information if the trade yields expected profits net of transaction costs, an individual bond's trading costs represent a threshold that must be exceeded before its return will reflect new information. The premise of this approach is that if the value of the information is insufficient to exceed the costs of trading, then the marginal investor will either reduce trading or not trade, causing a zero return.

Chen, Lesmond, and Wei [13] extend the LOT approach to corporate bonds by applying a two-factor return-generating model to estimate bond trading costs:

$$
\begin{aligned}
R^{*}_{j,t} &= \beta_{j1} \mathit{Duration}^{*}_{j,t} \Delta R_{ft} \\
&\qquad + \beta_{j2} \mathit{Duration}^{*}_{j,t} \Delta \mathit{S\&P\ Index}_t + \varepsilon_{j,t} \ ,
\end{aligned}
$$

where $R^{*}_{j,t}$ is the unobserved "true" return for bond j on day t that investors would bid given zero trading costs. The daily change in the 10-year risk-free interest rate, ΔR_{ft}, is the factor that is more important for investment grade bonds, while the second factor, $\Delta \mathit{S\&P\ Index}_t$, the daily return on the Standard & Poor's 500 equity index, is more important for high-yield bonds. Both factors are scaled by $\mathit{Duration}_{j,t}$, the bond's duration.

Chen, Lesmond, and Wei [13] then apply the Amihud and Mendelson [3] liquidity premium to bonds. In Amihud and Mendelson [3], observed asset prices differ from true values because of a liquidity premium that compensates investors for liquidity costs. Chen, Lesmond, and Wei [13] state the liquidity effects on bond returns as:

$$
R_{j,t} = R^{*}_{j,t} - \alpha_{i,j} \ ,
$$

where $R_{j,t}$ is the measured return, $\alpha_{2,j}$ is the effective buy-side cost, and $\alpha_{1,j}$ is the effective sell-side cost for bond j. The resulting liquidity constraint is:

$$
\begin{array}{lll}
R_{j,t} = R^{*}_{j,t} - \alpha_{1j} & \text{if } R^{*}_{j,t} < \alpha_{1j} & \text{and } \alpha_{1j} < 0 \\
R_{j,t} = 0 & \text{if } \alpha_{1j} \le R^{*}_{j,t} \le \alpha_{2j} & \\
R_{j,t} = R^{*}_{j,t} - \alpha_{2j} & \text{if } R^{*}_{j,t} > \alpha_{2j} & \text{and } \alpha_{2j} > 0 \, .
\end{array}
$$

Combining the liquidity constraint with the return generating model, Chen, Lesmond, and Wei [13] us a maximum likelihood method outlined in LOT to estimate transaction costs. They specify the log-likelihood function as:

$$\begin{aligned}\mathrm{Ln}L = {} & \sum_{1} \mathrm{Ln}\frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \\ & - \sum_{1} \frac{1}{2\sigma_j^2}\left(R_j + \alpha_{1,j} - \beta_{j1}\mathit{Duration}_{j,t} * \Delta R_{ft}\right. \\ & \qquad \left. - \beta_{j2}\mathit{Duration}_{j,t} * \Delta \mathit{S\&P\ Index}\right)^2 \\ & + \sum_{2} \mathrm{Ln}\frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \\ & + \sum_{2} \frac{1}{2\sigma_j^2}\left(R_j + \alpha_{2,j} - \beta_{j1}\mathit{Duration}_{j,t} * \Delta R_{ft}\right. \\ & \qquad \left. - \beta_{j2}\mathit{Duration}_{j,t} * \Delta \mathit{S\&P\ Index}\right)^2 \\ & + \sum_{0} \mathrm{Ln}\left(\Phi_{2,j} - \Phi_{1,j}\right) ,\end{aligned}$$

where $\Phi_{i,j}$ represents the cumulative distribution function for each bond-year evaluated at:

$$\frac{\left(\alpha_{1,j} - \beta_{j1}\mathit{Duration}_{j,t} * \Delta R_{ft}\right)}{\sigma_j} - \frac{\left(\beta_{j2}\mathit{Duration}_{j,t} * \Delta \mathit{S\&P\ Index}_t\right)}{\sigma_j} .$$

Σ_1 (region 1) represents the negative nonzero measured returns, Σ_2 (region 2) represents the positive nonzero measured returns, and Σ_0 (region 0) represents the zero measured returns. The difference in the buy-side and sell-side cost estimates, $\alpha_{2,j} - \alpha_{1,j}$, represents round-trip trading costs.

The model's implicit assumption that information motivates bond trades and that information is efficiently impounded into bond prices is supported by the results of Hotchkiss and Ronen [31]. The error term captures noise trading and trades due to unanticipated public information.

In addition to LOT estimates, Chen, Lesmond, and Wei [13] use bid-ask spreads and zero-returns as liquidity cost measures. The bid-ask spreads the use are bond-year proportional bid-ask spreads, calculated as the average of quarterly proportional spreads. Quarterly proportional spreads are calculated from quarterly bid-ask spreads obtained from Bloomberg consensus quotes among market participants, divided by the average bid and ask price. Zero-returns are simply the percentage of days with returns equal to zero.

They find that liquidity costs are related to credit rating. Liquidity costs are much higher for high-yield bonds than for investment grade bonds. They also find that liquidity costs are related to maturity. Liquidity costs for long-maturity bonds are higher than for short-maturity bonds.

They also find that yield spreads generally increase with maturity for investment grade bonds. But, they find that yield spreads generally decrease with maturity for high-yield bonds. They point out the endogeneity issue stemming from the Helwege and Turner [28] finding that relatively safer firms within the same high-yield credit rating category tend to issue longer-term bonds. This endogeneity issue causes the average yield spread to decline with maturity for high-yield bonds.

To investigate whether liquidity is priced in corporate yield spreads, Chen, Lesmond, and Wei [13] first run the following regression specification for investment grade bonds and high-yield bonds separately:

$$\begin{aligned}\mathit{Yield\ Spread}_{it} = {} & \eta_0 + \eta_1\mathit{Liquidity}_{it} + \eta_2\mathit{Maturity}_{it} \\ & + \eta_3\mathit{Amount\ Outstanding}_{it} + \eta_4\mathit{Coupon}_{it} \\ & + \eta_5\mathit{Treasury\ Rate}_{it} + \eta_6 10\mathit{Yr} - 2\mathit{Yr\ Treasury\ Rate}_{it} \\ & + \eta_7\mathit{EuroDollar}_{it} + \eta_8\mathit{Volatility}_{it} + \eta_9\mathit{Bond\ Rating}_{it} \\ & + \eta_{10}\mathit{PreTax\ Coverage\ Dummy}_{it} \\ & + \eta_{11}\mathit{Operating\ Income/Sales}_{it} + \eta_{12}\mathit{Debt/Assets}_{it} \\ & + \eta_{13}\mathit{Debt/Capitalization}_{it} + \varepsilon_{it} ,\end{aligned}$$

where the subscript *it* refers to bond *i* in year *t*. *Liquidity* refers to the three liquidity cost measures – bid-ask spread, zero-returns, or the LOT estimate. Additional variables control for bond-specific, firm-specific, and macroeconomic factors. *Maturity* is the number of years until the bond matures relative to the year being analyzed, and *Amount Outstanding* is natural logarithm of the dollar amount outstanding, *Coupon* is the bond coupon rate. *Treasury Rate* is the 1-year Treasury Note rate, *10Yr-2Yr Treasury Rate* is the difference between the 10-year and 2-year Treasury rates, and *Eurodollar* is the 30-day Eurodollar rate minus the 3-month T-Bill rate. *Volatility* is the equity volatility for each issuer and *Bond Rating* is a credit rating scale that ranges from 1 (AAA rating) to 10 (BBB- rating) for investment grade bonds and from 1(BB+ rating) to 12 (D rating) for high-yield bonds. *Pre-Tax Coverage Dummy* represents four dummy variables corresponding to groupings of pre-tax income, *Operating Income/Sales*, *Debt/Assets*, and *Debt/Capitalization* are each firm's respective accounting ratios.

They find that all three liquidity measures are positively related to the yield spread in both the investment grade and high-yield samples. The liquidity coefficients are statistically significant at the 1% level in every scenario. This provides strong evidence that liquidity is priced in corporate yield spreads. This finding is robust to controlling for issuer influences with issuer fixed-effects regressions. The only caveat is that they achieve slightly weaker results for the zero-return liquidity cost measure than for bid-ask spreads and LOT estimates. This finding is also robust to controlling for potential endogeneity problems arising from the contemporaneous measurement of the yield spread, liquidity costs, and credit rating. They perform this robustness check by employing a simultaneous equations model using three equations that correspond to each of the potentially endogenous variables:

$$\begin{aligned}&\mathit{Yield\ Spread}_{it} = \eta_0 + \eta_1 \mathit{Liquidity}_{it} + \eta_2 \mathit{Maturity}_{it}\\ &+ \eta_3 \mathit{Coupon}_{it} + \eta_4 \mathit{Treasury\ Rate}_{it}\\ &+ \eta_5 10\mathit{Yr} - 2\mathit{Yr\ Treasury\ Rate}_{it} + \eta_6 \mathit{EuroDollar}_{it}\\ &+ \eta_7 \mathit{Volatility}_{it} + \eta_8 \mathit{Bond\ Rating}_{it}\\ &+ \eta_9 \mathit{PreTax\ Coverage\ Dummy}_{it}\\ &+ \eta_{10} \mathit{Operating\ Income/Sales}_{it} + \eta_{11} \mathit{Debt/Assets}_{it}\\ &\qquad\qquad + \eta_{12} \mathit{Debt/Capitalization}_{it} + \varepsilon_{it}\ ,\end{aligned}$$

$$\begin{aligned}&\mathit{Liquidity}_{it} = \eta_0 + \eta_1 \mathit{Maturity}_{it} + \eta_2 \mathit{Age}_{it}\\ &+ \eta_3 \mathit{Amount\ Outstanding}_{it} + \eta_4 \mathit{Bond\ Rating}_{it}\\ &\quad + \eta_5 \mathit{Bond\ Volatility}_{it} + \eta_6 \mathit{Yield\ Spread}_{it} + \varepsilon_{it}\ ,\end{aligned}$$

$$\begin{aligned}&\mathit{Credit\ Rating}_{it} = \eta_0 + \eta_1 \mathit{Treasury\ Rate}_{it}\\ &+ \eta_2 10\mathit{Yr} - 2\mathit{Yr\ Treasury\ Rate}_{it}\\ &+ \eta_3 \mathit{PreTax\ Coverage\ Dummy}_{it}\\ &+ \eta_4 \mathit{Operating\ Income/Sales}_{it}\\ &+ \eta_5 \mathit{Debt/Assets}_{it} + \eta_6 \mathit{Debt/Capitalization}_{it}\\ &\qquad\qquad + \eta_7 \mathit{Yield\ Sprad}_{it} + \varepsilon\ .\end{aligned}$$

The model is estimated using twostage least squares. The simultaneous equation model estimation results show that the potential endogeneity does not affect the relation between liquidity and yield spreads for either the investment grade or the high-yield bonds.

Thus, Chen, Lesmond, and Wei [13] find extremely consistent and robust evidence that liquidity is a key determinant in corporate yield spreads. This finding provides at least a partial explanation for the findings of Collin-Dufresne, Goldstein, and Martin [15] and others who show that default risk does not completely explain corporate yield spreads.

Green, Hollifield, and Schurhoff (2007b)

Green, Hollifield, and Schurhoff [23] examine secondary market trading in newly issued municipal bonds for the first 60 trading days of their lives. They begin by descriptively documenting the price behavior of newly issued municipal bonds. They show that municipal bonds are underpriced when issued. But, unlike equities, the average price rises slowly over a period of several days. Green, Hollifield, and Schurhoff [23] also find that the observed price patterns are complex. High levels of price dispersion are observed for small trade sizes in the aftermarket for new municipal bond issues. While some small traders purchase bonds on attractive terms, others do not. In contrast, there is very little price dispersion for large trade sizes. Virtually all of the large traders purchase bonds on attractive terms.

They argue that the price level and dispersion patterns are the result of bond dealers discriminating between informed and uninformed customers. Accordingly, Green, Hollifield, and Schurhoff [23] develop and estimate a mixed distribution model for the markups that uninformed and informed investors pay when they purchase newly issued bonds. Their model incorporates investor search costs, i. e., the costs in terms of time and effort needed for investors to become informed about new bond issues.

The mixed distribution model of Green, Hollifield, and Schurhoff [23] is analogous to economic models of costly consumer search, such as the gametheoretic model of Shilony [43] that focuses on advertising and price competition among retail stores in homogeneous product markets. Shilony [43] assumes that all stores are required to advertise, but the advertising is segmented (e. g., signs are posted on store windows). Consumers have a preference for the particular store that that offers them free access to the advertising (e. g., the store right outside their house or the one that they regularly visit) and they will pay more for a product at this store even if it does not offer the lowest price.

The institutional mechanisms of the primary market and the structure of the secondary market for municipal bonds fits particularly well with the informational interpretation of Shilony's [43] model. Every investor has free information about the price that will be charged by his broker. Also, because all firms must disseminate their last sale information on a real-time basis, the investor can choose to look on www.investinginbonds.com or some other free website to find the range of prices charged by all brokers. But, this last sale information does not identify which broker charged the lowest price. To find this out, the investor must incur some cost.

Green, Hollifield, and Schurhoff [23] begin with the assumption that there are both observable and unobservable sources of heterogeneity in the costs investors face in gathering and using information about prices of new municipal issues. They assume that for investor i, the difference between the benefit and the cost of learning about a new issue is z_i^* with:

$$z_i^* = w_i\delta + \mu_i \ ,$$

where w_i is a vector of conditioning variables, δ is a parameter vector, and μ_ι is an error term. The error term is observed by the investor, but not by the econometrician. Investor i becomes informed about the price of a new issue if and only if $z_i^* \geq 0$. They do not observe z_i^*, but they do observe w_i and the price the investor pays for the bond.

An investor who is uninformed about the reoffering price for a new bond is willing to pay the percentage markup y_U of:

$$y_{Ui} = x_i\beta + \varepsilon_{Ui} \ ,$$

where x_i is a vector of conditioning variables, β is a parameter vector, and ε_{Ui} is an error term. Similarly, an investor who is informed about the underwriter's pricing of a new bond is willing to pay the percentage markup y_I of:

$$y_{Ii} = x_i\gamma + \varepsilon_{Ii} \ ,$$

where x_i is a vector of conditioning variables, γ is a parameter vector, and ε_{Ii} is an error term. The uncertainty about the percentage markup is expected to be lower when the investor is informed than when the investor is uninformed:

$$\sigma_I < \sigma_U \ .$$

They use this condition to empirically identify the informed versus uninformed distributions from which the observed markups, y_i, are drawn:

$$y_i = \begin{cases} y_{Ui} & \text{if} \quad z_i^* < 0 \ , \\ y_{Ii} & \text{if} \quad z_i^* \geq 0 \ . \end{cases}$$

Green, Hollifield, and Schurhoff [23] use iterated expectations to show that investors take the markup into account when deciding whether to become informed about an upcoming bond issue or not:

$$\begin{aligned} E\left(y_i\,|w_i, x_i\right) = {} & E\left(y_i\,|\text{Informed}_i, w_i, x_i\right) \\ & \Pr\left(\text{Informed}_i\,|w_i\right) \\ & + E\left(y_i\,|\text{Uninformed}_i, w_i, x_i\right) \Pr\left(\text{Uninformed}_i\,|w_i\right) . \end{aligned}$$

They estimate their model under the assumption that the error terms are drawn independently and identically from a multivariate normal distribution:

$$\begin{pmatrix} u_i \\ \varepsilon_{Ui} \\ \varepsilon_{Ii} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_U\sigma_U & \rho_I\sigma_I \\ \rho_U\sigma_U & \sigma_U^2 & 0 \\ \rho_I\sigma_I & 0 & \sigma_I^2 \end{bmatrix} \right),$$

where ρ_U is the correlation between u_i and ε_{Ui} and ρ_I is the correlation between u_i and ε_{Ii}. Denoting the cumulative standard normal distribution as Φ and the standard normal density as φ, Green, Hollifield, and Schurhoff [23] show that the condition that investor i becomes informed if and only if $z_i^* \geq 0$ implies that:

$$\begin{aligned} \Pr\left(\text{Informed}_i\,|w_i\right) &= \Pr\left(z_i^* \geq 0\,|w_i\right) \\ &= \Pr\left(u_i \geq -w_i\delta\,|w_i\right) \\ &= \Phi\left(w_i\delta\right) \ , \end{aligned}$$

and

$$\Pr\left(\text{Uninformed}_i\,|w_i\right) = 1 - \Phi\left(w_i\delta\right) \ .$$

By combining equations and using the distributional assumptions of the error terms, Green, Hollifield, and Schurhoff [23] show that

$$E\left(y_i\,|\text{Informed}_i, w_i, x_i\right) = x_i\gamma + \rho_I\sigma_I\frac{\phi\left(w_i\delta\right)}{\Phi\left(w_i\delta\right)} \ ,$$

and

$$E\left(y_i\,|\text{Uninformed}_i, w_i, x_i\right) = x_i\beta + \rho_U\sigma_U\frac{-\phi\left(w_i\delta\right)}{1-\Phi\left(w_i\delta\right)} .$$

Therefore, the expected markup is:

$$\begin{aligned} E\left(y_i\,|w_i, x_i\right) = {} & \left(x_i\gamma + \rho_I\sigma_I\frac{\phi\left(w_i\delta\right)}{\Phi\left(w_i\delta\right)}\right)\Phi\left(w_i\delta\right) \\ & + \left(x_i\beta + \rho_U\sigma_U\frac{-\phi\left(w_i\delta\right)}{1-\Phi\left(w_i\delta\right)}\right)1 - \Phi\left(w_i\delta\right) \ . \end{aligned}$$

Green, Hollifield, and Schurhoff [23] estimate this equation as a switching regression. Their results are consistent with two pricing regimes for newly issued municipal bonds. Uninformed investors pay higher average prices than informed investors and there is very little variation in prices paid by informed investors. They also find that the upward trend in prices after issuance is related to the change in the mix if informed and uninformed investors. Informed investors buy soon after issuance, while uninformed investors buy later on.

With respect to the decision about whether to become informed about a new municipal bond issue, they find that large buyers are more likely to become informed than small buyers. To examine how much money is left on the table by uninformed investors who pay large markups to dealers, Green, Hollifield, and Schurhoff [23] classify each transaction into either the Informed or Uninformed regime. The classification is based on whether the expected benefit from becoming informed about a new bond issue is greater than the cost of doing so:

$$\begin{aligned}\text{Informed}_i = 1 &\Leftrightarrow E\left(z_i^* \,|y_i, w_i, x_i\right) \geq 0\,,\\ \text{Uninformed}_i = 1 &\Leftrightarrow E\left(z_i^* \,|y_i, w_i, x_i\right) < 0\,.\end{aligned}$$

The difference in the expected markup between an informed investor and an uninformed investor is:

$$\begin{aligned}E\left(y_{Ui}y_{Ii}\,|\text{Uninformed}_i, w_i, x_i\right) &= x_i\left(\beta - \gamma\right)\\ &\quad + \left(\rho_U\sigma_U - \rho_I\sigma_I\right)\frac{-\phi\left(w_i\delta\right)}{1-\Phi\left(w_i\delta\right)}\,.\end{aligned}$$

They define the money left on the table in each transaction with an uninformed investor as

$$\Delta_i = \begin{cases}\max\left\{E\left(y_{Ui}y_{Ii}\,|\text{Uninformed}_i, w_i, x_i\right), 0\right\}, & \\ & \text{if Uninformed}_i = 1\,,\\ 0\,, & \text{else}\,.\end{cases}$$

They denote the estimates of Δ_i as $\widehat{\Delta}_i$ and obtain a cumulative measure across all sales transactions i in a given bond issue j, and then across all issues in a bond deal:

$$\widehat{\text{Money Left on the Table}} = \sum_{\text{Issues}\,j}\,\sum_{i\in j}\widehat{\Delta}_i\,.$$

Green, Hollifield, and Schurhoff [23] find that money left on the table by uninformed investors represents a significant fraction of the overall expected profits to the underwriters and dealers.

Bibliography

Primary Literature

1. Alexander G, Edwards A, Ferri M (2000a) The determinants of trading volume of high-yield corporate bonds. J Financ Mark 3:177–204
2. Alexander G, Edwards A, Ferri M (2000b) What does Nasdaq's high-yield bond market reveal about bondholderstockholder conflicts? Financ Manag 29:23–39
3. Amihud Y, Mendelson H (1986) Asset pricing and the bid-ask spread. J Financ Econ 17:223–249
4. Amihud Y, Mendelson H (1991) Liquidity, maturity, and yields on US Treasury securities. J Financ 46:1411–1425
5. Barclay M, Hendershott T, Kotz K (2006) Automation versus intermediation: Evidence from Treasuries going off the run. J Financ 61:2395–2414
6. Bernhardt D, Dvoracek V, Hughson E, Werner I (2005) Why do larger orders receive discounts on the London Stock Exchange? Rev Financ Stud 18:1343–1368
7. Bessembinder H, Maxwell W, Venkataraman K (2006) Optimal market transparency: Evidence from the initiation of trade reporting in corporate bonds. J Financ Econ 82:251–288
8. Biais B, Glosten L, Spatt C (2005) Market microstructure: A survey of microfoundations, empirical results, and policy implications. J Financ Mark 8:217–264
9. Biais B, Green R (2005) The microstructure of the bond market in the 20th century. Working paper, Carnegie Mellon University
10. Blume M, Keim D, Patel S (1991) Returns and volatility of low-grade bonds. J Financ 46:49–74
11. Carlin B (2007) Strategic price complexity in retail financial markets. Working paper, UCLA
12. Chakravarty S, Sarkar A (2003) Trading costs in three US bond markets. J Fixed Income 13:39–48
13. Chen L, Lesmond D, Wei J (2007) Corporate yield spreads and bond liquidity. J Financ 52:119–149
14. Cornell B, Green K (1991) The investment performance of low-grade bond funds. J Financ 46:29–48
15. Collin-Dufresne P, Goldstein R, Martin S (2001) The determinants of credit spread changes. J Financ 41:2177–2207
16. Downing C, Zhang F (2004) Trading activity and price volatility in the municipal bond market. J Financ 59:899–931
17. Edwards A, Harris L, Piwowar M (2007) Corporate Bond Market Transaction Costs and Transparency. J Financ 62:1421–1451
18. Elton E, Green C (1998) Tax and liquidity effects in pricing government bonds. J Financ 53:1533–1562
19. Fenn G (2000) Speed of issuance and the adequacy of disclosure in the 144a high-yield debt market. J Financ Econ 56:383–405
20. Goldstein M, Hotchkiss E, Sirri E (2007) Transparency and liquidity: A controlled experiment on corporate bonds. Rev Financ Stud 20:235–273
21. Green R (2007) Issuers, underwriter syndicates, and aftermarket transparency. Presidential Address to the American Finance Association
22. Green R, Hollifield B, Schurhoff N (2007a) Financial intermediation and the costs of trading in an opaque market. Rev Financ Stud 20:275–314
23. Green R, Hollifield B, Schurhoff N (2007b) Dealer intermediation and price behavior in the aftermarket for new bond issues. J Financ Econ 86:643–682
24. Harris L (2003) Trading and exchanges: market microstructure for practitioners. Oxford University Press, New York
25. Harris L, Piwowar M (2004) Municipal bond liquidity. Working paper, Available at SSRN: http://ssrn.com/abstract=503062
26. Harris L, Piwowar M (2006) Secondary trading costs in the municipal bond market. J Financ 61:1361–1397
27. Hasbrouck J (1993) Assessing the quality of a security market: A new approach to transaction-cost measurement. Rev Financ Stud 6:191–212
28. Helwege J, Turner C (1999) The slope of the credit yield curve for speculative-grade issuers. J Financ 54:1869–1884
29. Hong G, Warga A (2000) An empirical study of bond market transactions. Financ Anal J 56:32–46
30. Hong G, Warga A (2004) Municipal marketability. J Fixed Income 14:86–95
31. Hotchkiss E, Ronen T (2002) The informational efficiency of the corporate bond market: An intraday analysis. Rev Financ Stud 15:1325–1354
32. Hotchkiss E, Warga A, Jostova G (2002) Determinants of corporate bond trading: A comprehensive analysis. Working paper, Boston College

33. Huang R, Stoll H (1997) The components of the bid-ask spread: A general approach. R Financ Stud 10:995–1034
34. Kalimipalli M, Warga A (2002) Bid/ask spread, volatility and volume in the corporate bond market. J Fixed Income 12:31–42
35. Karolyi GA (2004) The world of cross-listings and cross-listings of the world: Challenging conventional wisdom. Working paper, Ohio State University
36. Lesmond D, Ogden J, Trzcinka C (1999) A new estimate of transaction costs. Rev Financ Stud 12:1113–1141
37. Levitt A (1998) The importance of transparency in America's debt market. Remarks at the Media Studies Center, New York, September 9, 1998
38. Madhavan A (2000) Market microstructure: A survey. J Financ Mark 3:205–208
39. Madhavan A, Richardson M, Roomans M (1997) Why do security prices change? A transaction-level analysis of NYSE stocks. R Financ Stud 10:1035–1064
40. Reiss P, Werner I (1996) Transaction costs in dealer markets: Evidence from the London Stock Exchange. In: Lo A (ed) The industrial organization and regulation of the securities industry. University of Chicago Press, Chicago
41. Sarig O, Warga A (1989) Bond price data and bond market liquidity. J Financ Quant Anal 24:367–378
42. Schultz P (2001) Corporate bond trading costs: A peek behind the curtain. J Financ 56:677–698
43. Shilony Y (1977) Mixed pricing in oligopoly. J Econ Theory 14:373–88
44. Warga A (1991) Corporate bond price discrepancies in the dealer and exchange markets. J Fixed Income 1:7–16
45. Warga A (1992) Bond returns, liquidity and missing data. J Financ Quant Anal 27:605–616

Books and Reviews

Greene W (1993) Econometric Analysis, 2nd edn. Macmillan, New York

O'Hara M (1997) Market microstructure theory. Basil Blackwell, Cambridge

Shultz B (1946) The securities market and how it works. Harper, New York

US Securities and Exchange Commission (2004) Report on transactions in municipal securities. http://www.sec.gov/news/studies/munireport2004.pdf

Speeches, Public Statements, and Testimony by US Bond Market Regulators and Participants

September 9, 1998, The Importance of Transparency in America's Debt Market, Remarks of SEC Chairman Arthur Levitt at the Media Studies Center, New York, NY, http://www.sec.gov/news/speech/speecharchive/1998/spch218.htm

April 23, 1999, Remarks of SEC Chairman Arthur Levitt Before the Bond Market Association, San Francisco, CA, http://www.sec.gov/news/speech/speecharchive/1999/spch268.htm

October 28, 1999, Electronic Trading Technology's Impact on the Fixed-Income Markets, Remarks of SEC Commissioner Laura S. Unger at The Bond Market Association, Fifth Annual Legal and Compliance Seminar, New York, NY, http://www.sec.gov/news/speech/speecharchive/1999/spch313.htm

January 8, 2002, Remarks Before the Bond Market Association Legal and Compliance Conference, by Annette L. Nazareth, SEC Director, Division of Market Regulation, http://www.sec.gov/news/speech/spch532.htm

April 25, 2002, Remarks Before the Annual Meeting of the Bond Market Association, by SEC Chairman Harvey L. Pitt, New York, NY, http://www.sec.gov/news/speech/spch553.htm

February 3, 2004, Legal & Compliance Conference – The Bond Market Association, by SEC Commissioner Cynthia A. Glassman, New York, NY, http://www.sec.gov/news/speech/spch020304cag.htm

June 17, 2004, US Senate Committee on Banking, Housing, and Urban Affairs, Hearing on "An Overview of the Regulation of the Bond Markets", http://banking.senate.gov/public

September 9, 2004, Remarks before the 2004 Bond Attorney's Workshop of the National Association of Bond Lawyers, by Martha Mahan Haines, SEC Assistant Director, Office of Municipal Securities, Division of Market Regulation, Chicago, IL, http://www.sec.gov/news/speech/spch090904mmh.htm

October 1, 2004, Keynote Address before the NASD Conference on Fixed Income Markets, by Annette L. Nazareth, SEC Director, Division of Market Regulation, New York City, NY, http://www.sec.gov/news/speech/spch100104aln.htm

December 16, 2004, Second MTS Conference on Financial Markets: "The Organization and Performance of Fixed-Income Markets", by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Vienna, Austria, http://www.sec.gov/news/speech/spch121604cs.htm

February 1, 2005, Remarks before the Bond Market Association 2005 Legal & Compliance Conference, by SEC Commissioner Paul S. Atkins, New York, NY, http://www.sec.gov/news/speech/spch020105psa.htm

February 2, 2005, Keynote Address at the Bond Market Association 2005 Legal & Compliance Conference, by NASD Vice Chairman Doug Shulman, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P013225

April 20, 2005, Remarks before the Bond Market Association, by SEC Chairman William H. Donaldson, New York, NY, http://www.sec.gov/news/speech/spch042005whd.htm

April 20, 2005, Navigating the Changing Climate in Fixed-Income Products, Remarks of NASD President Doug Shulman before the Bond Market Association, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P013842

May 6, 2005, Broad Themes in Market Microstructure, by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Cambridge, MA, http://www.sec.gov/news/speech/spch050605css.htm

June 21, 2005, Developing Bond Markets in APEC: Key Lessons from the US Experience, Remarks of SEC Commissioner Roel C. Campos before the ABAC/ADBI/PECC Conference, Tokyo, Japan, http://www.sec.gov/news/speech/spch062105rcc-2.htm

September 13, 2005, Remarks for Promethee: Transatlantic Dialogue and Regulatory Convergence: Panel on Financial Markets, by Ethiopis Tafara, SEC Director, Office of International Affairs, Paris, France, http://www.sec.gov/news/speech/spch091305et.htm

September 15, 2005, Keynote Address before the 30th Bond Attorney's Workshop of the National Association of Bond Lawyers, by SEC Commissioner Roel C. Campos, Chicago, IL, http://www.sec.gov/news/speech/spch091505rcc.htm

November 17, 2005, Address by NASD President Doug Shulman to the NASD Fall Securities Conference, San Francisco, CA, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P015554

January 6, 2006, Discussion: An Overview of Bond Market Transparency, by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Boston, MA, http://www.sec.gov/news/speech/spch010606css.htm

February 7, 2006, Remarks before the TBMA Legal and Compliance Conference, by SEC Commissioner Annette L. Nazareth, New York, NY, http://www.sec.gov/news/speech/spch020706aln.htm

May 19, 2006, The Continuing Evolution of the Bond Market and Other Observations: Remarks Before the Bond Market Association's 30th Annual Conference, by SEC Commissioner Cynthia A. Glassman, New York, NY, http://www.sec.gov/news/speech/2006/spch051906cag.htm

May 19, 2006, Remarks Before the Bond Market Association's 30th Annual Conference, by NASD President Doug Shulman, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P016651

Miscellaneous Sources of Information on International Corporate Bond Markets

April, 2004, Markets in Financial Instruments Directive (MiFID), Directive 2004/39/EC of the European Parliament and of the Council, Article 65, http://europa.eu.int/eur-lex/pri/en/oj/dat/2004/l_145/l_14520040430en00010044.pdf

May 2004, International Organization of Securities Commissions (IOSCO), "Transparency of Corporate Bond Markets", Report of the Technical Committee of IOSCO, http://www.iosco.org/library/pubdocs/pdf/IOSCOPD168.pdf

September 2005, UK Financial Services Authority (FSA), "Trading Transparency in the UK Secondary Bond Markets", FSA Discussion Paper 05/5, http://www.fsa.gov.uk/pubs/discussion/dp05_05.pdf

November, 2005, "Developing Bond Markets in Asia Conference", jointly by the Asian Office of the Bank for International Settlements (BIS) and the People's Bank of China (PBC), Kunming, China, http://www.bis.org/publ/bppdf/bispap26.htm

May 2006, Centre for Economic Policy Research (CEPR), "European corporate bond markets: Transparency, liquidity, efficiency", http://www.cepr.org/PRESS/TT_CorporateFULL.pdf

Correlated Equilibria and Communication in Games

FRANÇOISE FORGES
Ceremade, Université Paris-Dauphine, Paris, France

Article Outline

Glossary

Bayesian game An interactive decision problem consisting of a set of n players, a set of types for every player, a probability distribution which accounts for the players' beliefs over each others' types, a set of actions for every player and a *von Neumann–Morgenstern utility function* defined over n-tuples of types and actions for every player.

Nash equilibrium In an n-person strategic form game, a strategy n-tuple from which unilateral deviations are not profitable.

von Neumann–Morgenstern utility function A *utility function* which reflects the individual's preferences over lotteries. Such a utility function is defined over outcomes and can be extended to any lottery λ by taking expectation with respect to λ.

Pure strategy (or simply strategy) A mapping which, in an interactive decision problem, associates an action with the information of a player whenever this player can make a choice.

Sequential equilibrium A refinement of the *Nash equilibrium* for n-person multistage interactive decision problems, which can be loosely defined as a strategy n-tuple together with beliefs over past information for every player, such that every player maximizes his expected utility given his beliefs and the others' strategies, with the additional condition that the beliefs satisfy (possibly sophisticated) Bayes updating given the strategies.

Strategic (or normal) form game An interactive decision problem consisting of a set of n players, a set of strategies for every player and a (typically, *von Neumann–Morgenstern) utility function* defined over n-tuples of strategies for every player.

Utility function A real valued mapping over a set of outcomes which reflects the preferences of an individual by associating a utility level (a "payoff") with every outcome.

Definition of the Subject

The correlated equilibrium is a game theoretic solution concept. It was proposed by Aumann [1,2] in order to capture the strategic correlation opportunities that the players face when they take into account the extraneous environment in which they interact. The notion is illustrated in Sect. "Introduction". A formal definition is given in Sect. "Correlated Equilibrium: Definition and Basic Properties". The correlated equilibrium also appears as the appropriate solution concept if preplay communication is allowed between the players. As shown in Sect. "Correlated Equilibrium and Communication", this property can be given several precise statements according to the constraints imposed on the players' communication, which can go from plain conversation to exchange of messages through noisy channels. Originally designed for static games with complete information, the correlated equilibrium applies to any strategic form game. It is geometrically and computationally more tractable than the better known Nash equilibrium. The solution concept has been extended to dynamic games, possibly with incomplete information. As an illustration, we define in details the communication equilibrium for Bayesian games in Sect. "Correlated Equilibrium in Bayesian Games".

Introduction

Example

Consider the two-person game known as "chicken", in which each player i can take a "pacific" action (denoted as p^i) or an "aggressive" action (denoted as a^i):

$$\begin{array}{ccc} & p^2 & a^2 \\ p^1 & (8,8) & (3,10) \\ a^1 & (10,3) & (0,0) \end{array}$$

The interpretation is that player 1 and player 2 simultaneously choose an action and then get a payoff, which is determined by the pair of chosen actions according to the previous matrix. If both players are pacific, they both get 8. If both are aggressive, they both get 0. If one player is aggressive and the other is pacific, the aggressive player gets 10 and the pacific one gets 3. This game has two pure Nash equilibria (p^1, a^2), (a^1, p^2) and one mixed Nash equilibrium in which both players choose the pacific action with probability 3/5, resulting in the expected payoff 6 for both players. A possible justification for the latter solution is that the players make their choices as a function of independent extraneous random signals. The assumption of independence is strong. Indeed, there may be no way to prevent the players' signals from being correlated.

Consider a random signal which has no effect on the players' payoffs and takes three possible values: low, medium or high, occurring each with probability 1/3. Assume that, before the beginning of the game, player 1 distinguishes whether the signal is high or not, while player 2 distinguishes whether the signal is low or not. The relevant interactive decision problem is then the extended game in which the players can base their action on the private information they get on the random signal, while the payoffs only depend on the players' actions. In this game, suppose that player 1 chooses the aggressive action when the signal is high and the pacific action otherwise. Similarly, suppose that player 2 chooses the aggressive action when the signal is low and the pacific action otherwise. We show that these strategies form an equilibrium in the extended game. Given player 2's strategy, assume that player 1 observes a high signal. Player 1 deduces that the signal cannot be low so that player 2 chooses the pacific action; hence player 1's best response is to play aggressively. Assume now that player 1 is informed that the signal is not high; he deduces that with probability 1/2, the signal is medium (i. e., not low) so that player 2 plays pacific and with probability 1/2, the signal is low so that player 2 plays aggressive. The expected payoff of player 1 is 5.5 if he plays pacific and 5 if he plays aggressive; hence, the pacific action is a best response. The equilibrium conditions for player 2 are symmetric. To sum up, the strategies based on the players' private information form a Nash equilibrium in the extended game in which an extraneous signal is first selected. We shall say that these strategies form a "correlated equilibrium". The corresponding probability distribution over the players' actions is

$$\begin{array}{ccc} & p^2 & a^2 \\ p^1 & \frac{1}{3} & \frac{1}{3} \\ a^1 & \frac{1}{3} & 0 \end{array} \tag{1}$$

and the expected payoff of every player is 7. This probability distribution can be used directly to make private recommendations to the players before the beginning of the game (see the canonical representation below).

Correlated Equilibrium: Definition and Basic Properties

Definition

A game in strategic form $G = (N, (\Sigma^i)_{i \in N}, (u^i)_{i \in N})$ consists of a set of players N together with, for every player $i \in N$, a set of strategies (for instance, a set of actions) Σ^i and a (von Neumann–Morgenstern) utility function $u^i : \Sigma \to \mathbb{R}$, where $\Sigma = \prod_{j \in N} \Sigma^j$ is the set of all strategy profiles. We assume that the sets N and Σ^i, $i \in N$, are finite.

A correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i \in N})$ is described by a finite set of signals Ω, a probability distribution $\boldsymbol{q}$ over

Ω and a partition $\mathcal{P}^i$ of Ω for every player $i \in N$. Since Ω is finite, the probability distribution $\boldsymbol{q}$ is just a real vector $\boldsymbol{q} = (\boldsymbol{q}(\omega))_{\omega\in\Omega}$ such that $\boldsymbol{q}(\omega) \geq 0$ and $\sum_{\omega\in\Omega} \boldsymbol{q}(\omega) = 1$.

From G and d, we define the extended game G_d as follows:

- ω is chosen in Ω according to $\boldsymbol{q}$
- every player i is informed of the element $P^i(\omega)$ of $\mathcal{P}^i$ which contains ω
- G is played: every player i chooses a strategy σ^i in Σ^i and gets the utility $u^i(\sigma)$, $\sigma = (\sigma^j)_{j\in N}$.

A (pure) strategy for player i in G_d is a mapping $\alpha^i \colon \Omega \to \Sigma^i$ which is $\mathcal{P}^i$-measurable, namely, such that $\alpha^i(\omega') = \alpha^i(\omega)$ if $\omega' \in P^i(\omega)$. The interpretation is that, in G_d, every player i chooses his strategy σ^i as a function of his private information on the random signal ω which is selected before the beginning of G.

According to Aumann [1], a *correlated equilibrium* of G is a pair (d, α), which consists of a correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$ and a Nash equilibrium $\alpha = (\alpha^i)_{i\in N}$ of G_d. The equilibrium conditions of every player i, conditionally on his private information, can be written as:

$$\begin{aligned}&\sum_{\omega'\in P^i(\omega)} q(\omega'|P^i(\omega))u^i(\alpha(\omega'))\\ &\quad\geq \sum_{\omega'\in P^i(\omega)} q(\omega'|P^i(\omega))u^i(\tau^i, \alpha^{-i}(\omega'))\,,\\ &\qquad\qquad \forall i \in N,\ \forall \tau^i \in \Sigma^i,\ \forall \omega \in \Omega\colon q(\omega) > 0\,, \end{aligned} \tag{2}$$

where $\alpha^{-i} = (\alpha^j)_{j\neq i}$.

A mixed Nash equilibrium $\rho = (\rho^i)_{i\in N}$ of G can be viewed as a correlated equilibrium of G. By definition, every ρ^i is a probability distribution over Σ^i, the finite set of pure strategies of player i. Let us consider the correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$ in which $\Omega = \Sigma = \prod_{j\in N} \Sigma^j$, $\boldsymbol{q}$ is the product probability distribution induced by the mixed strategies (i. e., $\boldsymbol{q}((\sigma^j)_{j\in N}) = \prod_{j\in N} \rho^j(\sigma^j)$) and for each i, $\mathcal{P}^i$ is the partition of Ω generated by Σ^i (i. e., for $\omega, \nu \in \Omega$, $\nu \in P^i(\omega) \Leftrightarrow \nu^i = \omega^i$). Let $\alpha^i \colon \Sigma \to \Sigma^i$ be the projection over Σ^i (i. e., $\alpha^i(\sigma) = \sigma^i$). The correlation device d and the strategies α^i defined in this way form a correlated equilibrium. As we shall see below, this correlated equilibrium is "canonical" .

Canonical Representation

A *canonical* correlated equilibrium of G is a correlated equilibrium in which $\Omega = \Sigma = \prod_{j\in N} \Sigma^j$ while for every player i, the partition $\mathcal{P}^i$ of Σ is generated by Σ^i and $\alpha^i \colon \Sigma \to \Sigma^i$ is the projection over Σ^i. A canonical correlated equilibrium is thus fully specified by a probability distribution $\boldsymbol{q}$ over Σ. A natural interpretation is that a mediator selects $\sigma = (\sigma^j)_{j\in N}$ according to $\boldsymbol{q}$ and privately recommends σ^i to player i, for every $i \in N$. The players are not forced to obey the mediator, but σ is selected in such a way that player i cannot benefit from deviating unilaterally from the recommendation σ^i, i. e., $\tau^i = \sigma^i$ maximizes the conditional expectation of player i's payoff $u^i(\tau^i, \sigma^{-i})$ given the recommendation σ^i. A probability distribution $\boldsymbol{q}$ over Σ thus defines a canonical correlated equilibrium if and only if it satisfies the following linear inequalities:

$$\begin{aligned}&\sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^{-i}|\sigma^i)u^i(\sigma^i, \sigma^{-i})\\ &\quad\geq \sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^{-i}|\sigma^i)u^i(\tau^i, \sigma^{-i})\,,\\ &\qquad\qquad \forall i \in N,\ \forall \sigma^i \in \Sigma^i\colon \boldsymbol{q}(\sigma^i) > 0,\ \forall \tau^i \in \Sigma^i\end{aligned}$$

or, equivalently,

$$\begin{aligned}&\sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^i, \sigma^{-i})u^i(\sigma^i, \sigma^{-i})\\ &\quad\geq \sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^i, \sigma^{-i})u^i(\tau^i, \sigma^{-i})\,,\\ &\qquad\qquad \forall i \in N,\ \forall \sigma^i,\ \tau^i \in \Sigma^i\end{aligned} \tag{3}$$

The equilibrium conditions can also be formulated *ex ante*:

$$\begin{aligned}&\sum_{\sigma\in\Sigma} \boldsymbol{q}(\sigma)u^i(\sigma) \geq \sum_{\sigma\in\Sigma} \boldsymbol{q}(\sigma)u^i(\alpha^i(\sigma^i), \sigma^{-i})\,,\\ &\qquad\qquad \forall i \in N,\ \forall \alpha^i\colon \Sigma^i \to \Sigma^i\end{aligned}$$

The following result is an analog of the "revelation principle" in mechanism design (see, e. g., Myerson [41]): *let (α, d) be a correlated equilibrium associated with an arbitrary correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$. The corresponding "correlated equilibrium distribution", namely, the probability distribution induced over Σ by $\boldsymbol{q}$ and α, defines a canonical correlated equilibrium*. For instance, in the introduction, (1) describes a canonical correlated equilibrium.

Duality and Existence

From the linearity of (3), duality theory can be used to study the properties of correlated equilibria, in particular to prove their existence without relying on Nash's [45] theorem and its fixed point argument (recall that every mixed Nash equilibrium is a correlated equilibrium). Hart and Schmeidler [30] establish the existence of a correlated equilibrium by constructing an auxiliary two person zero-sum game and applying the minimax theorem. Nau and McCardle [47] derive another elementary proof of existence from an extension of the "no arbitrage opportunities" axiom that underlies subjective probability theory. They introduce jointly coherent strategy profiles, which do not expose the players as a group to arbitrage from an outside observer. They show that a strategy profile is jointly coherent if and only if it occurs with positive probability in some correlated equilibrium. From a technical point of view, both proofs turn out to be similar. Myerson [44] makes further use of the linear structure of correlated equilibria by introducing dual reduction, a technique to replace a finite game with a game with fewer strategies, in such a way that any correlated equilibrium of the reduced game induces a correlated equilibrium of the original game.

Geometric Properties

As (3) is a system of linear inequalities, the set of all correlated equilibrium distributions is a convex polytope. Nau et al. [49] show that if it has "full" dimension (namely, dimension $|\Sigma| - 1$), then all Nash equilibria lie on its relative boundary. Viossat [60] characterizes in addition the class of games whose correlated equilibrium polytope contains a Nash equilibrium in its relative interior. Interestingly, this class of games includes two person zero-sum games but is not defined by "strict competition" properties. In two person games, all extreme Nash equilibria are also extreme correlated equilibria [13,25]; this result does not hold with more than two players. Finally, Viossat [59] proves that having a unique correlated equilibrium is a robust property, in the sense that the set of n person games with a unique correlated equilibrium is open. The same is not true for Nash equilibrium (unless $n = 2$).

Complexity

From (3), correlated equilibria can be computed by linear programming methods. Gilboa and Zemel [24] show more precisely that the complexity of standard computational problems is "NP-hard" for the Nash equilibrium and *polynomial* for the correlated equilibrium. Examples of such problems are: "Does the game G have a Nash (resp., correlated) equilibrium which yields a payoff greater than r to every player (for some given number r)?" and "Does the game G have a unique Nash (resp., correlated) equilibrium?". Papadimitriou [50] develops a polynomial-time algorithm for finding correlated equilibria, which is based on a variant of the existence proof of Hart and Schmeidler [30].

Foundations

By re-interpreting the previous canonical representation, Aumann [2] proposes a decision theoretic foundation for the correlated equilibrium in games with complete information, in which Σ^i for $i \in N$, stands merely for a set of *actions* of player i. Let Ω be the space of all states of the world; an element ω of Ω thus specifies all the parameters which may be relevant to the players' choices. In particular, the action profile in the underlying game G is part of the state of the world. A partition $\mathcal{P}^i$ describes player i's information on Ω. In addition, every player i has a prior belief, i. e., a probability distribution, $\boldsymbol{q}^i$ over Ω. Formally, the framework is similar as above except that the players possibly hold different beliefs over Ω. Let $\alpha^i(\omega)$ denote player i's action at ω; a natural assumption is that player i knows the action he chooses, namely that α^i is $\mathcal{P}^i$-measurable. According to Aumann [2], player i is *Bayes-rational* at ω if his action $\alpha^i(\omega)$ maximizes his expected payoff (with respect to $\boldsymbol{q}^i$) given his information $P^i(\omega)$. Note that this is a separate rationality condition for every player, not an equilibrium condition. Aumann [2] proves the following result: *Under the common prior assumption (namely, $\boldsymbol{q}^i = \boldsymbol{q}$, $i \in N$), if every player is Bayes-rational at every state of the world, the distribution of the corresponding action profile α is a correlated equilibrium distribution*. The key to this decision theoretic foundation of the correlated equilibrium is that, under the common prior assumption, Bayesian rationality amounts to (2).

If the common prior assumption is relaxed, the previous result still holds, with *subjective* prior probability distributions, for the *subjective* correlated equilibrium which was also introduced by Aumann [1]. The latter solution concept is defined in the same way as above, by considering a device $(\Omega, (\boldsymbol{q}^i)_{i \in N}, (\mathcal{P}^i)_{i \in N})$, with a probability distribution $\boldsymbol{q}^i$ for every player i, and by writing (2) in terms of $\boldsymbol{q}^i$ instead of $\boldsymbol{q}$. Brandenburger and Dekel [10] show that (a refinement of) the subjective correlated equilibrium is equivalent to (correlated) rationalizability, another well-established solution concept which captures players' minimal rationality. Rationalizable strategies reflect that the players commonly know that each of them makes an optimal choice given some belief. Nau and McCardle [48] reconcile objective and subjective correlated equilibrium by proposing the no arbitrage principle as a unified approach to individual and interactive decision problems. They argue that the objective correlated equilibrium concept applies to a game that is revealed by the players' choices, while the subjective correlated equilibrium concept applies to the “true game”; both lead to the same set of jointly coherent outcomes.

Correlated Equilibrium and Communication

As seen in the previous section, correlated equilibria can be achieved in practice with the help of a mediator and emerge in a Bayesian framework embedding the game in a full description of the world. Both approaches require to extend the game by taking into account information which is not generated by the players themselves. Can the players reach a correlated equilibrium without relying on any extraneous correlation device, by just communicating with each other before the beginning of the game?

Consider the game of “chicken” presented in the introduction. The probability distribution

$$\begin{array}{ccc} & p^2 & a^2 \\ p^1 & 0 & \frac{1}{2} \\ a^1 & \frac{1}{2} & 0 \end{array} \tag{4}$$

describes a correlated equilibrium, which amounts to choosing one of the two pure Nash equilibria, with equal probability. Both players get an expected payoff of 6.5. Can they safely achieve this probability distribution if no mediator tosses a fair coin for them? The answer is positive, as shown by Aumann et al. [3]. Assume that before playing “chicken”, the players independently toss a coin and simultaneously reveal to each other whether heads or tails obtains. Player 1 tells player 2 “h^1” or “t^1” and, at the same time, player 2 tells player 1 “h^2” or “t^2”. If both players use a fair coin, reveal correctly the result of the toss and play (p^1, a^2) if both coins fell on the same side (i. e., if (h^1, h^2) or (t^1, t^2) is announced) and (a^1, p^2) otherwise (i. e., if (h^1, t^2) or (t^1, h^2) is announced), they get the same effect as a mediator using (4). Furthermore, none of them can gain by unilaterally deviating from the described strategies, even at the randomizing stage: the two relevant outcomes, $[(h^1, h^2)$ or $(t^1, t^2)]$ and $[(h^1, t^2)$ or $(t^1, h^2)]$, happen with probability 1/2 provided that one of the players reveals the toss of a fair coin. This procedure is known as a “jointly controlled lottery” . An important feature of the previous example is that, in the correlated equilibrium described by (4), the players know each other's recommendation. Hence, they can easily reproduce (4) by exchanging messages that they have selected independently. In the correlated equilibrium described by the probability distribution (1), the private character of recommendations is crucial to guarantee that (p^1, p^2) be played with positive probability. Hence one cannot hope that a simple procedure of direct preplay communication be sufficient to generate (1). However, the fact that direct communication is necessarily public is typical of two-person games.

Given the game $G = (N, (\Sigma^i)_{i \in N}, (u^i)_{i \in N})$, let us define a (bounded) “cheap talk” extension ext(G) of G as a game in which T stages of costless, unmediated preplay communication are allowed before G is played. More precisely, let M_t^i be a finite set of messages for player i, $i \in N$, at stage t, $t = 1, 2, \ldots T$; at every stage t of ext(G), every player i selects a message in M_t^i; these choices are made simultaneously before being revealed to a subset of players at the end of stage t. The rules of $ext(G)$ thus determine a set of “senders” for every stage t (those players i for whom M_t^i contains more than one message) and a set of “receivers” for every stage t. The players perfectly recall their past messages. After the communication phase, they choose their strategies (e. g., their actions) as in G; they are also rewarded as in G, independently of the preplay phase, which is thus “cheap” . Communication has an indirect effect on the final outcome in G, since the players make their decisions as a function of the messages that they have exchanged. Specific additional assumptions are often made on ext(G), as we will see below.

Let us fix a cheap talk extension ext(G) of G and a Nash equilibrium of $ext(G)$. As a consequence of the previous definitions, the distribution induced by this Nash equilibrium over Σ defines a correlated equilibrium of G (this can be proved in the same way as the canonical representation of correlated equilibria stated in Sect. “Correlated Equilibrium: Definition and Basic Properties”). The question raised in this section is whether the reverse holds.

If the number of players is two, the Nash equilibrium distributions of cheap talk extensions of G form a subset of the correlated equilibrium distributions: the convex hull of Nash equilibrium distributions. Indeed, the players have both the same information after any direct exchange of messages. Conversely, by performing repeated jointly controlled lotteries like in the example above, the players can achieve any convex combination (with rational weights) of Nash equilibria of G as a Nash equilibrium of a cheap talk extension of G. The restriction on probability distributions whose components are rational numbers is only needed as far as we focus on *bounded* cheap talk extensions.

Bárány [4] establishes that, if the number of players of G is at least four, every (rational) correlated equilibrium distribution of G can be realized as a Nash equilibrium of a cheap talk extension ext(G), provided that ext(G) allows the players to publicly check the record of communication under some circumstances. The equilibria of ext(G) constructed by Bárány involve that a receiver gets the same message from two different senders; the message is nevertheless not public thanks to the assumption on the number of players. At every stage of ext(G), every player can ask for the revelation of all past messages, which are assumed to be recorded. Typically, a receiver can claim that the two senders' messages differ. In this case, the record of communication surely reveals that either one of the senders or the receiver himself has cheated; the deviator can be punished (at his minmax level in G) by the other players.

The punishments in Bárány's [4] Nash equilibria of ext(G) need not be credible threats. Instead of using double senders in the communication protocols, Ben-Porath [5,6] proposes a procedure of random monitoring, which prescribes a given behavior to every player in such a way that unilateral deviations can be detected with probability arbitrarily close to 1. This procedure applies if there are at least three players, which yields an analog of Bárány's result already in this case. If the number of players is exactly three, Ben-Porath [6] needs to assumes, as Bárány [4], that public verification of the record of communication is possible in ext(G) (see Ben-Porath [7]). However, Ben-Porath concentrates on (rational) correlated equilibrium distributions which allow for strict punishment on a Nash equilibrium of G; he constructs *sequential* equilibria which generate these distributions in ext(G), thus dispensing with incredible threats. At the price of raising the number of players to five or more, Gerardi [22] proves that every (rational) correlated equilibrium distribution of G can be realized as a sequential equilibrium of a cheap talk extension of G which does not require any message recording. For this, he builds protocols of communication in which the players base their decisions on majority rule, so that no punishment is necessary.

We have concentrated on two extreme forms of communication: mediated communication, in which a mediator performs lotteries and sends private messages to the players and cheap talk, in which the players just exchange messages. Many intermediate schemes of communication are obviously conceivable. For instance, Lehrer [36] introduces (possibly multistage) "mediated talk": the players send private messages to a mediator, but the latter can only make deterministic public announcements. Mediated talk captures real-life communication procedures, like elections, especially if it lasts only for a few stages. Lehrer and Sorin [37] establish that whatever the number of players of G, every (rational) correlated equilibrium distribution of G can be realized as a Nash equilibrium of a single stage mediated talk extension of G. Ben-Porath [5] proposes a variant of cheap talk in which the players do not only exchange verbal messages but also "hard" devices such as urns containing balls. This extension is particularly useful in two-person games to circumvent the equivalence between the equilibria achieved by cheap talk and the convex hull of Nash equilibria. More precisely, the result of Ben-Porath [5] stated above holds for two-person games if the players first check together the content of different urns, and then each player draws a ball from an urn that was chosen by the other player, so as to guarantee that one player only knows the outcome of a lottery while the other one only knows the probabilities of this lottery.

The various extensions of the basic game G considered up to now, with or without a mediator, implicitly assume that the players are fully rational. In particular, they have unlimited computational abilities. By relaxing that assumption, Urbano and Vila [55] and Dodis et al. [12] build on earlier results from cryptography so as to implement any (rational) correlated equilibrium distribution through unmediated communication, including in two-person games.

As the previous paragraphs illustrate, the players can modify their intitial distribution of information by means of many different communication protocols. Gossner [26] proposes a general criterion to classify them: a protocol is "secure" if under all circumstances, the players cannot mislead each other nor spy on each other. For instance, given a cheap talk extension ext(G), a protocol P describes, for every player, a strategy in ext(G) and a way to interpret his information after the communication phase of ext(G). P induces a correlation device $d(P)$ (in the sense of Sect. "Correlated Equilibrium: Definition and Basic Properties"). P is secure if, *for every game* G and every Nash equilibrium α of $G_{d(P)}$, the following procedure is a Nash equilibrium of ext(G): communicate according to the strategies described by P in order to generate $d(P)$ and make the final choice, in G, according to α. Gossner [26] gives a tractable characterization of secure protocols.

Correlated Equilibrium in Bayesian Games

A Bayesian game $\Gamma = (N, (T^i)_{i \in N}, p, (A^i)_{i \in N}, (v^i)_{i \in N})$ consists of: a set of players N; for every player $i \in N$, a set of types T^i, a probability distribution p^i over $T = \prod_{j \in N} T^j$, a set of actions A^i and a (von Neumann–Morgenstern) utility function $v^i : T \times A \to \mathbb{R}$, where $A = \prod_{j \in N} A^j$. For simplicity, we make the common prior assumption: $p^i = p$ for every $i \in N$. All sets are assumed finite. The interpretation is that a virtual move of nature chooses $t = (t^j)_{j \in N}$ according to p; player i is only informed of his own type t^i; the players then choose simultaneously an action. We will focus on two possible extensions of Aumann's [1] solution concept to Bayesian games: the strategic form correlated equilibrium and the communication equilibrium. Without loss of generality, the definitions below are given in "canonical form" (see Sect. "Correlated Equilibrium: Definition and Basic Properties").

Strategic Form Correlated Equilibrium

A (pure) strategy of player i in Γ is a mapping $\sigma^i : T^i \to A^i$, $i \in N$. The strategic form of Γ is a game $G(\Gamma)$, like the game G considered in Sect. "Correlated Equilibrium: Definition and Basic Properties", with sets of pure strategies $\Sigma_i = A_i^{T_i}$ and utility functions u^i over $\Sigma = \prod_{j \in N} \Sigma^j$ computed as expectations with respect to p: $u^i(\sigma) = E[v^i(t, \sigma(t))]$, with $\sigma(t) = (\sigma^i(t^i))_{i \in N}$. A strategic form correlated equilibrium, or simply, a correlated equilibrium, of a Bayesian game Γ is a correlated equilibrium, in the sense of Sect. "Correlated Equilibrium: Definition and Basic Properties", of $G(\Gamma)$. A canonical correlated equilibrium of Γ is thus described by a probability distribution Q over Σ, which selects an N-tuple of pure strategies $(\sigma^i)_{i \in N}$. This lottery can be thought of as being performed by a mediator who privately recommends σ^i to player i, $i \in N$, before the beginning of Γ, i. e., before (or in any case, independently of) the chance move choosing the N -tuple of types. The equilibrium conditions express that, once he knows his type t^i, player i cannot gain in unilaterally deviating from $\sigma^i(t^i)$.

Communication Equilibrium

Myerson [41] transforms the Bayesian game Γ into a mechanism design problem by allowing the mediator to collect information from the players before making them recommendations. Following Forges [15] and Myerson [42], a canonical communication device for Γ consists of a system $\boldsymbol{q}$ of probability distributions $\boldsymbol{q} = (\boldsymbol{q}(.|t))_{t \in T}$ over A. The interpretation is that a mediator invites every player i, $i \in N$, to report his type t^i, then selects an N-tuple of actions a according to $\boldsymbol{q}(.|t)$ and privately recommends a^i to player i. The system $\boldsymbol{q}$ defines a communication equilibrium if none of the players can gain by unilaterally lying on his type or by deviating from the recommended action, namely if

$$\begin{aligned}&\sum_{t^{-i} \in T^{-i}} p(t^{-i}|t^i) \sum_{a \in A} q(a|t) v^i(t, a)\\ &\geq \sum_{t^{-i} \in T^{-i}} p(t^{-i}|t^i) \sum_{a \in A} q(a|s^i, t^{-i}) v^i(t, \alpha^i(a^i), a^{-i}) \,,\\ &\qquad \forall i \in N,\ \forall t^i,\ s^i \in T^i,\ \forall \alpha^i : A^i \to A^i\end{aligned}$$

Correlated Equilibrium, Communication Equilibrium and Cheap Talk

Every correlated equilibrium of the Bayesian game Γ induces a communication equilibrium of Γ, but the converse is not true, as the following example shows.

Consider the two-person Bayesian game in which $T^1 = \{s^1, t^1\}$, $T^2 = \{t^2\}$, $A^1 = \{a^1, b^1\}$, $A^2 = \{a^2, b^2\}$, $p(s^1) = p(t^1) = \frac{1}{2}$ and payoffs are described by

		a^2	b^2
s^1	a^1	$(1, 1)$	$(-1, -1)$
	b^1	$(0, 0)$	$(0, 0)$

		a^2	b^2
t^1	a^1	$(0, 0)$	$(0, 0)$
	b^1	$(-1, -1)$	$(1, 1)$

In this game, the communication equilibrium $\boldsymbol{q}(a^1, a^2|s^1) = \boldsymbol{q}(b^1, b^2|t^1) = 1$ yields the expected payoff of 1 to both players. However the maximal expected payoff of every player in a correlated equilibrium is 1/2. In order to see this, one can derive the strategic form of the game (in which player 1 has four strategies and player 2 has two strategies). Let us turn to the game in which player 1 can cheaply talk to player 2 just after having learned his type. In this new game, the following strategies form a Nash equilibrium: player 1 truthfully reveals his type to player 2 and plays a^1 if s^1, b^1 if t^1; player 2 chooses a^2 if s^1, b^2 if t^1 . These strategies achieve the same expected payoffs as the communication equilibrium.

As in Sect. "Correlated Equilibrium and Communication", one can define cheap talk extensions ext(Γ) of Γ. A wide definition of ext(Γ) involves an *ex ante* preplay phase, before the players learn their types, and an *interim* preplay phase, after the players learn their types but before they choose their actions. Every Nash equilibrium of ext(Γ) induces a communication equilibrium of Γ. In order to investigate the converse, namely whether cheap talk can simulate mediated communication in a Bayesian game, two approaches have been developed. The first one (Forges [17], Gerardi [21,22], Vida [58]) proceeds in two steps, by reducing communication equilibria to correlated equilibria before applying the results obtained for strategic form games (see Sect. "Correlated Equilibrium and Communication"). The second approach (Ben-Porath [6], Krishna [33]) directly addresses the question in a Bayesian game.

By developing a construction introduced for particular two person games (Forges [14]), Forges [17] shows that every communication equilibrium outcome of a Bayesian game Γ with at least four players can be achieved as a correlated equilibrium outcome of a two stage *interim* cheap talk extension $\mathrm{ext}_{\mathrm{int}}(\Gamma)$ of Γ. No punishment is necessary in $\mathrm{ext}_{\mathrm{int}}(\Gamma)$: at the second stage, every player gets a message from three senders and uses majority rule if the messages are not identical. Thanks to the underlying correlation device, each receiver is able to privately decode his message. Vida [58] extends Forges [17] to Bayesian games with three or even two players. In the proof, he constructs a correlated equilibrium of a long, but almost surely finite, *interim* cheap talk extension of Γ, whose length depends both on the signals selected by the correlation device and the messages exchanged by the players. No recording of messages is necessary to detect and punish a cheating player. If there are at least four players in Γ, once a communication equilibrium of Γ has been converted into a correlated equilibrium of $\mathrm{ext}_{\mathrm{int}}(\Gamma)$, one can apply Bárány's [4] result to $\mathrm{ext}_{\mathrm{int}}(\Gamma)$ in order to transform the correlated equilibrium into a Nash equilibrium of a further, *ex ante*, cheap talk preplay extension of Γ. Gerardi [21] modifies this ex ante preplay phase so as to postpone it at the *interim* stage. This result is especially useful if the initial move of nature in Γ is just a modelling convenience. Gerardi [22] also extends his result for at least five person games with complete information (see Sect. "Correlated Equilibrium and Communication") to any Bayesian game with full support (i. e., in which all type profiles have positive probability: $p(t) > 0$ for every $t \in T$) by proving that every (rational) communication equilibrium of Γ can be achieved as a *sequential* equilibrium of a cheap talk extension of Γ.

Ben-Porath [6] establishes that if Γ is a three (or more) person game with full support, every (rational) communication equilibrium of Γ which strictly dominates a Nash equilibrium of Γ for every type t^i of every player i, $i \in N$, can be implemented as a Nash equilibrium of an *interim* cheap talk extension of Γ in which public verification of past record is possible (see also Ben-Porath [7]). Krishna [33] extends Ben-Porath's [5] result on two person games (see Sect. "Correlated Equilibrium and Communication") to the incomplete information framework. The other results mentioned at the end of Sect. "Correlated Equilibrium and Communication" have also been generalized to Bayesian games (see [26,37,56]).

Related Topics and Future Directions

In this brief article, we concentrated on two solution concepts: the strategic form correlated equilibrium, which is applicable to any game, and the communication equilibrium, which we defined for Bayesian games. Other extensions of Aumann's [1]solution concept have been proposed for Bayesian games, as the agent normal form correlated equilibrium and the (possibly belief invariant) Bayesian solution (see Forges [18,19] for definitions and references). The Bayesian solution is intended to capture the players' rationality in games with incomplete information in the spirit of Aumann [2] (see Nau [46] and Forges [18]). Lehrer et al. [38] open a new perspective in the understanding of the Bayesian solution and other equilibrium concepts for Bayesian games by characterizing the classes of equivalent information structures with respect to each of them. Comparison of information structures, which goes back to Blackwell [8,9] for individual decision problems, was introduced by Gossner [27] in the context of games, both with complete and incomplete information. In the latter model, information structures basically describe how extraneous signals are selected as a function of the players' types; two information structures are equivalent with respect to an equilibrium concept if, in every game, they generate the same equilibrium distributions over outcomes.

Correlated equilibria, communication equilibria and related solution concepts have been studied in many other classes of games, like multistage games (see, e. g., [15,42]), repeated games with incomplete information (see, e. g., [14,16]) and stochastic games (see, e. g., [53,54]). The study of correlated equilibrium in repeated games with imperfect monitoring, initiated by Lehrer [34,35], proved to be particularly useful and is still undergoing. Lehrer [34] showed that if players are either fully informed of past actions or get no information (" standard-trivial" information structure), correlated equilibria are equivalent to Nash equilibria. In other words, all correlations can be generated internally, namely by the past histories, on which players have differential information. The schemes of internal correlation introduced to establish this result are widely applicable and inspired those of Lehrer [36] (see Sect. "Correlated Equilibrium and Communication"). In general repeated games with imperfect monitoring, Renault and Tomala [52] characterize communication equilibria but the amount of correlation that the players can achieve in a Nash equilibrium is still an open problem (see, e. g., [28,57] for recent advances).

Throughout this article, we defined a correlated equilibrium as a *Nash* equilibrium of an extension of the game under consideration. The solution concept can be strengthened by imposing some refinement, i. e., further rationality conditions, to the Nash equilibrium in this definition (see, e. g., [11,43]). Refinements of communication equilibria have also been proposed (see, e. g., [22,23,42]). Some authors (see, e. g., [39,40,51]) have also developed notions of *coalition proof* correlated equilibria, which resist not only to unilateral deviations, as in this article, but even to multilateral ones. A recurrent difficulty is that, for many of these stronger solution concepts, a useful canonical representation (as derived in Sect. "Correlated Equilibrium: Definition and Basic Properties") is not available.

Except for two or three references, we deliberately concentrated on the results published in the game theory and mathematical economics literature, while substantial achievements in computer science would fit in this survey. Both streams of research pursue similar goals but rely on different formalisms and techniques. For instance, computer scientists often make use of cryptographic tools which are not familiar in game theory. Halpern [29] gives an idea of recent developments at the interface of computer science and game theory (see in particular the section "implementing mediators") and contains a number of references.

Finally, the assumption of full rationality of the players can also be relaxed. Evolutionary game theory has developed models of learning in order to study the long term behavior of players with bounded rationality. Many possible dynamics are conceivable to represent more or less myopic attitudes with respect to optimization. Under appropriate learning procedures, which express for instance that agents want to minimize the regret of their strategic choices, the empirical distribution of actions converge to correlated equilibrium distributions (see, e. g., [20,31,32] for a survey). However, standard procedures, as the "replicator dynamics", may even eliminate all the strategies which have positive probability in a correlated equilibrium (see [61]).

Acknowledgment

The author wishes to thank Elchanan Ben-Porath, Frédéric Koessler, R. Vijay Krishna, Ehud Lehrer, Bob Nau, Indra Ray, Jérôme Renault, Eilon Solan, Sylvain Sorin, Bernhard von Stengel, Tristan Tomala, Amparo Urbano, Yannick Viossat and, especially, Olivier Gossner and Péter Vida, for useful comments and suggestions.

Bibliography

Primary Literature

1. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–96
2. Aumann RJ (1987) Correlated equilibrium as an expression of Bayesian rationality. Econometrica 55:1–18
3. Aumann RJ, Maschler M, Stearns R (1968) Repeated games with incomplete information: an approach to the nonzero sum case. Reports to the US Arms Control and Disarmament Agency, ST-143, Chapter IV, 117–216 (reprinted In: Aumann RJ, Maschler M (1995) Repeated Games of Incomplete Information. M.I.T. Press, Cambridge)
4. Bárány I (1992) Fair distribution protocols or how players replace fortune. Math Oper Res 17:327–340
5. Ben-Porath E (1998) Correlation without mediation: expanding the set of equilibrium outcomes by cheap pre-play procedures. J Econ Theor 80:108–122
6. Ben-Porath E (2003) Cheap talk in games with incomplete information. J Econ Theor 108:45–71
7. Ben-Porath E (2006) A correction to "Cheap talk in games with incomplete information". Mimeo, Hebrew University of Jerusalem, Jerusalem
8. Blackwell D (1951) Comparison of experiments. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles, pp 93–102
9. Blackwell D (1953) Equivalent comparison of experiments. Ann Math Stat 24:265–272
10. Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. Econometrica 55:1391–1402
11. Dhillon A, Mertens JF (1996) Perfect correlated equilibria. J Econ Theor 68:279–302
12. Dodis Y, Halevi S, Rabin T (2000) A cryptographic solution to a game theoretic problem. CRYPTO 2000: 20th International Cryptology Conference. Springer, Berlin, pp 112–130
13. Evangelista F, Raghavan TES (1996) A note on correlated equilibrium. Int J Game Theory 25:35–41
14. Forges F (1985) Correlated equilibria in a class of repeated games with incomplete information. Int J Game Theory 14:129–150
15. Forges F (1986) An approach to communication equilibrium. Econometrica 54:1375–1385
16. Forges F (1988) Communication equilibria in repeated games with incomplete information. Math Oper Res 13:191–231
17. Forges F (1990) Universal mechanisms. Econometrica 58:1341–1364
18. Forges F (1993) Five legitimate definitions of correlated equilibrium in games with incomplete information. Theor Decis 35:277–310
19. Forges F (2006) Correlated equilibrium in games with incomplete information revisited. Theor Decis 61:329–344
20. Foster D, Vohra R (1997) Calibrated learning and correlated equilibrium. Game Econ Behav 21:40–55
21. Gerardi D (2000) Interim pre-play communication. Mimeo, Yale University, New Haven
22. Gerardi D (2004) Unmediated communication in games with complete and incomplete information. J Econ Theory 114:104–131
23. Gerardi D, Myerson R (2007) Sequential equilibria in Bayesian games with communication. Game Econ Behav 60:104–134
24. Gilboa I, Zemel E (1989) Nash and correlated equilibria: some complexity considerations. Game Econ Behav 1:80–93
25. Gomez-Canovas S, Hansen P, Jaumard B (1999) Nash Equilibria from the correlated equilibria viewpoint. Int Game Theor Rev 1:33–44
26. Gossner O (1998) Secure protocols or how communication generates correlation. J Econ Theory 83:69–89
27. Gossner O (2000) Comparison of information structures. Game Econ Behav 30:44–63
28. Gossner O, Tomala T (2007) Secret correlation in repeated games with signals. Math Oper Res 32:413–424
29. Halpern JY (2007) Computer science and game theory. In: Durlauf SN, Blume LE (eds) The New Palgrave dictio-

nary of economics, 2nd edn. Palgrave Macmillan. The New Palgrave dictionary of economics online. http://www.dictionaryofeconomics.com/article?id=pde2008_C000566. Accessed 24 May 2008

30. Hart S, Schmeidler D (1989) Existence of correlated equilibria. Math Oper Res 14:18–25
31. Hart S, Mas-Colell A (2000) A simple adaptive procedure leading to correlated equilibrium. Econometrica 68:1127–1150
32. Hart S (2005) Adaptative heuristics. Econometrica 73:1401–1430
33. Krishna RV (2007) Communication in games of incomplete information: two players. J Econ Theory 132:584–592
34. Lehrer E (1991) Internal correlation in repeated games. Int J Game Theory 19:431–456
35. Lehrer E (1992) Correlated equilibria in two-player repeated games with non-observable actions. Math Oper Res 17:175–199
36. Lehrer E (1996) Mediated talk. Int J Game Theory 25:177–188
37. Lehrer E, Sorin S (1997) One-shot public mediated talk. Game Econ Behav 20:131–148
38. Lehrer E, Rosenberg D, Shmaya E (2006) Signaling and mediation in Bayesian games. Mimeo, Tel Aviv University, Tel Aviv
39. Milgrom P, Roberts J (1996) Coalition-proofness and correlation with arbitrary communication possibilities. Game Econ Behav 17:113–128
40. Moreno D, Wooders J (1996) Coalition-proof equilibrium. Games Econ Behav 17:80–112
41. Myerson R (1982) Optimal coordination mechanisms in generalized principal-agent problems. J Math Econ 10:67–81
42. Myerson R (1986a) Multistage games with communication. Econometrica 54:323–358
43. Myerson R (1986b) Acceptable and predominant correlated equilibria. Int J Game Theory 15:133–154
44. Myerson R (1997) Dual reduction and elementary games. Game Econ Behav 21:183–202
45. Nash J (1951) Non-cooperative games. Ann Math 54:286–295
46. Nau RF (1992) Joint coherence in games with incomplete information. Manage Sci 38:374–387
47. Nau RF, McCardle KF (1990) Coherent behavior in noncooperative games. J Econ Theory 50(2):424–444
48. Nau RF, McCardle KF (1991) Arbitrage, rationality and equilibrium. Theor Decis 31:199–240
49. Nau RF, Gomez-Canovas S, Hansen P (2004) On the geometry of Nash equilibria and correlated equilibria. Int J Game Theory 32:443–453
50. Papadimitriou CH (2005) Computing correlated equilibria in multiplayer games. Proceedings of the 37th ACM Symposium on Theory of Computing. STOC, Baltimore, pp 49–56
51. Ray I (1996) Coalition-proof correlated equilibrium: a definition. Game Econ Behav 17:56–79
52. Renault J, Tomala T (2004) Communication equilibrium payoffs in repeated games with imperfect monitoring. Game Econ Behav 49:313–344
53. Solan E (2001) Characterization of correlated equilibrium in stochastic games. Int J Game Theory 30:259–277
54. Solan E, Vieille N (2002) Correlated equilibrium in stochastic games. Game Econ Behav 38:362–399
55. Urbano A, Vila J (2002) Computational complexity and communication: coordination in two-player games. Econometrica 70:1893–1927
56. Urbano A, Vila J (2004a) Computationally restricted unmediated talk under incomplete information. J Econ Theory 23:283–320
57. Urbano A, Vila J (2004b) Unmediated communication in repeated games with imperfect monitoring. Game Econ Behav 46:143–173
58. Vida P (2007) From communication equilibria to correlated equilibria. Mimeo, University of Vienna, Vienna
59. Viossat Y (2005) Is having a unique equilibrium robust? to appear in J Math Econ
60. Viossat Y (2006) The geometry of Nash equilibria and correlated equilibria and a generalization of zero-sum games. Mimeo, S-WoPEc working paper 641, Stockholm School of Economics, Stockholm
61. Viossat Y (2007) The replicator dynamics does not lead to correlated equilibria. Game Econ Behav 59:397–407

Books and Reviews

Forges F (1994) Non-zero sum repeated games and information transmission. In: Megiddo N (ed) Essays in Game Theory in Honor of Michael Maschler. Springer, Berlin, pp 65–95

Mertens JF (1994) Correlated- and communication equilibria. In: Mertens JF, Sorin S (eds) Game Theoretic Methods in General Equilibrium Analysis. Kluwer, Dordrecht, pp 243–248

Myerson R (1985) Bayesian equilibrium and incentive compatibility. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) Social Goals and Social Organization. Cambridge University Press, Cambridge, pp 229–259

Myerson R (1994) Communication, correlated equilibria and incentive compatibility. In: Aumann R, Hart S (eds) Handbook of Game Theory, vol 2. Elsevier, Amsterdam, pp 827–847

Sorin S (1997) Communication, correlation and cooperation. In: Mas Colell A, Hart S (eds) Cooperation: Game Theoretic Approaches. Springer, Berlin, pp 198–218

Correlated Percolation

ANTONIO CONIGLIO[1,2,3], ANNALISA FIERRO[2]

[1] Dipartimento di Scienze Fisiche, Università di Napoli “Federico II”, Complesso Universitario di Monte Sant’Angelo, Naples, Italy

[2] INFM-CNR Coherentia, Università di Napoli “Federico II”, Naples, Italy

[3] INFN Udr di Napoli, Università di Napoli “Federico II”, Naples, Italy

Article Outline

Definition of the Subject

Cluster concepts have been extremely useful in elucidating many problems in physics. Percolation theory provides a generic framework to study the behavior of the cluster distribution. In most cases the theory predicts a geometrical transition at the percolation threshold, characterized in the percolative phase by the presence of a spanning cluster, which becomes infinite in the thermodynamic limit. Standard percolation usually deals with the problem when the constitutive elements of the clusters are randomly distributed. However correlations cannot always be neglected. In this case correlated percolation is the appropriate theory to study such systems. The origin of correlated percolation could be dated back to 1937 when Mayer [76] proposed a theory to describe the condensation from a gas to a liquid in terms of mathematical clusters (for a review of cluster theory in simple fluids see [88]). The location for the divergence of the size of these clusters was interpreted as the condensation transition from a gas to a liquid. One of the major drawbacks of the theory was that the cluster number for some values of thermodynamical parameters could become negative. As a consequence the clusters did not have any physical interpretation [50]. This theory was followed by Frenkel's phenomenological model [54], in which the fluid was considered as made of non interacting physical clusters with a given free energy. This model was later improved by Fisher [50], who proposed a different free energy for the clusters, now called droplets, and consequently a different scaling form for the droplet size distribution. This distribution, which depends on two geometrical parameters, σ and τ, has the nice feature that the mean droplet size exhibits a divergence at the liquid-gas critical point. Interestingly the critical exponents of the liquid gas critical point can be expressed in terms of the two parameters, σ and τ, and are found to satisfy the standard scaling relations proposed at that time in the theory of critical phenomena.

Introduction

Fisher's droplet model was very successful, to describe the behavior of a fluid or of a ferromagnet near the critical point, in terms of geometrical clusters. However the microscopic definition of such a cluster in a fluid or ferromagnet was still a challenge. While the exact definition in a continuum fluid model is still an open problem, a proper definition in the Ising model or lattice gas model has been provided. A first attempt to define a cluster in the Ising model which had the same properties of Fisher's droplet model was to consider a cluster as set of parallel spins. In two dimensions, in fact, these clusters seemed to have the properties of Fisher's droplets, i. e. the mean cluster size of these clusters were found to diverge at the Ising critical point on the basis of numerical analysis [8]. This result was later proved rigorously [36,37]. However the critical exponents for the mean cluster size in $2d$ was found to be larger than the corresponding critical exponent of the susceptibility [100], contrary to the requirement of Fisher's droplet model. Moreover numerical simulations in $3d$ and analytical results on the Bethe lattice showed that the critical point and the percolation point of such clusters were different. It was clear then that the clusters made of nearest neighbors parallel spins were too big to describe correlated regions. A different definition of clusters was then proposed [30] obtained by breaking the clusters of parallel spins by introducing fictitious bonds with a probability p_b between parallel spins. The new clusters are defined as a maximal set of parallel spins connected by bonds. For a particular choice of $p_\mathrm{b} \equiv p = 1 - \mathrm{e}^{-2J/k_\mathrm{B}T}$ it was shown that these clusters (Coniglio–Klein droplets) have the same properties of Fisher's droplets, namely their size diverges at the Ising critical point with Ising exponents. Note that the bonds are only fictitious and do not change the energy of the spins. They only have the role of breaking the clusters made of parallel spins. Some years earlier Kasteleyn and Fortuin defined a random cluster model, obtained starting from an Ising model and by changing the spin interaction J in $J = \infty$, with probability p, and J into $J = 0$, with probability $1 - p$. They showed that the partition function of this modified model, called the random cluster model, coincides with the partition function of the original Ising model. In the random cluster model the clusters are defined as a maximal set of spins connected by infinite interactions. Although these clusters have the same properties of the droplet model, they were defined in the random cluster model, and for this reason these clusters were not associated with the droplets of the Ising model. It was only after Swendsen and Wang [99] introduced a cluster dynamics based on the Kasteleyn and Fortuin formalism, that it was formally shown [40] that the distribution of the Coniglio–Klein (CK) droplets are the same as the distribution of Kasteleyn–Fortuin (KF) clusters in the random cluster model. For this reason often the CK droplets and the KF clusters are identified, however the different meanings should be kept in mind.

A further development was obtained when the fractal structure of the droplets was studied not only for the

Ising model but for the full hierarchy of the q-state Potts model, which in the limit $q = 1$ gives the random percolation problem. It was shown that the critical droplets of the Potts model have the same structure, made of links and blobs, as found for the clusters in the random percolation problem. One of the consequences of this study was a better understanding of scaling and universality in terms of geometrical cluster and fractal dimension [74].

The cluster approach to the phase transitions leads also to a deeper understanding of why critical exponents do not depend on dimensionality above the upper critical dimension, and coincide with mean field exponents. It was in fact suggested [26] that, at least for random percolation, the mean field behavior is due to the presence of an infinite multiplicity of critical clusters at the percolation point. This suggests that similar results may be also extended to thermal problems.

Although the original interest in the field of correlated percolation was the study of critical phenomena in terms of geometrical concepts, later it was suggested that correlated percolation could be applied to the sol-gel transition, in particular when correlation was too large to be neglected. In many cases in fact the sol-gel transition, which is based on long range connectivity and percolation transition, interferes with large density fluctuation or critical point. The interplay between percolation points and critical points gives rise to interesting phenomena which are well understood within the concepts of correlated percolation. Correlated percolation has been studied also in systems with different types of long range correlation [108], and has been applied to many other fields such as nuclear physics [69], Gauge Theory [52] and O(n) models [7], fragmentation [15], urban growth [70], random resistor network [6], interacting colloids [12], biological models [1].

In Sect. "Random Percolation" we introduce random percolation concepts. In Sect. "Percolation in the Ising Model" in the context of the Ising model it is shown how clusters have to be defined in order to describe correlated regions corresponding to spin fluctuations. In Subsects. "Ising Clusters" and "Ising Droplets" the Ising clusters and droplets are respectively introduced, and in Subsect. "Ising Droplets Above $d = 4$" it is shown how the mapping between thermal properties and connectivity breaks down below T_c above $d = 4$. In Subsect. "Generalization to the q-State Potts Model" the results found for the Ising model are extended to the q-state Potts model, and in Subsect. "Fractal Structure in the Potts Model: Links and Blobs" the fractal structure is studied in terms of links and blobs. In Subsect. "Fortuin–Kasteleyn–Random Cluster Model" Fortuin–Kasteleyn–Random Cluster Model is presented, and the connection with the Coniglio–Klein droplets is further developed in Appendix A. In Sect. "Hill's Clusters" the possibility to extend the definition of droplets to simple fluids is discussed. In Sect. "Clusters in Weak and Strong Gels" the mechanism, leading to the formation of bound states in gelling systems, is considered, and in Sect. "Scaling Behavior of the Viscosity" the effect that finite bond lifetime has on the behavior of viscosity in weak or colloidal gels. Finally future directions and open problems are discussed in Sect. "Future Directions".

Random Percolation

In this section we define some connectivity quantities and present some results in the context of random percolation, which we will use in the following sections, where the correlated percolation will be presented.

Consider a d-dimensional hypercubic lattice of linear dimension L. Suppose that each edge has a probability p of being occupied by a bond. For small values of p, small clusters made of sites connected by nearest-neighbor bonds are formed. Each cluster is characterized by its size or mass s, the number of sites in the cluster. For large values of p in addition to small clusters we expect a macroscopic cluster that connects the opposite boundaries. This spanning cluster becomes infinite as the system size becomes infinite. For an infinite system there exists a percolation threshold p_c below which only finite clusters are present.

In order to describe the percolation transition [13,58,95], one defines: an order parameter, $P_\infty(p)$, as the density of sites in the infinite cluster, the mean cluster size, $S(p)$, of the finite clusters, and the average number of clusters, $K(p)$.

These quantities can be related to the average number of clusters of s sites per site, $n(s, p)$, and near the percolation threshold the critical behavior is characterized by critical exponents:

$$K(p)|_{\text{sing}} = \sum n(s,p)|_{\text{sing}} \sim |p - p_c|^{2-\alpha_p} , \tag{1}$$

$$P_\infty(p) = 1 - \sum sn(s,p) \sim \begin{cases} 0 & \text{if } p < p_c \\ (p-p_c)^{\beta_p} & \text{if } p > p_c , \end{cases} \tag{2}$$

$$S(p) = \sum s^2 n(s,p) \sim |p - p_c|^{-\gamma_p} , \tag{3}$$

where the sum is over all finite clusters, and in Eq. (1) only the singular part has been considered. Finally one can define the pair connectedness function p^f_{ij} as the probability that *i* and *j* are in the same finite cluster through

$$\xi^2(p) = \frac{\sum r_{ij}^2 p_{ij}^f}{\sum p_{ij}^f} \,. \tag{4}$$

The connectedness length, $\xi(p)$, which is the critical radius of the finite clusters, diverges as

$$\xi \sim |p - p_c|^{-\nu_p} \,. \tag{5}$$

The critical exponents defined in Eqs. (1)–(4) are not all independent. Scaling relations can be derived among them as for ordinary second phase transitions. These scaling laws are intimately related to the property of the incipient infinite cluster of being a self similar fractal [74] to all length scales. The mass, s^*, of a typical cluster of linear dimension, ξ, scales as $s^* \sim \xi^{D_p}$, where D_p is the fractal dimension of the cluster.

Scaling and Hyperscaling

To obtain scaling laws, following Kadanoff's original idea, we perform [26] the following three steps: (i) divide the system into cells of linear dimension *b*, (ii) coarse grain by some suitable rule, (iii) rescale the lengths by a factor *b*. The result is a renormalized system where the size of the large clusters *s* has been reduced by factor b^{D_p} and all lengths by a factor *b*:

$$L' = \frac{L}{b} \,, \quad \xi = \frac{\xi'}{b} \,, \quad s' = \frac{s}{b^{D_p}} \,. \tag{6}$$

Assuming that the large clusters do not interpenetrate, the sum over the large clusters in an interval between $(s, s + \Delta s)$ must be the same before and after rescaling, i. e.

$$N(s,\xi)\Delta s \sim N(s',\xi')\Delta s' \tag{7}$$

where $N(s,\xi)/L^d = \overline{n}(s,\xi)$ is the number of clusters of *s* sites per unit volume. Dividing by the volume L^d, from (6) we obtain

$$\overline{n}(s,\xi) = b^{-d-D_p}\overline{n}(sb^{-D_p} \,,\ \xi b^{-1}) \,. \tag{8}$$

Choosing $b = s^{1/D_p}$ from (8) we obtain

$$n(s,p) = s^{-\tau_p} f((p - p_c)s^{\sigma_p}) \tag{9}$$

where $n(s,p) = \overline{n}(s,\xi)$ and

$$\tau_p = \frac{\mathrm{d}}{D_p} + 1, \quad \sigma_p = \frac{1}{\nu_p D_p} \,. \tag{10}$$

Equation (9) exhibits the scaling form postulated by Stauffer [13,58,95]. From (1), (2) and (10) we have:

$$2-\alpha_p = \frac{\tau_p - 1}{\sigma_p} \,, \quad \beta_p = \frac{\tau_p - 2}{\sigma_p} \,, \quad -\gamma_p = \frac{\tau_p - 3}{\sigma_p} \tag{11}$$

and

$$\tau_p = 2 + \frac{\beta_p}{\beta_p + \gamma_p} \,, \quad \sigma_p = \frac{1}{\beta_p + \gamma_p} \,, \tag{12}$$

from which the following scaling relations are obtained:

$$\alpha_p + 2\beta_p + \gamma_p = 2 \,, \tag{13}$$

$$\frac{1}{\nu_p}(\beta_p + \gamma_p) = D_p \,. \tag{14}$$

From (10), (11) one can also find relations which contain the Euclidean dimensionality *d* called hyperscaling relation:

$$2 - \alpha_p = \nu_p d \,, \tag{15}$$

$$d - \frac{\beta_p}{\nu_p} = D_p \,. \tag{16}$$

Equation (16) was originally suggested in [68]. In 2*d* exact results give $\tau_p = 187/91$ and $\sigma_p = 36/91$, and in 3*d* the best estimates $\tau_p \simeq 2.18$, $\sigma_p \simeq 0.45$. From mean field theory [57] we know that for any *d* above the upper critical dimension $d_c = 6$, the critical exponents coincide with the mean field ones, namely $-\alpha_p = \beta_p = \gamma_p = 1$, $\nu_p = \sigma_p = 1/2$ and $\tau_p = 5/2$. These exponents satisfy the scaling relation (13), but fail to satisfy the hyperscaling relation (15) except for $d = 6$.

Moreover, while Eq. (14), for any $d > 16$, shows that the fractal dimension is stacked at the value $D_p = 4$, the hyperscaling relation (16) breaks down for $d > 6$.

Breakdown of Hyperscaling

By following a less conventional scaling approach, here we want to propose a geometrical interpretation of hyperscaling, why it breaks down above d_c, and why the hyperscaling breakdown occurs when the mean field becomes valid [26].

Let us assume that the singular behavior comes only from the critical clusters. Say N_ξ is the number of such clusters in a volume of the order ξ^d. The singular part of the cluster number is given by

$$\frac{N_\xi}{\xi^d} \sim \xi^{\frac{\alpha_p - 2}{\nu_p}} \,. \tag{17}$$

At the same time, the density of sites in the infinite cluster $P_\infty \sim |p-p_c|^{\beta_p}$ scales as the total mass of the spanning clusters $N_\xi s^*$ in a volume of linear dimension ξ divided by the volume ξ^d, namely

$$\frac{N_\xi \xi^{D_p}}{\xi^d} \sim \xi^{\frac{-\beta_p}{\nu_p}} , \tag{18}$$

where we have used $s^* \sim \xi^{D_p}$. Similarly the mean cluster size:

$$\frac{N_\xi \xi^{2D_p}}{\xi^d} \sim \xi^{\frac{\gamma_p}{\nu_p}} . \tag{19}$$

These equations lead to the scaling relations, Eqs. (13) and (14). Now if N_ξ is of the order of unity, we recover the hyperscaling relations [48], Eqs. (15) and (16), while if N_ξ diverges hyperscaling breaks down. We know that for dimension d above $d_c = 6$, $n(s,p) \sim s^{-5/2} e^{-(p-p_c)^2 s}$ for large s. Therefore $N_\xi = \xi^d \sum n(s,p) \sim \xi^{d-6}$, where $\xi \sim |p-p_c|^{-1/2}$. This calculation shows that, above d_c, N_ξ diverges and hyperscaling breaks down, and from Eqs. (17) and (18) the hyperscaling relations are replaced by $2-\alpha_p = 6\nu_p$ and $D_p = 6 - \beta_p/\nu_p$, which in fact are satisfied for mean field exponents.

The more standard scaling approach of the previous section must be modified taking into account that for $d > 6$ the large number of clusters will be reduced by a factor b^{6-d}, then Eq. (7) will be modified as $N(s,\xi)\Delta s = b^{6-d} N(s', \xi')\Delta s'$ which still leads to all the Eqs. (7)–(14), except that d is replaced everywhere by 6. In particular, both Eqs. (14) and (16) give a fractal dimension $D_p = 4$.

The multiplicity of infinite clusters above d_c was numerically shown in [42,53]. The average (finite) number N_ξ of distinct clusters below d_c have been estimated theoretically and calculated numerically [3,63,94].

Consider now a critical cluster for $d > d_c$ just below p_c and its center of mass, 0. Say ξ_1 is the distance from 0, below which the cluster has not been penetrated by the other critical clusters. This length can be obtained by equating the mass density inside the region of radius ξ_1 to the mass density inside the region of radius ξ, $N_\xi \xi^{4-d} = \xi_1^{4-d}$, which gives $\xi_1 \sim \xi^{2/(d-4)}$.

If $\rho(r)$ is the density profile defined as the mass density of all the critical clusters at a distance r from 0, we expect that the density profile behaves as a power law r^{d-4} for $r < \xi_1$, as it should be for an object with fractal dimension $D_p = 4$ and as a constant for $r > \xi_1$ due to the penetration of the other critical clusters. Consequently we can make the following scaling Ansatz [2]:

$$\rho(r) = \frac{1}{r^{d-4}} f\left(\frac{r}{\xi_1}\right) , \tag{20}$$

where $f(x) \sim$ const for $x < 1$ and $\sim x^{d-4}$ for $x > 1$.

In conclusion, while for $d < 6$ the density of the order parameter fluctuates over a distance of the order ξ, for $d > 6$, where the mean field holds, the fluctuations are damped by the presence of infinitely many interpenetrating clusters, and the density of the order parameter crosses over from a power law (fractal) regime to a homogeneous regime at a distance $\xi_1 \ll \xi$.

The mean field solution is therefore a consequence of the presence of infinitely many interpenetrating clusters which suppress the spatial fluctuation of the order parameter. The condition for the validity of mean field theory is then given by $N_\xi \gg 1$. Using Eqs. (18) and (14) this condition implies

$$N_\xi^{-1} \sim \frac{\xi^{\frac{\gamma_p}{\nu_p}}}{\xi^d \xi^{\frac{-\beta_p}{\nu_p}}} \sim \frac{\langle \Delta M^2 \rangle}{\xi^d M^2} \ll 1 , \tag{21}$$

where M and $\langle \Delta M^2 \rangle$ are the order parameter and the fluctuations of the order parameter (here we used that the mean cluster size $S(p)$ has the same critical behavior as the fluctuations of the order parameter [18,34]). Interestingly enough Eq. (21) coincides with Ginzburg criterion for the validity of mean field theory.

Cluster Structure

Nodes and Links In the previous sections we have shown that the Incipient Infinite Cluster (IIC) is a fractal. Here we want to show in more details the internal structure of the IIC. A very useful nodes and links picture for the infinite cluster just above p_c was introduced by Skal and Shklowskii [89] and de Gennes [44]. In this picture the infinite cluster consists of a superlattice made of nodes, separated by a distance of the order of ξ, connected by macrobonds. Just below p_c the structure of the very large cluster, the IIC, was expected to have the same structure as the macrobonds.

Later on, in 1977, Stanley [90] made the important observation that in general for each configuration of bonds at p_c the IIC can be partitioned into three categories. By associating an electric unit resistance to each bond, and applying a voltage between the ends of the cluster, one distinguishes the dangling bonds which do not carry current (yellow bonds). The remaining bonds are the backbone bonds. The backbone can be partitioned in singly connected bonds (red bonds) and all the others, the multiply connected bonds, which lump together in "blobs" (blue bonds). The red bonds, which carry the whole current, have also the property that if one is cut the cluster breaks into two parts. This partition in three types of bonds is very general and can be done for any cluster or aggregate.

The next major problem was to determine whether the blobs are or not relevant. In the nodes and links picture the assumption is that the blobs are irrelevant and only links are relevant. A further elaboration [90] assumed that the backbone close to p_c would reduce to a self-avoiding walk chain, which implies that the blobs are not relevant. This self-avoiding walk Ansatz received a large amount of attention, since it predicted a value for the crossover exponent of the dilute Heisenberg ferromagnetic model near the percolation threshold in $2d$, in good agreement with the experimental data, although the prediction for the dilute Ising crossover exponent did not agree as well with the data [9].

Syerpinsky Gasket: A Model Without Links In 1981 a completely alternative model was proposed by Gefen et al. [55]. Based on the observation that in a computer simulation the red bonds were hardly seen, they proposed an alternative model, the Syerpinsky gasket, that represents the opposite extreme of the nodes and links picture. It has a self-similar structure but only multiply connected bonds are present. A great advantage of this model is that it can be solved exactly. It also gives good prediction for the fractal dimension of the backbone, but it fails to predict the correct value for the dilute Ising crossover exponent [9].

Nodes, Links and Blobs Motivated by all these conflicting models, some rigorous results were presented which led unambiguously to the nodes, links and blobs picture of the infinite cluster [23,24], in which both links and blobs are relevant below $d = 6$, while only links are relevant above $d = 6$ or in the mean field. In particular the following relation was proven for any p and for any lattice in any dimension:

$$p \frac{\mathrm{d} p_{ij}}{\mathrm{d} p} = \lambda_{ij} \qquad (22)$$

where p_{ij} is the probability that i and j are connected, λ_{ij} is the average number of red bonds between i and j, such that if one is cut, i and j would have been disconnected. From Eq. (22) it is possible to calculate the average number L_{ij} of red bonds between i and j under the condition that i and j are in the same cluster:

$$L_{ij} = \frac{\lambda_{ij}}{p_{ij}} . \qquad (23)$$

From the scaling form of $p_{ij} = r_{ij}^{-d+2-\eta_p} f(r_{ij}/\xi)$ it follows

$$L_{ij} = r_{ij}^{\frac{1}{\nu_p}} f_1\left(\frac{r_{ij}}{\xi}\right) , \qquad (24)$$

where $f_1(x)$ is related via Eqs. (22) and (23) to $f(x)$ and goes to a constant for $x \ll 1$. In particular, by putting $r_{ij} = \xi$ in Eq. (24) we obtain

$$L_{\mathrm{R}} \sim \xi^{\frac{1}{\nu_p}} , \qquad (25)$$

where $L_{\mathrm{R}} \equiv L(r_{ij} = \xi)$ is the average number of red bonds between two points separated by a distance of the order of ξ. From Eq. (25) it follows that the fractal dimension of the red bonds is $D_{\mathrm{R}} = 1/\nu_p$. An immediate consequence is that not only the red bonds are relevant but also the number of bonds L_{B} in the blobs diverge. For more details, see [23,24]. Later Eq. (25) was confirmed numerically by Pike and Stanley [83] in $d = 2$. Although the links are much less in number than the backbone bonds, they can be detected experimentally, in fact it can be shown [23,25] that only the links determine the critical behavior of the dilute Ising model at p_c leading to a crossover exponent 1 in any d. While for a dilute Heisenberg system the crossover exponent is related to the resistivity exponent, in agreement with the experimental data of [9].

In conclusion we can write the following relations

$$y_{\mathrm{H}} = D_p , \quad y_{\mathrm{T}} = D_{\mathrm{R}} \qquad (26)$$

where $y_{\mathrm{H}} = d - \beta_p/\nu_p$ is the so-called magnetic field scaling exponent and $y_{\mathrm{T}} = 1/\nu_p$ is the thermal scaling exponent in the renormalization group language. This result is quite interesting as it shows that the scaling exponents can be expressed in terms of geometric quantities: The fractal dimension D_p of the entire incipient infinite cluster, and the fractal dimension D_{R} of the subset made of red bonds.

Surfaces and Interfaces

The study of the structure of the surfaces and interfaces of the large clusters below p_c has not received as much attention as the study of the internal structure of the IIC. This problem is relevant to the study of the dielectric constant of random composite materials, the viscosity of a gel, the conductivity of a random superconducting network, and the relative termite diffusion model.

For simplicity, let us consider a random superconducting network in which superconducting bonds are present with probability p and normal bonds carrying a unit resistance with probability $1 - p$. For small values of p we have finite superconducting clusters in a background of normal resistor. As $p \to p_c$, the superconductivity Σ diverges. For a finite cell of linear dimension L just below p_c, the typical configurations are characterized by two very large clusters almost touching, each one attached to one of two opposite faces. Inside these clusters there are islands of normal resistors. If a unit voltage is applied between the opposite face of the hypercube, there is no current flowing through the bonds in the island. We call these "dead" bonds, analogous to the dead ends of the percolating cluster. The remaining normal bonds connect one superconducting cluster to the other. These bonds are made of "bridges", also called "antired" bonds, which have the property that if one is replaced by a superconducting bond, a percolating superconducting cluster is formed, and the remaining multiple "connecting" bonds. Similar to the red bonds, it can be proved [26] that the fractal dimensionality of the antired bonds is $1/\nu_p$. The proof is based on the following relation which can be proved for any lattice in any dimension

$$(1-p)\frac{\mathrm{d}p_{ij}}{\mathrm{d}p} = \mu_{ij} , \tag{27}$$

where p_{ij} is the pair connectedness function (the probability that sites i and j belong to the same cluster) and μ_{ij} is the average number of antired bonds between i and j. These are defined as non-active bonds, such that if one is made active, i and j become connected.

The above considerations suggest that just below p_c the system can be imagined as a superlattice made of large critical clusters whose centers are separated by a distance of the order ξ. The surfaces of these clusters almost touch, and are connected by bridges made of single bonds and other paths made of more than one bond [33].

Finally we mention the following result which relates the size of the critical cluster s^* and the size of the entire perimeter t^* [13,58,95]

$$t^* = \frac{1-p}{p}s^* - As^{*\sigma_p} , \tag{28}$$

where $\sigma_p = 1/(\nu_p D_p)$ is the critical exponent which appears in the cluster number Eq. (9) and A is a constant. The last term $s^{*\sigma_p}$ which appears also in Fisher's droplet model [50] is usually interpreted as the surface of the droplet. However, if it was a surface, σ_p should satisfy the following boundary $(d-1)/d \le \sigma_p \le 1$. The upper bound corresponds to the fully rarefied droplets and the lower bound to compact droplets. Surprisingly enough for the percolation problem σ_p is strictly smaller than $(d-1)/d$. This paradox can be solved by using a result [25], which shows that $As^{*\sigma_p}$ is equal to a number of antired bonds between critical clusters separated by a distance of order ξ. Since the subset of antired bonds is only a subset of the entire perimeter, it explains why $\sigma_p < (d-1)/d$. This result gives the best geometrical interpretation of the thermal scaling exponent y_{T}. It in fact shows that $y_{\mathrm{T}} = D_{\mathrm{AR}}$, where D_{AR} is the fractal dimension of the antired bonds namely that part of the surface which contributes to the surface tension.

Percolation in the Ising Model

In this section we want to extend the percolation problem to the case in which the particles are correlated. The simplest model to consider is the lattice gas or Ising model. In the following we will use the Ising terminology. We know that the Ising model exhibits a thermodynamic transition for zero external field, $H = 0$, at a critical temperature T_c. The question that we ask is how the percolation properties are modified due to the presence of correlation. We first consider the case when the clusters are made of nearest-neighbor down spins (Subsect. "Scaling and Hyperscaling"). Later in Subsect."Breakdown of Hyperscaling" we will modify the cluster definition in such a way that these new clusters describe the thermal fluctuations namely we require that the clusters satisfy the same properties as the droplets in Fisher's droplet model [50]. Namely, i) the size of the clusters must diverge at the Ising critical points, ii) the linear dimension of the clusters must diverge with the same exponent as the correlation length, and iii) the mean cluster size must diverge with the same exponent as the susceptibility.

These conditions are satisfied if the cluster size distribution for zero external field has the following form

$$n(s,T) = s^{-\tau} f((T-T_c)s^{\sigma}) . \tag{29}$$

The parameters σ and τ are related to critical exponents α, β and γ through Eqs. (10) and (11), where now α, β and γ are the Ising critical exponents. In particular for $d = 2$, $\sigma = 8/15 \simeq 0.53$ and $\tau = 31/15 \simeq 2.07$, and for $d = 3$, $\sigma \simeq 0.64$ and $\tau \simeq 2.21$.

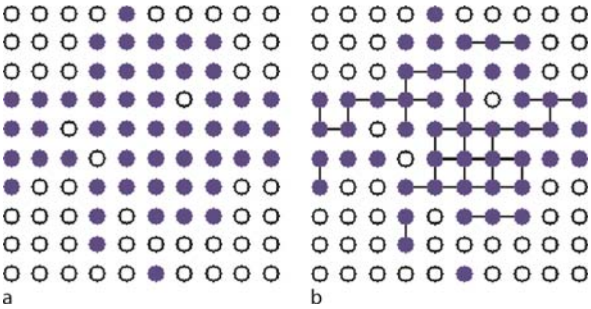
Ising Clusters

The Hamiltonian of the Ising model is given by:

$$\mathcal{H} = -J\sum_{\langle ij\rangle} S_i S_j - H\sum_i S_i \tag{30}$$

where $S_i = \pm 1$ are the spin variables, J is the interaction between two nearest-neighbor (nn) spins and H is the magnetic field.

From the thermodynamic point of view the only quantities of interest are those which can be obtained from the free energy and those were the only quantities that Onsager was concerned with in his famous solution of the 2d Ising model. However one can look at the Ising model from a different perspective by studying the connectivity properties using concepts such as clusters which have been



Correlated Percolation, Figure 1

a Ising configuration at T_c: "down" spins are represented by *filled circles*. **b** Correct clusters are obtained from the configuration given in (**a**) by putting bonds between occupied sites with probability $p = 1 - e^{-2\beta J}$

systematically elaborated in percolation theory [92]. There are two reasons for approaching the problem also from the connectivity point of view. One reason is that it gives a better understanding of the mechanism of the phase transition [50]. In fact, concepts like universality and scaling have been better understood in terms of geometrical clusters and fractal dimensions [27]. A second reason is that there are physical quantities amenable to experimental observations, which are associated to the connectivity properties and cannot be obtained from the free energy. It is very important to note however that the definition of connectivity, and therefore the definition of the cluster, is not always the same, but may depend on the particular observable associated with it.

In the Ising model, for a given configuration of spins it is rather natural to define a cluster as a maximal set of *nn* down parallel spins[1] (Fig. 1). For some time these clusters were believed to be responsible for the correlations present in the Ising model. This idea was also based on numerical results which showed evidence that in two dimensions the mean cluster size diverges at the thermal critical point [8]. However the idea that the clusters could describe thermal correlations was definitively abandoned when it was shown, by numerical simulations in the three-dimensional Ising model [78] and by exact solution on the Bethe lattice [22], that the percolation point appeared in the low density phase of down spins on the coexistence curve at a temperature T_p before the critical point T_c is reached ($T_p < T_c$).

Later it was suggested by topological arguments [21] that only in two dimensions does the critical point coincide with the percolation point, but not necessarily in higher dimensions. The arguments followed two steps: in the first step it was argued that an infinite cluster of up spins is a necessary condition for having a spontaneous magnetization. This implies a percolation transition of down spins on the coexistence curve $T_p \leq T_c$, in the second step it was argued that due to topological reasons in two dimensions it is not possible to have an infinite cluster of up spins coexisting with an infinite cluster of down spins, which implies $T_p \geq T_c$. Combining with the previous inequalities one obtains in two dimensions $T_p = T_c$.

Later these results were proven rigorously [36,37] along with many other results relating connectivity and thermodynamic quantities. For more details we refer to the original papers.

It is clear that the Ising clusters, defined as a group of *nn* parallel spins, do not have the property of describing correlated regions corresponding to spin fluctuations as originally expected. In fact even in two dimensions, where the thermal critical point coincides with the percolation point, the Ising clusters were not suitable for such description. Series expansion showed that the mean cluster size diverges with an exponent, $\gamma^* = 1.91 \pm 0.001$, rather different from the susceptibility exponent, $\gamma = 1.75$ [100]. Later it has been shown exactly that $\gamma^* = 91/48$ [97].

Ising Droplets

From the properties mentioned in Subsect. "Ising Clusters", it appears that the Ising clusters are too big to describe the proper droplets. The reason is that there are two contributions to the Ising clusters. One is due to correlations but there is another contribution purely geometrical due to the fact that two *nn* spins even in the absence of correlation have a finite probability of being parallel. The last contribution becomes evident in the limit of infinite temperature and zero external field. In this case, in fact, although there is no correlation and the susceptibility is zero, the cluster size is different from zero. In fact in 3*d* at infinite temperature there is even an infinite cluster of "up" and "down" spins.

Binder [8] proposed to cut the infinite cluster in order to have $T_p = T_c$ in $d = 3$, but he did not give the microscopic prescription to do it. Later Coniglio and Klein [30] proposed to reduce the cluster size by introducing fictitious bonds between *nn* parallel spins with probability p_b (Fig. 1). These new clusters are made of *nn* parallel spins connected by bonds. The original Ising cluster will either reduce its size or will break into smaller clusters. If $p_b = 1$, we obtain the Ising clusters again. This case is known as the site correlated percolation problem because one looks

[1] The Ising Hamiltonian, Eq. (30), is equivalent to the lattice gas Hamiltonian $\mathcal{H}_{LG} = -J' \sum_{\langle ij \rangle} n_i n_j - \mu \sum_i n_i$, with $n_i = (1 - S_i)/2$, $J' = 4J$ and $\mu = 2H - 4J$. In the lattice gas terminology an Ising cluster is a maximal set of *nn* occupied sites.

at the properties of the Ising clusters just as in the random percolation problem. The main difference is that in random percolation the occupied sites are randomly distributed, while in this case the down (or up) spins are correlated according to the Ising Hamiltonian. In the infinite temperature limit one recovers random percolation. The case $p_b \neq 1$ is called site-bond correlated percolation [39].

A Hamiltonian formalism was proposed to study site correlated percolation [79]. This formalism was generalized in [30] to study site-bond correlated percolation. In this case for the zero external field the Hamiltonian is given by the following dilute Ising s-state Potts Model (DIPM)[2]:

$$-\mathcal{H}_{\mathrm{DP}} = J_b \sum_{\langle ij \rangle} (\delta_{\sigma_i \sigma_j} - 1)(S_i S_j + 1) + J \sum_{\langle ij \rangle} S_i S_j , \quad (31)$$

where $\sigma_i = 1, \ldots, s$ are Potts variables and the sum is over all nearest neighbor sites. In the same way as the s-state Potts model in the limit $s = 1$ [110] describes the random bond percolation model, the DIPM describes percolation in the Ising model where the clusters are made of parallel spins connected by bonds with probability, $p_b = 1 - \mathrm{e}^{-2\beta J_b}$.

In particular the average number of clusters G, that plays the role of the free energy in the percolation problem is given by $G = \mathrm{d}F/\mathrm{d}s|_{s=1}$, where

$$-\beta F = \lim_{N \to \infty} \frac{1}{N} \ln \left(\sum_{\{\sigma_i S_i\}} \mathrm{e}^{-\beta \mathcal{H}_{\mathrm{DP}}} \right) . \quad (32)$$

At that time the DIPM was investigated in a different context by Berker et al. [81]. The model exhibits the interesting properties that by choosing $J_b = J$ it coincides with a pure $s+1$-state Potts model. Therefore in the limit $s = 1$ the DIPM coincides with the $s = 2$ Potts model namely with the Ising model. Consequently F becomes the Ising model free energy and G has a singularity at the Ising critical point. This argument immediately suggested that the site-bond correlated percolation for $J_b = J$ namely with the bond probability given by

$$p_b \equiv p = 1 - \mathrm{e}^{-2\beta J} , \quad (33)$$

should reproduce the same critical behavior of the Ising model. Namely the percolation quantities become critical at the Ising critical point in the same way as the corresponding thermal quantities.

In fact using real space renormalization group arguments, it was possible to show that the size of the clusters of parallel spins connected by bonds with probability, p_b, given by Eq. (33), diverges at the Ising critical point with Ising exponents, exhibiting thus the same properties as the droplets in Fisher's model. These clusters were called droplets to distinguish them from the Ising clusters.

Droplets in Two and Three Dimensions

This site-bond correlated percolation problem has been studied by real space renormalization groups in two dimensions [30,35], by ϵ expansion, near six dimensions [31] and by Monte Carlo in two and three dimensions [64,82,85,92].

The renormalization group analysis shows that in $2d$ the Ising critical point is a percolation point for down or up spins connected by bonds for all values of bond probability such that $1 \leq p_b < 1 - \mathrm{e}^{-2\beta J}$. The fractal dimension $D^* = (\gamma^*/\nu + 2)/2 = 187/96$ [97] being higher than the fractal dimension $D = (\gamma/\nu + 2)/2 = 15/8$ for the value of $p_b \equiv p = 1 - \mathrm{e}^{-2\beta J}$.

In the renormalization group language this means that there are two fixed points, one corresponding to the universality class of the Ising cluster, the other one corresponding to the droplets. In the first one, the variable J_b is irrelevant, namely the scaling exponent associated with it, $y_b < 0$. In the second fixed point associated with the droplets instead $y_b > 0$. The result is that the Ising critical point is a percolation point for a range of values of p_b, at the first sight seems counter-intuitive. In fact if the Ising critical point corresponds to the onset of percolation for Ising clusters ($p_b = 1$), one would expect that for $p_b < 1$ the clusters would not percolate anymore. The puzzle can be clarified by studying the fractal structure of the Ising clusters and the droplets at T_c [27]. In fact it can be shown that y_b is the scaling exponent of the red bonds, namely $L_R \sim l^{y_b}$ where L_R is the number of red bonds between two connected sites separated by a distance of the order l, consequently the droplets, characterized by $y_b > 0$, are made of links and blobs, like in random percolation. Due to the presence of links the cluster breaks apart and does not percolate anymore as the bond probability decreases. On the contrary the Ising clusters ($p_b = 1$), characterized by $y_b < 0$, are made only of blobs and no links, therefore by decreasing the bond probability the infinite cluster does not break and still percolates, until $p_b = p$.

In $3d$ at the Ising critical point, T_c, there is an analogous line of anomalous percolation points for clusters of down spins connected by bonds, for all values of bond probability such that $1 \leq p_b < 1 - \mathrm{e}^{-2\beta J}$, although the

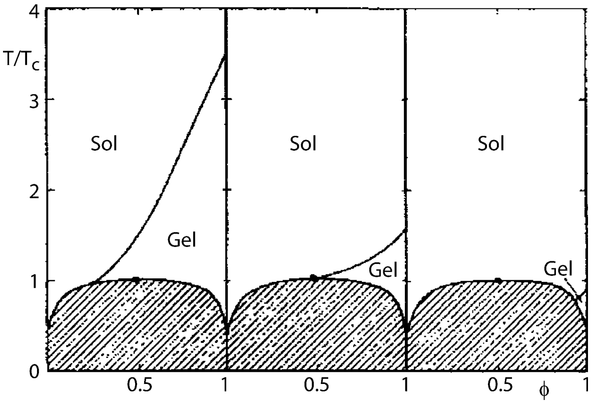
[2] Originally in [30] the Hamiltonian of the DIPM, $\mathcal{H}_{\mathrm{DP}}$, was expressed in terms of the lattice gas variables n_i, and the Ising droplets were defined as nn occupied sites connected by bonds, corresponding to nn down spins.

probability P_∞ for a down spin to be in the infinite cluster is different from zero. More precisely the quantity $p_{ij} - P_\infty^2$ decays as a power law, where p_{ij} is the probability that i and j are connected. For more details see [29]. As p_b decreases towards $p = 1 - e^{-2\beta J}$ there is a crossover towards a different power law characterized by the Ising exponent, while P_∞ goes to 0.

Droplets in an External Field

By keeping the same definition of droplets given above, in the case of an Ising model in an external field $H > 0$ one finds a phase diagram in the H, T plane or in the M, T plane, with a percolation line of "down" spins ending at the Ising critical point (see Fig. 2). Along the percolation line one finds critical exponents in the universality class of random percolation with a cross-over to Ising critical exponents as the Ising critical point is approached. The Ising critical point being a higher order critical point for the percolation transition. This percolation line, also known as the Kertesz line, has received some attention [66,93,95,105] (see for more details the review by Sator [88]). Although the Ising free energy has no singularity along this line some physical interpretation is given to the Kertesz line [17].

On the other hand this line disappears if the droplet definition is modified in the presence of an external field [40,104], according to Kasteleyn and Fortuin formalism [65] and the Swendsen and Wang approach [99]. In this approach the field is treated as a new interaction between each spin and a ghost site. Consequently for positive H (negative H) an "up" ("down") spin can be connected to the ghost spin with a probability $p_H = 1 - e^{-2\beta|H|}$. Droplets now are defined as a maximal set of spins connected by bonds where as before the bonds between nearest-neighbor parallel spins have probability p_b given by Eq. (33) and p_H between spins and the ghost spin. Note that two far away spins can be easily connected through the ghost spin. In this way the presence of a positive (negative) magnetic field implies always the presence of an infinite cluster of "up" ("down") spins.



Correlated Percolation, Figure 2
Monte Carlo simulations of the 3*d* lattice gas model for three values of the bond probability $p_b = 1 - e^{-2c\beta J}$ with the constant $c = 2.25, 1, 0.564$ from *left* to *right*. Φ is the density of down spins. The Gel and the Sol indicates the percolation and non percolation phase. From [59]

Exact Relations Between Connectivity and Thermal Properties

Interestingly it was also shown [40] that the droplets so defined have the same statistics as the clusters in the random cluster model introduced by Kasteleyn and Fortuin (KF) [65] (see Subsect. "Fortuin–Kasteleyn–Random Cluster Model"), although the CK droplets and the KF clusters have a different meaning. Using the relations between the connectivity properties of the random cluster model and the thermal properties of the Ising model, it was finally possible to prove that in any dimension and for any temperature T and external field $H > 0$, provided that the extension of the droplet definition in the external field is considered, the following relations between connectivity and thermal properties hold [40]:

$$\begin{cases} \rho_\infty = m \\ p_{ij} = g_{ij} \end{cases} \tag{34}$$

where ρ_∞ is the density of up spins in the percolating droplet, m is the magnetization per site, p_{ij} is the probability that i and j are connected (through both finite or infinite droplet) and $g_{ij} = \langle S_i S_j \rangle$.

In particular, for $T > T_c$ and zero external field $H \to 0$, we have that the magnetization $m = 0$ and g_{ij} coincides with the spin-spin pair correlation function. Consequently $\rho_\infty = 0$, namely the probability for a spin to be in an infinite droplet is zero, and therefore p_{ij} coincides with the probability that two spins i and j are in the same finite droplet. For $T < T_c$ instead we have $\rho_\infty = m > 0$, and $p_{ij} = p_{ij}^f + p_{ij}^\infty$, where p_{ij}^f (p_{ij}^∞) is the probability that spins in i and j are in a finite (infinite) droplet. From Eq. (34) it follows for $T < T_c$:

$$p_{ij}^f + p_{ij}^\infty - \rho_\infty^2 = \langle S_i S_j \rangle - m^2 \,. \tag{35}$$

By summing over i and j we have

$$S + (\Delta\rho_\infty)^2 = \chi \,, \tag{36}$$

where S is the mean cluster size of the finite clusters, $(\Delta\rho_\infty)^2$ is the fluctuation of the density of the infinite cluster and χ is the susceptibility. These exact results show that above T_c mean cluster size and susceptibility coincide, while below T_c there are two contributions to the susceptibility, one due to the mean cluster size and the second related to the fluctuation of the density of the infinity cluster. Monte Carlo calculations [85] show that both terms have the same critical behavior as also occurs in random percolation [18,34], so the mean cluster size S diverges like the susceptibility. We expect that this is the case for dimensions up to $d = 4$, the upper critical dimensionality of the Ising model. In the mean field, as we will see, the mean cluster size below T_c diverges with an exponent different from the susceptibility.

One very interesting application based on the KF approach was produced by Swendsen and Wang [99,106], who elaborated a cluster dynamics which drastically reduced the slowing down near the critical point of the Ising and Potts model (see also [109] for further developments).

The droplet definition can be extended to the *nn* antiferromagnetic Ising model [4] and to the Ising model with any ferromagnetic interaction J_{ij} between sites i and j [64]. In this case the CK clusters are defined as set of parallel spins connected by bonds present between i and j with probability $p_{ij} = 1 - \exp[-2\beta J_{ij}]$. It can be shown that also in this case the relations (34) between connectivity and thermal quantities hold.

Ising Droplets Above $d = 4$

In Subsect. "Ising Droplets" we have reported the relations Eqs. (34) and (35), which are exact and are valid in any dimension including the mean field. As a matter of fact in the mean field the percolation order parameter and the magnetization are identical and go to zero with the exponent $\beta = 1/2$, while the mean cluster size above T_c coincides with the susceptibility and diverges with the exponent $\gamma = 1$. The same is true for the connectedness length above T_c, which coincides with the correlation length, and diverges with an exponent $\nu = 1/2$. However below T_c the mean cluster size diverges with an exponent $\gamma' = 1/2$ and the correlation length with an exponent $\nu' = 1/4$. The result is a consequence that the two terms in Eq. (35), the probability that two sites are in the same finite droplet, p^f_{ij}, and the correlation of the infinite droplet density at site i and j, $p^\infty_{ij} - \rho^2_\infty$, do not scale in the same way, giving rise to two lengths, diverging respectively with exponents ν' and ν.

These somehow anomalous results are probably a consequence that the Ising model has an upper critical dimension $d_c = 4$ while the DIPM which describes the droplet problem has an upper critical dimension $d_c = 6$ [31]. In fact there are arguments that for $4 \le d \le 6$ below T_c the critical exponents are $\nu' = 1/(d-2)$, $\gamma' = 2/(d-2)$, $\beta = 1/2$, $\eta = 0$ and fractal dimension $D_p = 1/2(d+2)$, with an upper critical dimension $d_c = 6$. Of course for $T > T_c$ the exponents are $\gamma = 1$, $\nu = 1/2$ and $\eta = 0$.

Due to the breakdown of the mapping between thermal fluctuations and mean cluster size below T_c above $d = 4$, it is not possible to extend easily the geometrical picture, employed in random percolation, to explain the breakdown of hyperscaling in the Ising model. For a study of droplets inside the metastable region see [5].

Generalization to the q-State Potts Model

All the results found for the Ising case have been extended [32] to the q-state Potts model. This model is defined by the following Hamiltonian:

$$-\mathcal{H}_q = qJ \sum_{\langle ij \rangle} \delta_{\sigma_i \sigma_j} , \qquad (37)$$

where the spin variables σ_i can assume q values, $\sigma_i = 1, \ldots, q$. This model coincides with the Ising model for $q = 2$, reproduces the random percolation problem in the limit $q = 1$ and the tree percolation model in the limit $q = 0$ [110]. The geometrical approach developed in the previous sections for the Ising model, can be extended to the q-state Potts model. In particular one can define the site-bond Potts correlated percolation, where clusters are made of *nn* spins in the same state, connected by bonds with bond probability p_b. By choosing $p_b = p = 1 - e^{-q\beta J}$, it is possible to show that these clusters percolate at the Potts critical temperature $T_c(q)$, with percolation exponents identical to the thermal exponents and therefore behave as the critical droplets.

The formalism is based on the following diluted Potts model [32,102]:

$$-\mathcal{H}^q_{\mathrm{DP}} = J_b \sum_{\langle ij \rangle} (\delta_{\tau_i \tau_j} - 1)\delta_{\sigma_i \sigma_j} + qJ \sum_{\langle ij \rangle} \delta_{\sigma_i \sigma_j} , \quad (38)$$

where the second term, which controls the distribution of spin variables, is the q-state Potts Hamiltonian, whereas the first term contains auxiliary Potts variables $\tau_i = 1, 2, \ldots, s$ and controls the bonds distribution.

As in the Ising case, Hamiltonian (37) in the limit $s \to 1$ describes the site-bond Potts correlated percolation problem with p_b given by $p_b = 1 - e^{-q\beta J_b}$. The droplets are obtained in the particular case $J_b = qJ$. For this value in fact Hamiltonian (37) for $s \to 1$ coincides with the q-state Potts model.

Correlated Percolation, Table 1

Fractal dimensions, for $d = 2$, of the whole cluster (D), of the Hull (D_H), and of the red bonds (D_R) for the Potts droplets. It is also reported the thermal power exponent y_T

q	D	y_T	D_H	D_R
0	2	0	2	5/4
1	91/48	3/4	7/4	3/4
2	15/8	1	5/3	13/24
3	28/15	6/5	8/5	7/20
4	15/8	3/2	3/2	0

Once the Ising and Potts model has been mapped onto a percolation problem, we can extend some of the results of random percolation to thermal problems.

Fractal Structure in the Potts Model: Links and Blobs

Like in random percolation, also in the q-state Potts model it can be shown that at $T_\mathrm{c}(q)$ the critical droplets have a fractal structure made of links and blobs, with a fractal dimension $D(q) = d - \beta(q)/\nu(q)$, where $\beta(q)$ and $\nu(q)$ are respectively the order parameter and correlation length exponent. Therefore $D(q)$ coincides with the magnetic scaling exponent $y_\mathrm{H}(q)$. However the fractal dimension of the red bonds $D_\mathrm{R}(q)$ does not coincide with the thermal scaling exponent $y_\mathrm{T}(q)$, associated with the thermal variable J, like in random percolation. Instead $D_\mathrm{R}(q)$ is found to coincide with the bond probability scaling exponent y_b associated with the variable J_b in Hamiltonian (37) [27].

Like for random percolation the fractal dimension of the red bonds coincides with the fractal dimension of the antired bonds. Using the mapping from the Potts model to the Coulomb gas [87], it is possible to obtain the exact value of the fractal dimension of the red bonds and of the external perimeter or hull [27]. For further exact results see also [10,97].

From Table 1 it appears that the exact value of $D(q)$ does not vary substantially with q, for $d = 2$. This observation can be understood by noting that, using this geometrical approach, the driving mechanism of the critical behavior can be viewed as coalescence of clusters just like in random percolation. Then one would expect for any q that the fractal dimension should be close to the fractal dimension of the critical clusters in the percolation problem. This also explains the observation of Suzuky [98], known as strong universality, that for a large class of models the ratio γ/ν or β/ν do not vary appreciably. Since these ratios of critical exponents for fixed d depend only on the magnetic scaling exponent, which is identical to the fractal dimension, the strong universality is a consequence of the quasi-universal feature of the fractal dimension as discussed above. Unlikely the fractal dimension of the whole cluster, $D_\mathrm{R}(q)$ and $D_\mathrm{H}(q)$ do change substantially and characterize the different models as function of q. Particularly sensitive to q is the fractal dimension of the red bonds, which has its largest value at $q = 0$ (tree percolation), where the backbone is made only of links. As q approaches q_c the cluster becomes less ramified until the red bonds vanish ($D_\mathrm{R}(4) = 0$). This results in a drastic structural change from a links and blobs picture to a blobs picture only, anticipating a first order transition. Interestingly, the fractal dimension of the red bonds for $q = 0$, $D_\mathrm{R} = 5/4$, has been related to the abelian sandpile model [47]. The reason why $D_\mathrm{R}(q)$ is so model-dependent is due to the fact that the fractal set of the red bonds is only a small subset of the entire droplet, and therefore this "detail" is strongly model dependent. Also the thermal exponent $y_\mathrm{T}(q)$ is strongly model-dependent, however so far the geometrical characterization in terms of a fractal dimension for such exponent has not been found except $q = 1$ (random percolation).

Fortuin–Kasteleyn–Random Cluster Model

We will present here the random cluster model introduced by Kasteleyn and Fortuin. Let us consider the q-state Potts model on a d-dimensional hypercubic lattice. By freezing and deleting each interaction of the Hamiltonian (see "Appendix: Random Cluster Model and Ising Droplets"), they managed to write the partition function of the Potts model, $Z = \sum_{\{\sigma_i\}} \mathrm{e}^{-\beta \mathcal{H}_q}$, in the following way

$$Z = \sum_C p^{|C|}(1-p)^{|A|} q^{N_C} , \qquad (39)$$

where C is a configuration of bonds defined in the same hypercubic lattice, just like a bond configuration in the standard percolation model, $|C|$ and $|A|$ are respectively the number of bonds present and absent in the configuration C, and N_C is the number of clusters in the configuration C.

In conclusion, in the KF formalism the partition function of the Potts model is identical to the partition function (39) of a correlated bond percolation model [62,65] where the weight of each bond configuration C is given by

$$W(C) = p^{|C|}(1-p)^{|A|} q^{N_C} \qquad (40)$$

which coincides with the weight of the random percolation except for the extra factor q^{N_C}. They called this particular correlated bond percolation model the random cluster model. Clearly for $q = 1$ the cluster model coincides with the random percolation model.

Kasteleyn and Fortuin have related the percolation quantities associated with the random cluster model to the corresponding thermal quantities in the q-state Potts model [65]. In particular for the Ising case, $q = 2$,

$$|\langle S_i \rangle| = \langle \gamma_i^{\infty} \rangle_W \tag{41}$$

and

$$\langle S_i S_j \rangle = \langle \gamma_{ij} \rangle_W \, , \tag{42}$$

where $\langle \ldots \rangle$ is the Boltzmann average and $\langle \ldots \rangle_W$ is the average over bond configurations in the bond correlated percolation with weights given by (40). Here $\gamma_i^{\infty}(C)$ is equal to 1 if the spin at i belongs to the infinite cluster, 0 otherwise; $\gamma_{ij}(C)$ is equal to 1 if the spins at sites i and j belong to the same cluster, 0 otherwise.

Interestingly the connectivity properties in the KF random cluster model can be related to the CK droplets:

$$\rho_{\infty} = \langle \gamma_i^{\infty} \rangle_W \, , \tag{43}$$

$$p_{ij} = \langle \gamma_{ij} \rangle_W \, , \tag{44}$$

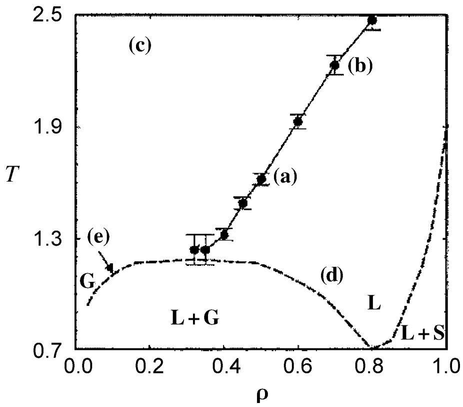
where ρ_{∞} and p_{ij} are defined in Subsect. "Ising Droplets Above $d = 4$". From Eqs. (41)–(44) it follows Eqs. (34).

Hill's Clusters

In this section we discuss the possibility to extend the definition of droplets to simple fluids. In 1955 Hill [60] introduced the concept of physical clusters in a fluid in an attempt to explain the phenomenon of condensation from a gas to a liquid. In a fluid made of particles interacting via a pair potential $u(r)$ physical clusters are defined as a group of particles pairwise bounded. A pair of particles is bounded if in the reference frame of their center of mass their total energy is less than zero. Namely their relative kinetic energy plus the potential energy is less than zero. The probability that two particles at distance r are bounded can be calculated [60] and is given by

$$p_{\mathrm{H}}(r) = \frac{4}{\pi} \int_0^{\sqrt{-\beta u(r)}} x^2 \mathrm{e}^{-x^2} \mathrm{d}x \, . \tag{45}$$

More recently it was noted [16] that the bond probability Eq. (45) calculated for the interaction of the three-dimensional nn lattice gas model is almost coincident with the bond probability p of Eq. (33). This implies that Hill's physical clusters for the $3d$ lattice gas almost coincide with the droplets defined by Coniglio and Klein, and in fact Hill's clusters percolate along a line almost indistinguishable from the droplets percolation line (see Fig. 2).



Correlated Percolation, Figure 3
Phase diagram of the Lennard–Jones fluid using molecular dynamics. The full line corresponds to percolation of cluster following Hill's definition. From [17]

In order to calculate percolation quantities in a fluid, in [38] the authors developed a theory based on Mayer's expansion. In particular, using this theory they calculated analytically for a potential made of a hard core plus an attractive interaction, the percolation line of Hill's physical clusters in a crude mean field approximation and compared with the liquid gas coexistence curve. They found that the percolation line ended just below the critical point in the low density phase but not exactly at the critical point. For further developments of the theory see [56].

Very recently, Campi et al. [17], using molecular dynamics have calculated the percolation line of Hill's physical clusters for a Lennard–Jones potential. The results showed a percolation line ending close or at the critical point (Fig. 3) suggesting that Hill's clusters are good candidates to describe the density fluctuations like the droplets in the lattice gas model, although there is no proof of relations analogous to those valid for the droplets in the lattice gas such as Eq. (34), which would prove that their size would diverge exactly at the critical point with thermal exponents.

Although Hill's clusters may represent the critical fluctuation near the critical point, we may wonder whether they have a physical meaning away from the critical point. In particular we may wonder whether we can detect experimentally the percolation line in the phase diagram. In a Lennard–Jones fluid, molecular dynamics shows that quantities such as viscosity or the diffusion coefficient do not seem to exhibit any anomalous behavior through the percolation line [17]. In some colloids instead the percolation line is detected through a steep increase of the viscosity. What would be the difference in the two cases? The difference may rely on their lifetime. The possibility to detect the percolation line of these clusters is expected to depend on the lifetime of the clusters which in turn depends on the bond lifetime. The larger the cluster lifetime, the larger the increase of the viscosity, and the better the percolation line can be detected. In Sect. "Scaling Behavior of the Viscosity" we will discuss the behavior of the viscosity as function of the lifetime of the clusters.

Clusters in Weak and Strong Gels

In the previous section we have shown the case in which the probability of having a bond between two particles coincides with the probability that the two particles form a bound state defined according to Hill's criterion. Now we want to show another mechanism leading to the formation of bound states, which is more appropriate to gels. The importance of connectivity in gels was first emphasized by Flory [51]. The application of percolation theory to gels was later suggested by de Gennes [43] and Stauffer [91,96]. Here we consider a system made of monomers in a solvent. Following [39] we shall assume that the monomers can interact with each other in two ways. One is the usual van der Waals interaction, and the other is a directional interaction that leads to a chemical bond. A simple model for such a system is a lattice gas model where an occupied site represents a monomer and an empty site a solvent. For simplicity we can put the monomer-solvent interaction and the solvent-solvent interaction equal to zero, and include such interactions in an effective monomer-monomer interaction. The monomer-monomer interaction ε_{ij} can reasonably be approximated by a nearest neighbor interaction

$$\varepsilon_{ij} = \begin{cases} -W \\ -E \end{cases} \tag{46}$$

where $-W$ is the van der Waals type of attraction and $-E$ is the bonding energy. Of course, this second interaction, which is the chemical interaction, occurs only when the monomers are in particular configurations. For simplicity we can suppose that there is one configuration which corresponds to the interaction of strength E, and Ω configurations which correspond to the interaction of strength W. We expect $E \gg W$ and $\Omega \gg 1$. It can be easily calculated [39] that such a system is equivalent to a lattice gas model with an effective *nn* interaction $-\varepsilon$ given by

$$e^{\beta\varepsilon} = e^{\beta E} + \Omega e^{\beta W} . \tag{47}$$

Therefore from the static point of view the system exhibits a coexistence curve and a critical temperature which characterizes the thermodynamics of the system. However the system microscopically behaves rather different from a standard lattice gas. In fact in a configuration in which two monomers are *nn*, in a standard lattice gas they feel one interaction, while in the system considered here with some probability p_b they feel a strong chemical interaction $-E$ and with probability $1 - p_b$ they feel a much smaller interaction $-W$. The probability p_b can be easily calculated and is given by

$$p_b = \frac{e^{\beta E}}{e^{\beta E} + \Omega e^{\beta W}} . \tag{48}$$

In conclusion, the system from the static point of view is equivalent to a lattice gas with interaction ε given by (47). However we can also study the percolation line of the clusters made by monomers connected by chemical bonds. This can be done by introducing bonds between *nn* particles in the lattice gas with *nn* interactions, the bonds being present with probability p_b given by Eq. (48). By changing the solvent the effective interaction W changes and one can realize three cases topologically similar to those of Fig. 2, where the percolation line ends at the critical point or below the critical point in the low density or high density phase (for more details see [39]).

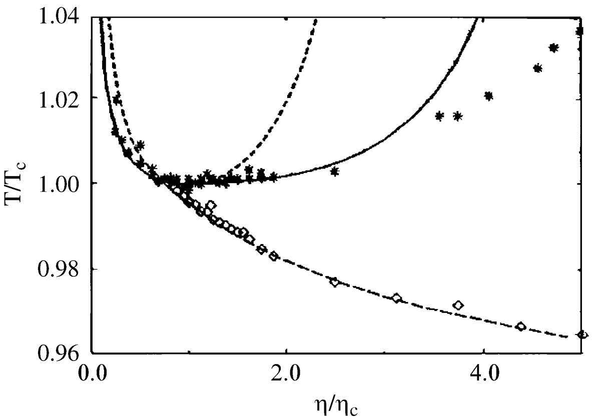
The lifetimes of the bonds are of the order of $e^{\beta E}$. Since E is very large the lifetime could be very large. For an infinite bond lifetime the bonded clusters are permanent and the viscosity diverges due to the divergence of the mean cluster size (see for example [96]), and the percolation line can be easily detected. We consider three particular physical systems which could be rather emblematic of a general situation where the percolation line has been detected:

a) Microemulsions of water in oil [20]
b) Triblock copolymers in unicellular systems [71,72]
c) Gelatin water methanol systems [101].

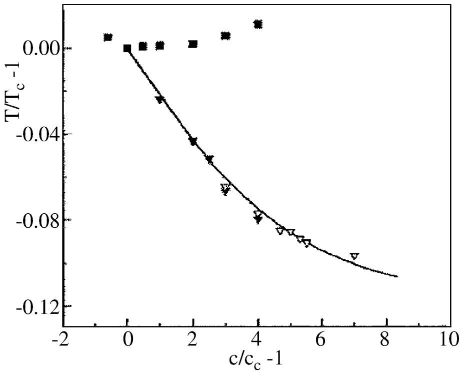
In Figs. 4, 5 and 6 we show respectively the phase diagram of the systems a), b), c), where the coexistence curve in the temperature-concentration diagram is shown together with "percolation lines".

In particular, in a) the system consists of three components AOT/water/decane. For the temperature and the concentration of interest, the system can be considered as made of small droplets of oil surrounded by water in a solvent. The droplets interact via a hard core potential plus short range attractive interaction. Because of the entropic nature of the attractive interaction, the coexistence curve is "upside-down" with the critical point being the minimum instead of the maximum (Fig. 4). The broken line is characterized by a steep increase of conductivity.

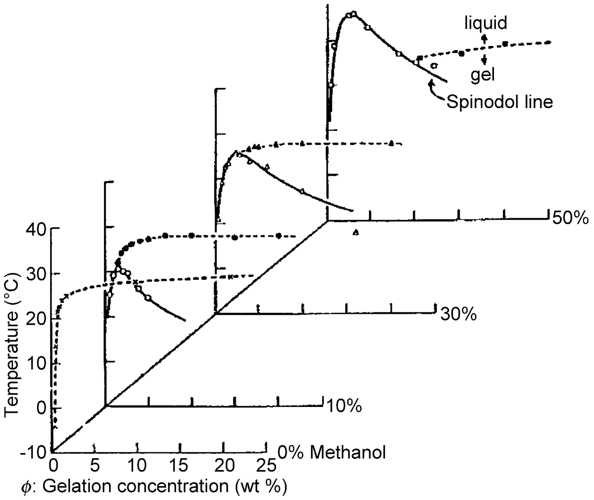
In b) the system is made of triblock copolymers unicellular in water solution, c is the volume fraction of the uni-



Correlated Percolation, Figure 4
Experimental points in AOT/water/decane from [20] together with the *coexistence curve* and *spinodal curve* based on the Baxter's model. The *percolation line* where the conductivity exhibits a steep increase has been fitted with the Baxter's model



Correlated Percolation, Figure 5
***L*/64 water system. Experimental points of the *coexistence curve* and *percolation line*, where the viscosity exhibits a steep increase. From [72]**



Correlated Percolation, Figure 6
Sol-Gel transition temperature (*solid symbols*) and the spinodal temperature (*open symbols*) of gelatin-water-methanol mixtures as function of gelatin concentration. At the sol-gel transition the viscosity diverges. From [101]

cells (Fig. 5). The line is characterized by a steep increase in the viscosity.

In c) the system is made of gelatin dissolved in water + methanol; ϕ is the gelatin concentration. The broken lines are characterized by the divergence of the viscosity and correspond to the sol-gel transition. Each line represents a different value of the methanol concentration, which has been chosen in such a way that the line ends at the consolute point or, below it, in the low or high density phase (Fig. 6).

In all these experiments the consolute point is characterized by a thermodynamical singularity, where the correlation length and compressibility diverge. The other lines are usually ascribed to a "percolation" transition. However it is important to precisely identify which are the relevant clusters in the three different systems. Also we would like to understand why, in system c), the viscosity diverges at the percolation transition, while in b) it reaches a plateau, and why in a) and b) the "percolation" lines end on the coexistence curve close to the critical point in the low density region. It is also important to realize that for each phenomenon is very important to define the proper cluster, which is responsible for the physical phenomenon. In the conductivity experiments in microemulsion the proper clusters are made of "touching" spheres similar to nearest neighbor particles in a lattice gas model. The viscoelastic properties of microemulsions may be more suitably described by clusters made of spheres pairwise bonded.

From the cluster properties of the lattice gas model we expect the infinite cluster is a necessary condition for a critical point therefore the percolation line ends just below the critical point in the low density region, as observed in the experiments described above and more recently in numerical simulations of models of interacting colloids [84].

In weak reversible gelatin the clusters are made of monomers (or polymers) bonded by a strong interaction which leads to chemical bond. In this case the bond probability can be changed by changing the solvent and therefore the percolation line, by properly choosing the solvent, can end on the coexistence curve at or below the critical point.

The reason why the viscosity in the gel experiments diverges at the percolation point, while it reaches a plateau in colloids, is due to the lifetime of the bonds which is much longer in the first system than in the second [46,73,86]. In low density colloids the proper cluster to describe coloidal gelation also appear to be related to strong bonds with large bond lifetime [14,41,86]. When the relaxation time is much smaller than the bond lifetime the dynamics is dominated by the clusters, otherwise a crossover is expected towards a regime due to the crowding of the particles [41]. Percolation line of clusters pairwise bonded can also be defined in fluids, but due to the negligible lifetime cannot be detected.

Scaling Behavior of the Viscosity

If the lifetime of the chemical bonds is infinite, the viscosity exhibits a divergence at the percolation threshold as recently shown in different models [11,45,103]

$$\eta \sim \xi^{\tilde{k}} \,, \tag{49}$$

where ξ is the linear dimension of the critical cluster which diverges at the percolation threshold with the exponent ν.

The relation between the diffusion coefficient $D(R)$ of a cluster of radius R and the viscosity η would be given by the Stokes–Einstein relation for a cluster radius much larger than ξ

$$D(R) \sim \frac{1}{R\eta} \,. \tag{50}$$

For cluster radius R smaller than ξ it has been proposed [75] that the viscosity will depend also on R in such a way as to satisfy a generalized Stokes–Einstein relation Eq. (50) with $\eta = \eta(R)$. When $R = \xi$ the viscosity $\eta(\xi) = \eta$, and from Eq. (50) one obtains the following scaling behavior for R:

$$D(R) \sim R^{-(1+\tilde{k})} \tag{51}$$

therefore the relaxation time $\tau(R)$ for a cluster of radius R is

$$\tau(R) \sim R^{1+\tilde{k}} \,. \tag{52}$$

If τ is the lifetime of a typical cluster, then a cluster of radius R will contribute to the viscosity if $\tau(R) < \tau$, and therefore:

$$\eta \sim \xi^{\tilde{k}} f\left(\frac{\tau}{\xi^{1+\tilde{k}}}\right) \sim \begin{cases} \xi^{\tilde{k}} & \tau > \xi^{1+\tilde{k}} \\ \tau^{\frac{\tilde{k}}{1+\tilde{k}}} & \tau < \xi^{1+\tilde{k}} \end{cases} \tag{53}$$

which implies that the viscosity will exhibit a steep increase followed by a plateau. The higher τ is, the higher the plateau.

The viscosity data on microemulsion (Fig. 5) shows in fact such a plateau, suggesting that the mechanism for the appearance of the plateau is linked to the bond lifetime which in turn is related to the cluster relaxation time.

Future Directions

In conclusion, we have discussed the interplay between the percolation line and critical point in systems where thermal correlations play an important role. The problem of defining the droplets in spin models is satisfactorily solved. However there are still some open problems. Above $d = 4$ in the Ising model the definition of droplets presents some difficulties, probably related to the upper critical dimension for the percolation problem. That critical dimension is six. This type of difficulty does not allow for a trivial extension of the arguments used in the random percolation problem, to explain the hyperscaling breakdown. Another open problem is the characterization of the thermal scaling exponent $1/\nu$, in terms of the fractal dimension of some subset of the critical droplet, as occurs in the random percolation problem.

In the last decade the KF, CK approach has been extended to frustrated systems. Interestingly this approach has led to a new frustrated percolation model, with unusual properties relevant to spin glasses and other glassy systems [77,80]. However the precise definition of clusters, which are able to characterize the critical droplets for spin glasses, is still missing.

Although some advances have been obtained towards a droplet definition in Lennard Jones systems [17], a general definition for continuum models of fluids still needs to be formulated.

Appendix: Random Cluster Model and Ising Droplets

In 1969 Kasteleyn and Fortuin (KF) [65] introduced a correlated bond percolation model, called the random cluster model, and showed that the partition function of this percolation model was identical to the partition function of the q-state Potts model. They also showed that the thermal quantities in the Potts model could be expressed in terms of connectivity properties of the random cluster model. Much later in 1980 Coniglio and Klein [30] independently have used a different approach with the aim to define the proper droplets in the Ising model. It was only later that it was realized that the two approaches were related, although the meaning of the clusters in the two approaches is different. We will discuss these two approaches here, and show that their statistical properties are the same.

Random Cluster Model

Let us consider an Ising system of spins $S_i = \pm 1$ on a lattice with nearest-neighbor interactions and, when needed, let us assume periodic boundary conditions in both directions. All interactions have strength J and the Hamiltonian is

$$\mathcal{H}(\{S_i\}) = -\sum_{\langle i,j\rangle} J(S_i S_j - 1) , \tag{54}$$

where $\{S_i\}$ represents a spin configuration and the sum is over *nn* spins. The main point in the KF approach is to replace the original Ising Hamiltonian with an annealed diluted Hamiltonian

$$\mathcal{H}'(\{S_i\}) = -\sum_{\langle i,j\rangle} J'_{ij}(S_i S_j - 1) , \tag{55}$$

where

$$J'_{ij} = \begin{cases} J' & \text{with probability } p \\ 0 & \text{with probability } (1-p) . \end{cases} \tag{56}$$

The parameter p is chosen such that the Boltzmann factor associated with an Ising configuration of the original model coincides with the weight associated with a spin configuration of the diluted Ising model

$$\begin{aligned} \mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} &\equiv \prod_{\langle i,j\rangle} \mathrm{e}^{\beta J(S_i S_j - 1)} \\ &= \prod_{\langle i,j\rangle} \left(p\mathrm{e}^{\beta J'(S_i S_j - 1)} + (1-p) \right) , \end{aligned} \tag{57}$$

where $\beta = 1/k_{\mathrm{B}}T$, k_{B} is the Boltzmann constant and T is the temperature. In order to satisfy (57) we must have

$$\mathrm{e}^{\beta J(S_i S_j - 1)} = p\mathrm{e}^{\beta J'(S_i S_j - 1)} + (1-p) . \tag{58}$$

We take now the limit $J' \mapsto \infty$. In such a case $\mathrm{e}^{\beta J'(S_i S_j - 1)}$ equals the Kronecker delta $\delta_{S_i S_j}$ and from (58) p is given by

$$p = 1 - \mathrm{e}^{-2\beta J} . \tag{59}$$

From (57), by performing the products we can write

$$\mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} = \sum_C W_{\mathrm{KF}}(\{S_i\}, C) , \tag{60}$$

where

$$W_{\mathrm{KF}}(\{S_i\}, C) = p^{|C|}(1-p)^{|A|} \prod_{\langle i,j\rangle \in C} \delta_{S_i S_j} . \tag{61}$$

Here C is a configuration of interactions where $|C|$ is the number of interactions of strength $J' = \infty$ and $|A|$ the number of interactions of strength 0. $|C| + |A| = |E|$, where $|E|$ is the total number of edges in the lattice.

$W_{\mathrm{KF}}(\{S_i\}, C)$ is the statistical weight associated a) with a spin configuration $\{S_i\}$ and b) with a set of interactions in the diluted model where $|C|$ edges have ∞ strength interactions, while all the other edges have 0 strength interactions. The Kronecker delta indicates that two spins connected by an ∞ strength interaction must be in the same state. Therefore the configuration C can be decomposed in clusters of parallel spins connected by infinite strength interactions.

Finally the partition function of the Ising model Z is obtained by summing the Boltzmann factor (60) over all the spin configurations. Since each cluster in the configuration C gives a contribution of 2, we obtain:

$$Z = \sum_C p^{|C|}(1-p)^{|A|} 2^{N_C} , \tag{62}$$

where N_C is the number of clusters in the configuration C.

In conclusion, in the KF formalism the partition function (62) is equivalent to the partition function of a correlated bond percolation model [62,65] where the weight of each bond configuration C is given by

$$W(C) = \sum_{\{S_i\}} W_{\mathrm{KF}}(\{S_i\}, C) = p^{|C|}(1-p)^{|A|} 2^{N_C} \tag{63}$$

which coincides with the weight of the random percolation except for the extra factor 2^{N_C}. Clearly all percolation quantities in this correlated bond model are weighted according to Eq. (63) coincide with the corresponding percolation quantities of the KF clusters made of parallel spins connected by an ∞ strength interaction, whose statistical weight is given by (61). Moreover using (61) and (60) Kasteleyn and Fortuin have proved that [65]

$$|\langle S_i \rangle| = \langle \gamma_i^\infty \rangle_W \tag{64}$$

and

$$\langle S_i S_j \rangle = \langle \gamma_{ij} \rangle_W , \tag{65}$$

where $\langle \dots \rangle$ is the Boltzmann average and $\langle \dots \rangle_W$ is the average over bond configurations in the bond correlated percolation with weights given by (63). Here $\gamma_i^\infty(C)$ is equal to 1 if the spin at site i belongs to the spanning cluster, 0 otherwise; $\gamma_{ij}(C)$ is equal to 1 if the spins at sites i and j belong to the same cluster, 0 otherwise.

Connection Between the Ising Droplets and the Random Cluster Model

In the approach followed by Coniglio and Klein [30], given a configuration of spins, one introduces at random connecting bonds between *nn* parallel spins with probability p_b, antiparallel spins are not connected with probability 1. Clusters are defined as maximal sets of parallel spins connected by bonds. The bonds here are fictitious, they are introduced only to define the clusters and do not modify the interaction energy as in the FK approach. For a given realization of bonds we distinguish the subsets C and B of *nn* parallel spins respectively connected and not connected by bonds and the subset D of *nn* antiparallel spins. The union of C, B and D coincides with the total set of *nn* pair of spins E. The statistical weight of a configuration of spins and bonds is [28,40]

$$W_\mathrm{CK}(\{S_i\}, C) = p_\mathrm{b}^{|C|}(1-p_\mathrm{b})^{|B|}\mathrm{e}^{-\beta\mathcal{H}(\{S_i\})}\,, \tag{66}$$

where $|C|$ and $|B|$ are the number of *nn* pairs of parallel spins respectively in the subset C and B not connected by bonds.

For a given spin configuration, using the Newton binomial rule, we have the following sum rule

$$\sum_C p_\mathrm{b}^{|C|}(1-p_\mathrm{b})^{|B|} = 1\,. \tag{67}$$

From Eq. (67) follows that the Ising partition function, Z, may be obtained by summing (66) over all bond configurations and then over all spin configurations.

$$Z = \sum_{\{S_i\}}\sum_C W_\mathrm{CK}(\{S_i\}, C) = \sum_{\{S_i\}} \mathrm{e}^{-\beta\mathcal{H}(\{S_i\})}\,. \tag{68}$$

The partition function of course does not depend on the value of p_b which controls the bond density. By tuning p_b instead it is possible to tune the size of the clusters. For example by taking $p_\mathrm{b} = 1$ the clusters would coincide with nearest-neighbor parallel spins, while for $p_\mathrm{b} = 0$ the clusters are reduced to single spins. By choosing the droplet bond probability $p_\mathrm{b} = 1 - \mathrm{e}^{-2\beta J} \equiv p$ and observing that $\mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} = \mathrm{e}^{-2\beta J|D|}$, where $|D|$ is the number of antiparallel pairs of spins, the weight (66) simplifies and becomes:

$$W_\mathrm{CK}(\{S_i\}, C) = p^{|C|}(1-p)^{|A|}\,, \tag{69}$$

where $|A| = |B| + |D| = |E| - |C|$.

From (69) we can calculate the weight $W(C)$ that a given configuration of connecting bonds C between *nn* parallel spins occurs. This configuration C can occur in many spin configurations. So we have to sum over all spin configurations compatible with the bond configuration C, namely

$$W(C) = \sum_{\{S_i\}} W_\mathrm{CK}(\{S_i\}, C) \prod_{\langle i,j\rangle \in C} \delta_{S_i S_j}\,, \tag{70}$$

where, due to the product of the Kronecker delta, the sum is over all spin configurations compatible with the bond configuration C. From (59) and (70) we have

$$\begin{aligned} W(C) &= \sum_{\{S_i\}} p^{|C|}(1-p)^{|A|} \prod_{\langle i,j\rangle \in C} \delta_{S_i S_j} \\ &= p^{|C|}(1-p)^{|A|}2^{N_C}\,. \end{aligned} \tag{71}$$

Consequently in (68) by taking first the sum over all spins compatible with the configuration C, the partition function Z can be written as in the KF formalism (62).

$$Z = \sum_C p^{|C|}(1-p)^{|A|}2^{N_C}\,. \tag{72}$$

In spite of the strong analogies the CK clusters and the KF clusters have a different meaning. In the CK formalism the clusters are defined directly in a given configuration of the Ising model as parallel spin connected by fictitious bonds, while in the KF formalism clusters are defined in the equivalent random cluster model. However, due to the equality of the weights (69) and (61) the statistical properties of both clusters are identical [40] and due to the relations between (61) and (63) both coincide with those of the correlated bond percolation whose weight is given by (63). More precisely, any percolation quantity $g(C)$ which depends only on the bond configuration has the same average

$$\langle g(C)\rangle_\mathrm{KF} = \langle g(C)\rangle_\mathrm{CK} = \langle g(C)\rangle_W\,, \tag{73}$$

where $\langle\ldots\rangle_\mathrm{KF}$, $\langle\ldots\rangle_\mathrm{CK}$ are the average over spin and bond configurations with weights given by (61) and (69) respectively and $\langle\ldots\rangle_W$ is the average over bond configurations in the bond correlated percolation with weights given by (63). In view of (73) it follows [40]

$$|\langle S_i\rangle| = \langle\gamma_i^\infty\rangle_\mathrm{CK} \tag{74}$$

and

$$\langle S_i S_j\rangle = \langle\gamma_{ij}\rangle_\mathrm{CK}\,. \tag{75}$$

We end this section noting that in order to generate an equilibrium CK droplet configuration in a computer simulation, it is enough to equilibrate a spin configuration of the Ising model and then introduce at random fictitious bonds between parallel spins with a probability given by (59).

Bibliography

Primary Literature

1. Abete T, de Candia A, Lairez D, Coniglio A (2004) Phys Rev Lett 93:228301
2. Aharony A, Gefen Y, Kapitulnik A (1984) J Phys A 17:l197; Alexander S, Grest GS, Makanishi H, Witten TA (1984) J Phys A 17:L185
3. Aizenman M (1997) Nucl Phys B 485:551
4. Amitrano C, di Liberto F, Figari R, Peruggi F (1983) J Phys A Math Gen 16:3925
5. Anghel M, Tobochnik J, Klein W, Gould H, Alexander FJ, Johnson G (2000) Phys Rev Lett 85:1270; Padoa SC, Sciortino F, Tartaglia P (1998) Phys Rev E 57:3797; Neerman DW, Coniglio A, Klein W, Stauffer D (1984) J Stat Phys 36:477
6. Bastiaansen PJM, Knops HJF (1997) J Phys A Math Gen 30:1791
7. Bialas P, Blanchard P, Fortunato S, Gandolfo D, Satz H (2000) Nucl Phys B 583:368; Blanchard P, Digal S, Fortunato S, Gandolfo D, Mendes T, Satz H (2000) J Phys A Math Gen 33:8603
8. Binder K (1976) Ann Phys NY 98:390
9. Birgeneau RJ, Cowley RA, Shirane G, Guggenheim HJ (1976) Phys Rev Lett 37:940; Birgeneau RJ, Cowley RA, Shirane G, Guggenheim HJ (1980) Phys Rev B 21:317
10. Blote HWJ, Knops YMM, Nienhuis B (1992) Phys Rev Lett 68:3440
11. Broderix K, Löwe H, Müller P, Zippelius A (2001) Phys Rev E 63:011510
12. Bug ALR, Safran SA, Grest GS, Webman I (1985) Phys Rev Lett 55:1896; Safran SA, Webman I, Grest GS (1985) Phys Rev A 32:506
13. Bunde A, Havlin S (1991) Percolation I. In: Bunde A, Havlin S (eds) Fractals and disordered systems. Springer, New York, pp 51–95
14. Campbell AI, Anderson VJ, van Duijneveldt JS, Bartlett P (2005) Phys Rev Lett 94:208301
15. Campi X, Krivine H (2005) Phys Rev C 72:057602; Krivine H, Campi X, Sator N (2003) Phys Rev C 67:044610; Mader CM, Chappars A, Elliott JB, Moretto LG,, Phair L, Wozniak GJ (2003) Phys Rev C 68:064601
16. Campi X, Krivine H, Puente A (1999) Physica A 262:328
17. Campi X, Krivine H, Sator N (2001) Physica A 296:24
18. Chayes JT, Chayes L, Grimmet GR, Kesten H, Schonmann R (1989) Ann Probab 17:1277
19. Chayes L, Coniglio A, Machta J, Shtengel K (1999) J Stat Phys 94:53
20. Chen SH, Rouch J, Sciortino F, Tartaglia P (1994) J Phys Cond Matter 6:10855
21. Coniglio A (1975) J Phys A 8:1773
22. Coniglio A (1976) Phys Rev B 13:2194
23. Coniglio A (1981) Phys Rev Lett 46:250
24. Coniglio A (1982) J Phys A 15:3829
25. Coniglio A (1983) In: Proceedings of Erice school on ferromagnetic transitions. Springer, New York
26. Coniglio A (1985) Finely divided matter. In: Boccara N, Daoud M (eds) Proc. les Houches Winter conference. Springer, New York
27. Coniglio A (1989) Phys Rev Lett 62:3054
28. Coniglio A (1990) In: Stanley HE, Ostrowsky W (eds) Correlation and connectivity – Geometric aspects of physics, chemistry and biology, vol 188. NATO ASI series. Kluwer, Dordrecht
29. Coniglio A, Figari R (1983) J Phys A Math Gen 16:L535
30. Coniglio A, Klein W (1980) J Phys A 13:2775
31. Coniglio A, Lubensky T (1980) J Phys A 13:1783
32. Coniglio A, Peruggi F (1982) J Phys A 15:1873
33. Coniglio A, Stanley HE (1984) Phys Rev Lett 52:1068
34. Coniglio A, Stauffer D (1980) Lett Nuovo Cimento 28:33
35. Coniglio A, Zia RVP (1982) J Phys A Math Gen 15:L399
36. Coniglio A, Nappi C, Russo L, Peruggi F (1976) Comm Math Phys 51:315
37. Coniglio A, Nappi C, Russo L, Peruggi F (1977) J Phys A 10:205
38. Coniglio A, De Angelis U, Forlani A, Lauro G (1977) J Phys A Math Gen 10:219; Coniglio A, De Angelis U, Forlani A (1977) J Phys A Math Gen 10:1123
39. Coniglio A, Stanley HE, Klein W (1979) Phys Rev Lett 42:518; Coniglio A, Stanley HE, Klein W (1982) Phys Rev B 25:6805
40. Coniglio A, di Liberto F, Monroy G, Peruggi F (1989) J Phys A 22:L837
41. Coniglio A, Abete T, de Candia A, del Gado E, Fierro A (2007) J Phys Condens Matter 19:205103; de Candia A, del Gado E, Fierro A, Sator N, Coniglio A (2005) Phys A 358:239; Coniglio A, de Arcangelis L, del Gado E, Fierro A, Sator N (2004) J Phys Condens Matter 16:S4831; Gimel JC, Nicolai T, Durand D (2001) Eur Phys J E 5:415
42. de Arcangelis L (1987) J Phys A 20:3057
43. de Gennes PG (1975) J Phys Paris 36:1049; de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, Ithaca
44. de Gennes PG (1976) La Recherche 7:919
45. del Gado E, de Arcangelis L, Coniglio A (2000) Eur Phys J E 2:359
46. del Gado E, Fierro A, de Arcangelis L, Coniglio A (2004) Phys Rev E 69:051103
47. Dhar D (1999) Physica A 263:4
48. Dunn AG, Essam JW, Ritchie DS (1975) J Phys C 8:4219; Cox MAA, Essam JW (1976) J Phys C 9:3985
49. Essam JW (1980) Rep Prog Phys 43:833
50. Fisher ME (1967) Physics NY 3:225; Fisher ME (1967) J Appl Phys 38:981; Fisher ME, Widom B (1969) J Chem Phys 50:3756; Fisher ME (1971) In: Green MS (ed) Critical Phenomena. Proc. of the international school of physics "Enrico Fermi" course LI, Varenna on lake Como (Italy). Academic, New York, p 1
51. Flory PJ (1941) J Am Chem Soc 63:3083; Flory PJ (1979) Principles of polymer chemistry. Cornell University Press. Ithaca
52. Fortunato S, Satz H (2000) Nucl Phys B Proc Suppl 83:452
53. Fortunato S, Aharony A, Coniglio C, Stauffer D (2004) Phys Rev E 70:056116
54. Frenkel J (1939) J Chem Phys 7:200; Frenkel J (1939) J Chem Phys 7:538
55. Gefen Y, Aharony A, Mandelbrot BB, Kirkpatrick S (1981) Phys Rev Lett 47:1771
56. Given JA, Stell G (1991) J Phys A Math Gen 24:3369
57. Harris AB, Lubensky TC, Holcomb W, Dasgupta C (1975) Phys Rev Lett 35:327
58. Havlin S, Bunde A (1991) Percolation II. In: Bunde A, Havlin S (eds) Fractals and disordered systems. Springer, New York, pp 97–149
59. Heermann DW, Stauffer D (1981) Z Phys B 44:339
60. Hill TL (1955) J Chem Phys 23: 617
61. Hong DC, Stanley HE, Coniglio A, Bunde A (1986) Phys Rev B 33:4564

62. Hu CK (1984) Phys Rev B 29:5103; Hu CK (1992) Phys Rev Lett 69:2739; Hu CK, Mak KS (1989) Phys Rev B 40:5007
63. Hu CK, Lin CY (1996) Phys Rev Lett 77:8
64. Jan N, Coniglio A, Stauffer D (1982) J Phys A 15:L699
65. Kasteleyn PW, Fortuin CM (1969) J Phys Soc Japan Suppl 26:11; Fortuin CM, Kasteleyn PW (1972) Phys Utrecht 57:536
66. Kertesz J (1989) Physica A 161:58
67. Kertesz J, Coniglio A, Stauffer D (1983) Clusters for random and interacting percolation. In: Deutscher G, Zallen R, Adler J (eds) Percolation structures and processes, Annals of the Israel Physical Society 5. Adam Hilger, Bristol, pp 121–147. The Israel Physical Society, Jerusalem
68. Kirkpatrick S (1978) AIP Conference Proc. 40:99
69. Ma YG (1999) Phys Rev Lett 83:3617; Ma YG, Han DD, Shen WQ, Cai XZ, Chen JG, He ZJ Long JL, Ma GL, Wang K, Wei YB, Yu LP, Zhang HY, Zhong C, Zhou XF, Zhu ZY (2004) J Phys G Nucl Part Phys 30:13
70. Makse HA, Havlin S, Stanley HE (1995) Nature 377:608; Makse HA, Andrade JS Jr, Batty M, Havlin S, Stanley HE (1998) Phys Rev E 58:7054
71. Mallamace F, Chen SH, Liu Y, Lobry L, Micali N (1999) Physica A 266:123
72. Mallamace F, Gambadauro P, Micali N, Tartaglia P, Liao C, Chen SH (2000) Phys Rev Lett 57:5431
73. Mallamace F, Chen SH, Coniglio A, de Arcangelis L, del Gado E, Fierro A (2006) Phys Rev E 73:020402
74. Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco
75. Martin JE, Douglas A, Wilcoxon JP (1988) Phys Rev Lett 61:262
76. Mayer JE (1937) J Chem Phys 5:67; Mayer JE, Ackermann PG (1937) J Chem Phys 5:74; Mayer JE, Harrison SF (1938) J Chem Phys 6:87; Mayer JE, Mayer MG (1940) Statistical mechanics. Wiley, New York
77. Monroy G, Coniglio A, di Liberto F, Peruggi F (1991) Phys Rev B 44:12605; Coniglio A (2000) Physica A 281:129
78. Muller-Krhumbaar H (1974) Phys Lett A 50:27
79. Murata KK (1979) J Phys A 12: 81
80. Newman CM, Machta J, Stein DL (2007) J Stat Phys 130:113
81. Nienhuis B, Berker AN, Riedel EK, Shick M (1979) Phys Rev Lett 43:737
82. Odagaki T, Ogita N, Matsuda H (1975) J Phys Soc Japan 39:618
83. Pike R, Stanley HE (1981) J Phys A 14:L169
84. Romano F, Tartaglia P, Sciortino F (2007) J Phys Condens Matter 19:322101; Zaccarelli E (2007) J Phys Condens Matter 19:323101
85. Roussenq J, Coniglio A, Stauffer D (1982) J Phys Paris 43:L703
86. Saika-Voivod I, Zaccarelli E, Sciortino F, Buldyrev SV, Tartaglia P (2004) Phys Rev E 70:041401
87. Saleur H, Duplantier B (1987) Phys Rev Lett 58:2325; Duplantier B, Saleur H (1989) Phys Rev Lett 63:2536
88. Sator N (2003) Physics Reports 376:1
89. Skal AS, Shklovskii BI (1975) Sov Phys Semicond 8:1029
90. Stanley HE (1977) J Phys A 10:1211
91. Stauffer D (1976) J Chem Soc Faraday Trans 72:1354
92. Stauffer D (1981) J Phys Lett 42:L49
93. Stauffer D (1990) Physica A 168:614
94. Stauffer D (1997) Physica A 242:1. for a minireview on the multiplicity of the infinite clusters
95. Stauffer D, Aharony A (1994) Introduction to percolation theory. Taylor and Francis, London
96. Stauffer D, Coniglio A, Adam M (1982) Adv Pol Sci 44:103. For a review on percolation and gelation (special volume Polymer Networks, ed. Dusek K)
97. Stella AL, Vanderzande C (1989) Phys Rev Lett 62:1067
98. Suzuki M (1974) Progr Theor Phys Kyoto 51:1992
99. Swendsen RH, Wang JS (1987) Phys Rev Lett 58:86
100. Sykes MF, Gaunt DS (1976) J Phys A 9:2131
101. Tanaka T, Swislow G, Ohmine I (1979) Phys Rev Lett 42:1557
102. Temesvary T (1984) J Phys A Math Gen 17:1703; Janke W, Schakel AMJ (2004) Nucl Phys B 700:385; Qian X, Deng Y, Blote HWJ (2005) Phys Rev E 72:056132; Deng Y, Guo W, Blote HWJ (2005) Phys Rev E 72:016101; Balog I, Uzelac K (2007) Phys Rev E 76:011103
103. Vernon DC, Plischke M, Joos B (2001) Phys Rev E 64:031505
104. Wang JS (1981) Physica A 161:249
105. Wang JS (1989) Physica A 161:249
106. Wang JS, Swendsen R (1990) Physica A 167:565
107. Webman I, Safran S, Bug ALR (1986) Phys Rev A 33:2842
108. Weinrib A (1984) Phys Rev B 29:387; Weinrib A, Halperin BI (1983) Phys Rev B 27:413; Sahimi M, Knackstedt MA, Sheppard AP (2000) Phys Rev E 61:4920; Sahimi M, Mukhopadhyay (1996) Phys Rev E 54:3870; Makse HA, Havlin S, Schwartz M, Stanley HE (1996) Phys Rev E 53:5445
109. Wolff U (1988) Phys Rev Lett 60:1461; Wolff U (1989) Phys Lett B 228:379; Wolff U (1989) Phys Rev Lett 62:361
110. Wu F (1982) Rev Mod Phys 54:235

Books and Reviews

Grimmett G (1989) Percolation. Springer, Berlin

Sahimi M (1994) Application of percolation theory. Taylor and Francis, London

Correlations in Complex Systems

Renat M. Yulmetyev[1,2], Peter Hänggi[3]

[1] Department of Physics, Kazan State University, Kazan, Russia

[2] Tatar State University of Pedagogical and Humanities Sciences, Kazan, Russia

[3] University of Augsburg, Augsburg, Germany

Article Outline

Glossary

Correlation A correlation describes the degree of relationship between two or more variables. The correlations are viewed due to the impact of random factors and can be characterized by the methods of probability theory.

Correlation function The correlation function (abbreviated, as CF) represents the quantitative measure for the compact description of the wide classes of correlation in the complex systems (CS). The correlation function of two variables in statistical mechanics provides a measure of the mutual order existing between them. It quantifies the way random variables at different positions are correlated. For example in a spin system, it is the thermal average of the scalar product of the spins at two lattice points over all possible orderings.

Memory effects in stochastic processes through correlations Memory effects (abbreviated, as ME) appear at a more detailed level of statistical description of correlation in the hierarchical manner. ME reflect the complicated or hidden character of creation, the propagation and the decay of correlation. ME are produced by inherent interactions and statistical after-effects in CS. For the statistical systems ME are induced by contracted description of the evolution of the dynamic variables of a CS.

Memory functions Memory functions describe mutual interrelations between the rates of change of random variables on different levels of the statistical description. The role of memory has its roots in the natural sciences since 1906 when the famous Russian mathematician Markov wrote his first paper in the theory of Markov Random Processes. The theory is based on the notion of the instant loss of memory from the prehistory (memoryless property) of random processes.

Information measures of statistical memory in complex systems From the physical point of view time scales of correlation and memory cannot be treated as arbitrary. Therefore, one can introduce some statistical quantifiers for the quantitative comparison of these time scales. They are dimensionless and possess the statistical spectra on the different levels of the statistical description.

Definition of the Subject

As commonly used in probability theory and statistics, a correlation (also so called correlation coefficient), measures the strength and the direction of a linear relationship between two random variables. In a more general sense, a correlation or co-relation reflects the deviation of two (or more) variables from mutual independence, although correlation does not imply causation. In this broad sense there are some quantifiers which measures the degree of correlation, suited to the nature of data. Increasing attention has been paid recently to the study of statistical memory effects in random processes that originate from nature by means of non-equilibrium statistical physics. The role of memory has its roots in natural sciences since 1906 when the famous Russian mathematician Markov wrote his first paper on the theory of Markov Random Processes (MRP) [1]. His theory is based on the notion of an instant loss of memory from the prehistory (memoryless property) of random processes. In contrast, there are an abundance of physical phenomena and processes which can be characterized by statistical memory effects: kinetic and relaxation processes in gases [2] and plasma [3], condensed matter physics (liquids [4], solids [5], and superconductivity [6]) astrophysics [7], nuclear physics [8], quantum [9] and classical [9] physics, to name only a few. At present, we have a whole toolbox available of statistical methods which can be efficiently used for the analysis of the memory effects occurring in diverse physical systems. Typical such schemes are Zwanzig–Mori's kinetic equations [10,11], generalized master equations and corresponding statistical quantifiers [12,13,14,15,16,17,18], Lee's recurrence relation method [19,20,21,22,23], the generalized Langevin equation (GLE) [24,25,26,27,28,29], etc.

Here we shall demonstrate that the presence of statistical memory effects is of salient importance for the functioning of the diverse natural complex systems. Particularly, it can imply that the presence of large memory times scales in the stochastic dynamics of discrete time series can characterize catastrophical (or pathological for live systems) violation of salutary dynamic states of CS. As an example, we will demonstrate here that the emergence of strong memory time scales in the chaotic behavior of complex systems (CS) is accompanied by the likely initiation and the existence of catastrophes and crises (Earthquakes, financial crises, cardiac and brain attack, etc.) in many CS and especially by the existence of pathological states (diseases and illness) in living systems.

Introduction

A common definition [30] of a correlation measure $\rho(X, Y)$ between two random variables X and Y with the mean values $E(X)$ and $E(Y)$, and fluctuations $\delta X = X - E(X)$ and $\delta Y = Y - E(Y)$, dispersions $\sigma_X^2 = E(\delta X^2) = E(X^2) - E(X)^2$ and $\sigma_Y^2 = E(\delta Y^2) = E(Y^2) - E(Y)^2$ is defined by:

$$\rho(X, Y) = \frac{E(\delta X\, \delta Y)}{\sigma_X\, \sigma_Y}\,,$$

where E is the expected value of the variable. Therefore we can write

$$\rho(X, Y) = \frac{[E(XY) - E(X)\, E(Y)]}{(E(X^2) - E(X)^2)^{1/2}\, (E(Y^2) - E(Y)^2)^{1/2}}\,.$$

Here, a correlation can be defined only if both of the dispersions are finite and both of them are nonzero. Due to the Cauchy–Schwarz inequality, a correlation cannot exceed 1 in absolute value. Consequently, a correlation assumes it maximum at 1 in the case of an increasing linear relationship, or -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is either to -1 or 1, the stronger is the correlation between the variables. If the variables are independent then the correlation equals 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

Since the absolute value of the sample correlation must be less than or equal to 1 the simple formula conveniently suggests a single-pass algorithm for calculating sample correlations. The square of the sample correlation coefficient, which is also known as the coefficient of determination, is the fraction of the variance in σ_x that is accounted for by a linear fit of x_i to σ_y. This is written

$$R^2_{xy} = 1 - \frac{\sigma^2_{y|x}}{\sigma^2_y}\,,$$

where $\sigma^2_{y|x}$ denotes the square of the error of a linear regression of x_i on y_i in the equation $y = a + bx$,

$$\sigma^2_{y|x} = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

and σ^2_y denotes just the dispersion of y.

Note that since the sample correlation coefficient is symmetric in x_i and y_i, we will obtain the same value for a fit to y_i:

$$R^2_{xy} = 1 - \frac{\sigma^2_{x|y}}{\sigma^2_x}\,.$$

This equation also gives an intuitive idea of the correlation coefficient for random (vector) variables of higher dimension. Just as the above described sample correlation coefficient is the fraction of variance accounted for by the fit of a 1-dimensional linear submanifold to a set of 2-dimensional vectors (x_i, y_i), so we can define a correlation coefficient for a fit of an m-dimensional linear submanifold to a set of n-dimensional vectors. For example, if we fit a plane $z = a + bx + cy$ to a set of data (x_i, y_i, z_i) then the correlation coefficient of z to x and y is

$$R^2 = 1 - \frac{\sigma^2_{z|xy}}{\sigma^2_z}\,.$$

Correlation and Memory in Discrete Non-Markov Stochastic Processes

Here we present a non-Markov approach [31,32] for the study of long-time correlations in chaotic long-time dynamics of CS. For example, let the variable x_i be defined as the R-R interval or the time distance between nearest, so called R peaks occurring in a human electrocardiogram (ECG). The generalization will consist in taking into account non-stationarity of stochastic processes and its further applications to the analysis of the heart-rate-variability.

We should bear in mind, that one of the key moments of the spectral approach in the analysis of stochastic processes consists in the use of normalized time correlation function (TCF)

$$a_0(t) = \frac{\langle\langle \mathbf{A}(T)\, \mathbf{A}(T + t)\rangle\rangle}{\langle \mathbf{A}(T)^2\rangle}\,. \qquad (1)$$

Here the time T indicates the beginning of a time serial, $\mathbf{A}(t)$ is a state vector of a complex system as defined below in Eq. (5) at t, $|\mathbf{A}(t)|$ is the length of vector $\mathbf{A}(t)$, the double angular brackets indicate a scalar product of vectors and an ensemble averaging. The ensemble averaging is, of course needed in Eq. (1) when correlation and other characteristic functions are constructed. The average and scalar product becomes equivalent when a vector is composed of elements from a discrete-time sampling, as done later. Here a continuous formalism is discussed for convenience. However further, since Sect. "Correlation and Memory in Discrete Non-Markov Stochastic Processes" we shall consider only a case of discrete processes.

The above-stated designation is true only for stationary systems. In a non-stationary case Eq. (1) is not true and should be changed. The concept of TCF can be generalized in case of discrete non-stationary sequence of signals. For this purpose the standard definition of the correlation coefficient in probability theory for the two random signals X and Y must be taken into account

$$\rho = \frac{\langle\langle \mathbf{XY}\rangle\rangle}{\sigma_X\, \sigma_Y}\,, \quad \sigma_X = \langle|\mathbf{X}|\rangle\,, \quad \sigma_Y = \langle|\mathbf{Y}|\rangle\,. \qquad (2)$$

In Eq. (2) the multi-component vectors **X**, **Y** are determined by fluctuations of signals x and y accordingly, σ_X^2, σ_Y^2 represent the dispersions of signals **x** and **y**, and values $|\mathbf{X}|, |\mathbf{Y}|$ represent the lengths of vectors **X**, **Y**, correspondingly. Therefore, the function

$$a(T,t) = \frac{\langle\langle \mathbf{A}(T)\, \mathbf{A}(T+t)\rangle\rangle}{\langle|\mathbf{A}(T)|\rangle\, \langle|\mathbf{A}(T+t)|\rangle} \tag{3}$$

can serve as the generalization of the concept of TCF (1) for non-stationary processes $\mathbf{A}(T+t)$. The non-stationary TCF (3) obeys the conditions of the normalization and attenuation of correlation

$$a(T,0) = 1\,, \quad \lim_{t\to\infty} a(T,t) = 0\,.$$

Let us note, that in a real CS the second limit, typically, is not carried out due possible occurrence nonergodocity (meaning that a time average does not equal its ensemble average). According to the Eqs. (1) and (3) for the quantitative description of non-stationarity it is convenient to introduce a function of non-stationarity

$$\gamma(T,t) = \frac{\langle|\mathbf{A}(T+t)|\rangle}{\langle|\mathbf{A}(T)|\rangle} = \left\{\frac{\sigma^2(T+t)}{\sigma^2(T)}\right\}^{1/2}\,. \tag{4}$$

One can see that this function equals the ratio of the lengths of vectors of final and initial states. In case of stationary process the dispersion does not vary with the time (or its variation is very weak). Therefore the following relations

$$\sigma(T+t) = \sigma(T)\,, \quad \gamma(T,t) = 1 \tag{5}$$

hold true for the stationary process.

Due to the condition (5) the following function

$$\mathit{\Gamma}(T,t) = 1 - \gamma(T,t) \tag{6}$$

is suitable in providing a dynamic parameter of non-stationarity. This dynamic parameter can serve as a quantitative measure of non-stationarity of the process under investigation. According to Eqs. (4)–(6) it is reasonable to suggest the existence of three different elementary classes of non-stationarity

$$\begin{aligned} |\mathit{\Gamma}(T,t)| &= |1-\gamma(T,t)| \\ &= \left\{\begin{array}{ll} \ll 1, & \text{weak non-stationarity} \\ \sim 1, & \text{intermediate non-stationarity} \\ \gg 1, & \text{strong non-stationarity} \end{array}\right\}\,. \end{aligned} \tag{7}$$

The existence of dynamic parameter of non-stationarity makes it possible to determine, on-principle, the type of non-stationarity of the underlying process and to find its spectral characteristics from the experimental data base. We intend to use Eqs. (4), (6), (7) for the quantitative description of effects of non-stationarity in the investigated temporary series of R-R intervals of human ECG's for healthy people and patients after myocardial infarction (MI).

Statistical Theory of Non-Stationary Discrete Non-Markov Processes in Complex Systems

Here we shall extend the original results of the statistical theory of discrete non-Markov processes in complex systems, developed recently in [31], for the case of non-stationary processes. The theory [31] is developed on the basis of first principles and represents a discrete finite-difference analogy for complex systems of well known Zwanzig–Mori's kinetic equations [10,11,12,13,14,15,16,17,18] used in the statistical physics of condensed matter.

We examine a discrete stochastic process $X(T+t)$, where $t = m\tau$

$$\begin{aligned} X = \{&x(T), x(T+\tau), x(T+2\tau), \dots, x(T+k\tau), \\ &\dots, x(T+(N-1)\tau)\}\,, \end{aligned} \tag{8}$$

where T is the beginning of the time and τ is a discretization time. The normalized time correlation function (TCF)

$$a(t) = \frac{1}{(N-m)\,\sigma^2} \sum_{j=0}^{N-1-m} \delta x(T+j\tau)\,\delta x(T+(j+m)\,\tau) \tag{9}$$

yields a convenient measure to analyze the dynamic properties of complex systems. Herein, we used the variance σ^2, the fluctuation $\delta x(T+j\tau)$, which in terms of the the mean value $\langle x\rangle$ reads:

$$\begin{aligned} \delta x_j &= \delta x(T+j\tau) = x(T+j\tau) - \langle x\rangle\,, \\ \sigma^2 &= \frac{1}{(N-m)} \sum_{j=0}^{N-1-m} \{\delta x(T+j\tau)\}^2\,, \end{aligned} \tag{10}$$

$$\langle x\rangle = \frac{1}{(N-m)} \sum_{j=0}^{N-1-m} x(T+j\tau)\,. \tag{11}$$

The discrete time t is given as $t = m\tau$.

In general, the mean value, the variance and TCF in (9), (10) and (11) is dependent on the numbers m and N. All indicated values cease to depend on numbers m and N for stationary processes when $m \ll N$. The definition of TCF in Eq. (9) is true only for stationary processes.

Next, we shall try to take into account this important dependence. With this purpose we shall form two k-dimensional vectors of state by the process (8):

$$\begin{aligned}\mathbf{A}_k^0 &= (\delta x_0, \delta x_1, \delta x_2, \ldots, \delta x_{k-1})\,,\\ \mathbf{A}_{m+k}^m &= (\delta x_m, \delta x_{m+1}, \delta x_{m+2}, \ldots, \delta x_{m+k-1})\,.\end{aligned} \tag{12}$$

When a vector of a state is composed of elements from a discrete-time sampling, the average and scalar product in Eq. (1) become equivalent. In an Euclidean space of vectors of state (12) TCF $a(t)$

$$a(t) = \frac{\langle \mathbf{A}_{N-1-m}^0\, \mathbf{A}_{N-1}^m \rangle}{(N-m)\{\sigma(N-m)\}^2} = \frac{\langle \mathbf{A}_{N-1-m}^0\, \mathbf{A}_{N-1}^m \rangle}{|\mathbf{A}_{N-1-m}^0|^2} \tag{13}$$

describes the correlation of two different states of the system ($t = m\tau$). Here the brackets $\langle \ldots \rangle$ indicate the scalar product of the two vectors. The dimension dependence of the corresponding vectors is also taken into account in the variance $\sigma = \sigma(N-m)$. As a matter of fact TCF $a(t) = \cos\vartheta$, where ϑ is the angle between the two vectors from Eq. (12). Let's introduce a unit vector of dimension $(N-m)$ in the following way:

$$\mathbf{n} = \frac{\mathbf{A}_{N-1-m}^0}{\sqrt{(N-m)\sigma^2}}\,. \tag{14}$$

Then, the TCF $a(t)$ (9) is given by

$$a(t) = \langle \mathbf{n}(0)\, \mathbf{n}(t) \rangle\,. \tag{15}$$

From the above discussion it is evident that Eqs. (13)–(15) are true for the stationary processes only. In case of non-stationary processes it is necessary to redefine TCF, taking into account the non-stationarity in the variance σ^2 in a line with Eqs.(2)–(7). For this purpose we shall redefine a unit vector of the final state as following

$$\mathbf{n}(t) = \frac{\mathbf{A}_{N-1}^m(t)}{|\mathbf{A}_{N-1}^m(t)|}\,. \tag{16}$$

For non-stationary processes it is convenient to write the TCF as the scalar product of the two unit vectors of the initial and final states

$$a(t) = \langle \mathbf{n}(0)\, \mathbf{n}(t) \rangle = \frac{\langle \mathbf{A}_{N-1-m}^0(0)\, \mathbf{A}_{N-1}^m(t) \rangle}{|\mathbf{A}_{N-1-m}^0(0)|\, |\mathbf{A}_{N-1}^m(t)|}\,. \tag{17}$$

Now we shall turn to the the dynamics of a non-stationary stochastic process. The equation of motion of a the random process x_j can be written in a finite-difference form for $0 \le j \le N-1$ [31,32] in the following way

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} \Rightarrow \frac{\Delta \delta x_j}{\Delta t} = \frac{\delta x_j(t+\tau) - \delta x_j(t)}{\tau}\,. \tag{18}$$

Then it is convenient to define the discrete evolution single step operator $\hat{U}$ as following:

$$x(T+(j+1)\,\tau) = \hat{U}(T+(j+1)\,\tau,\, T+j\tau)\, x(T+j\tau)\,. \tag{19}$$

In the case of stationary process we can rewrite the equation of motion (18) in a more simple form

$$\frac{\Delta \delta x_j}{\Delta t} = \tau^{-1}\big\{\hat{U}(\tau) - 1\big\}\, \delta x_j\,. \tag{20}$$

The invariance of the mean value $\langle x \rangle$ is taken into account in an Eq. (20)

$$\langle x \rangle = \hat{U}(\tau)\langle x \rangle\,, \quad \big\{\hat{U}(\tau) - 1\big\}\langle x \rangle = 0\,. \tag{21}$$

In case of a non-stationary process it is necessary to turn to the equation of motion for vector of the final state $\mathbf{A}_{m+k}^m(t)$ $(k = N-1-m)$

$$\frac{\Delta \mathbf{A}_{m+k}^m(t)}{\Delta t} = i\hat{L}(t,\tau)\, \mathbf{A}_{m+k}^m(t)\,, \tag{22}$$

where Liouville's quasioperator is

$$\hat{L}(t,\tau) = (i\tau)^{-1}\big\{\hat{U}(t+\tau, t) - 1\big\}\,. \tag{23}$$

It is well known that, in general, a stochastic trajectory does not obey a linear equation, so the general evolution operator and Liouville's quasioperator should probably be non-linear. Furthermore, in statistical physics the Liouville's operator acts upon the probability densities of dynamical variables, as well upon the variables itself like in Mori's paper [12]. The evolution of the density would be indeed linear. But Mori used the Liouville operator in the quantum equation of motion in [12]. In line with Mori [12] Eqs. (20), (22) can be considered as formal and exact equations of the motion of a complex system.

Thus, due to the Eqs. (17), (22) and (23) we may take into account the non-stationarity of the stochastic process. Towards this goal let's introduce the linear projection operator in Euclidean space of the state vectors

$$\Pi \mathbf{A}(t) = \frac{\mathbf{A}(0)\rangle \langle \mathbf{A}(0)\, \mathbf{A}(t) \rangle}{|\mathbf{A}(0)|^2}\,, \qquad \Pi = \frac{\mathbf{A}(0)\rangle \langle \mathbf{A}(0)}{\langle \mathbf{A}(0)\, \mathbf{A}(0) \rangle}\,, \tag{24}$$

where angular brackets in numerator present the boundaries of action for the scalar product.

For the analysis of the dynamics of the stochastic process $\mathbf{A}(t)$ the vector $\mathbf{A}_k^0(0)$ from (12) can be considered as a vector of the initial state $\mathbf{A}(0)$, and vector $\mathbf{A}_{m+k}^m(t)$ from (12) at value $m + k = N - 1$ can be considered as the vector of the final state $\mathbf{A}(t)$.

It is necessary to note that the projection operator (24) has the required property of idem-potency $\Pi^2 = \Pi$. The presence of operator Π allows one to introduce the mutually supplementary projection operator P:

$$P = 1 - \Pi\,, \quad P^2 = P\,, \quad \Pi P = P\Pi = 0\,. \tag{25}$$

It is necessary to remark, that both projectors Π and P are linear and can be recorded for the fulfillment of operations in the particular Euclidean space. Due to the property (17) and Eq. (4) it is easy to obtain the required TCF:

$$\begin{aligned}\Pi\,\mathbf{A}(t) &= \Pi\,\mathbf{A}^m_{m+k}(t)\\ &= \mathbf{A}^0_k(0)\langle \mathbf{n}^0_k(0)\,\mathbf{n}^m_{k+m}(t)\rangle \gamma_1(t)\\ &= \mathbf{A}^0_k(0)\,a(t)\,\gamma_1(t)\,,\\ \gamma_1(t) &= \frac{|\mathbf{A}^m_{m+k}(t)|}{|\mathbf{A}^0_m(0)|}\,.\end{aligned} \tag{26}$$

Therefore the projector Π generates a unit vector along the vector of the final state $\mathbf{A}(t)$ and makes its projection onto the initial state vector $\mathbf{A}(0)$.

The existence of a pair of two mutually supplementary projection operators Π and P allows one to carry out the splitting of Euclidean space of vectors $A(\mathbf{A}(0), \mathbf{A}(t) \in A)$ into a straight sum of two mutually supplementary subspaces in the following way

$$A = A' \overset{\bullet}{+} A''\,, \quad A' = \Pi A\,, \quad A'' = PA\,. \tag{27}$$

Substituting Eq. (27) in Eq. (23) we find Liouville's quasioperator $\hat{L}$ in a matrix form

$$\hat{L} = \hat{L}_{11} + \hat{L}_{12} + \hat{L}_{21} + \hat{L}_{22}\,, \tag{28}$$

where the matrix elements are introduced

$$\begin{aligned}\hat{L}_{11} &= \Pi\hat{L}\Pi\,, \quad \hat{L}_{12} = \Pi\hat{L}P\,,\\ \hat{L}_{21} &= P\hat{L}\Pi\,, \quad \hat{L}_{22} = P\hat{L}P\,.\end{aligned} \tag{29}$$

The Euclidean space of values of Liouville's quasioperator $W = \hat{L}A$ will be generated by the vectors $\mathbf{W}$ of dimension $k-1$

$$\begin{aligned}&(\mathbf{W}(0) \in W\,,\ \mathbf{W}(t) \in W)\\ &W = W' \overset{\bullet}{+} W''\,, \quad W' = \Pi W\,, \quad W'' = PW\,.\end{aligned} \tag{30}$$

Matrix elements $\hat{L}_{ij}$ of the contracted description

$$\hat{L} = \begin{pmatrix} \hat{L}_{11} & \hat{L}_{12}\\ \hat{L}_{21} & \hat{L}_{22}\end{pmatrix} \tag{31}$$

are acting in the following way:

$\hat{L}_{11}$ – from a subspace A' to subspace W',

$\hat{L}_{12}$ – from A'' to W',

$\hat{L}_{21}$ – from W' to W'' and

$\hat{L}_{22}$ – from A'' to W''.

The projection operators Π and P provide the contracted description of the stochastic process. Splitting the dynamic Eq. (22) into two equations in the two mutually supplementary Euclidean subspaces (see, for example [11]), we find

$$\frac{\Delta \mathbf{A}'(t)}{\Delta t} = i\hat{L}_{11}\,\mathbf{A}'(t) + i\hat{L}_{12}\,\mathbf{A}''(t)\,, \tag{32}$$

$$\frac{\Delta \mathbf{A}''(t)}{\Delta t} = i\hat{L}_{21}\,\mathbf{A}'(t) + i\hat{L}_{22}\,\mathbf{A}''(t)\,. \tag{33}$$

Following [31,32] it is necessary to eliminate first the irrelevant part $\mathbf{A}''(t)$ in order to simplify Liouville's Eq. (22) and then to write a closed equation for relevant part $\mathbf{A}'(t)$. According to [32] that can be achieved by a series of successive steps (for example, see Eqs. (32)–(36) in [32]). First a solution to Eq. (33) for the first step can be obtained in a form

$$\begin{aligned}\frac{\Delta \mathbf{A}''(t)}{\Delta t} &= \frac{\mathbf{A}''(t+\tau) - \mathbf{A}''(t)}{\tau}\\ &= i\hat{L}_{21}\,\mathbf{A}'(t) + i\hat{L}_{22}\,\mathbf{A}''(t)\,,\\ \mathbf{A}''(t+\tau) &= \mathbf{A}''(t) + i\tau\,\hat{L}_{21}\,\mathbf{A}'(t) + i\tau\,\hat{L}_{22}\,\mathbf{A}''(t)\\ &= \{1 + i\tau\,\hat{L}_{22}\}\,\mathbf{A}''(t) + i\tau\,\hat{L}_{21}\,\mathbf{A}'(t)\\ &= U_{22}(t+\tau,t)\,\mathbf{A}''(t) + i\tau\,\hat{L}_{21}(t+\tau,t)\,\mathbf{A}'(t)\,.\end{aligned} \tag{34}$$

We next can derive a finite-difference kinetic equation of a non-Markov type for TCF $a(t = m\tau)$

$$\frac{\Delta a(t)}{\Delta t} = \lambda_1 a(t) - \tau\Lambda_1 \sum_{j=0}^{m-1} M_1(t - j\tau)\,a(j\tau)\,. \tag{35}$$

Here, λ_1 is a eigenvalue, Λ_1 is a relaxation parameter of Liouville's quasioperator $\hat{L}$

$$\begin{aligned}\lambda_1 &= i\,\frac{\langle \mathbf{A}^0_k(0)\,\hat{L}\,\mathbf{A}^0_k(0)\rangle}{|\mathbf{A}^0_k(0)|^2}\,,\\ \Lambda_1 &= \frac{\langle \mathbf{A}^0_k(0)\,\hat{L}_{12}\,\hat{L}_{21}\,\mathbf{A}^0_k(0)\rangle}{|\mathbf{A}^0_k(0)|^2} = \frac{\langle \mathbf{A}^0_k(0)\,\hat{L}^2\,\mathbf{A}^0_k(0)\rangle}{|\mathbf{A}^0_k(0)|^2}\,,\end{aligned} \tag{36}$$

The angular brackets indicate here a scalar product of new state vectors. Function $M_1(t - j\tau)$ on the right side of Eq. (35) represents a modified memory function (MF) of the first order

$$M_1(t - j\tau) = \frac{\gamma_1(t - j\tau)}{\gamma_1(t)}\,m_1(t - j\tau)\,. \tag{37}$$

For stationary processes the function $\gamma_1(t)$ approaches unity. Then the memory functions $M_1(t)$ and $m_1(t)$ coincide with each other. The latter equation is the first kinetic finite-difference equation for TCF. It is remarkable, that the non-Markovity, discretization and non-stationarity of stochastic process can be considered explicitly. Due to the presence of non-stationarity both in TCF and in the first memory function this equation generalizes our results recently obtained in [31].

Following the projection technique described above, we arrive at a chain of connected kinetic finite-difference equations of a non-Markov type for the normalized short memory functions $m_n(t)$ in Euclidean space of state vectors of dimension $(k-n)$ $(t = m\tau, n \geq 1)$

$$\begin{aligned} \frac{\Delta m_n(t)}{\Delta t} &= \lambda_{n+1}\, m_n(t) - \tau \Lambda_{n+1} \\ &\quad \times \sum_{j=0}^{m-1} m_{n+1}(j\tau)\, m_n(t-j\tau) \\ &\quad \times \left\{ \frac{\gamma_{n+1}(j\tau)\gamma_{n+1}(t-j\tau)}{\gamma_n(t)} \right\} , \end{aligned} \tag{38}$$

$$\begin{aligned} m_{n+1}(t) &= \frac{\langle \mathbf{W}_{n+1}(0)\, \mathbf{W}_{n+1}(t)\rangle}{|\mathbf{W}_{n+1}(0)||\mathbf{W}_{n+1}(t)|} , \\ \gamma_n(j\tau) &= \left\{ \frac{|\mathbf{W}_n(j\tau)|}{|\mathbf{W}_n(0)|} \right\} . \end{aligned} \tag{39}$$

Here, $\gamma_n(j\tau)$ is the nth order of the non-stationarity function.

The set of all memory functions $m_1(t), m_2(t), m_3(t), \ldots$ allows one to describe non-Markov processes and statistical memory effects in the considered non-stationary system. For the particular case we obtain a more simple form for the set of equations for the first three short memory functions, namely $(t = m\tau)$:

$$\begin{aligned} \frac{\Delta a(t)}{\Delta t} &= -\tau \Lambda_1 \sum_{j=0}^{m-1} m_1(j\tau) \left\{ \frac{\gamma_1(j\tau)\gamma_1(t-j\tau)}{\gamma_1(t)} \right\} \\ &\quad \times a(t-j\tau) + \lambda_1 a(t) , \\ \frac{\Delta m_1(t)}{\Delta t} &= -\tau \Lambda_2 \sum_{j=0}^{m-1} m_2(j\tau) \left\{ \frac{\gamma_2(j\tau)\gamma_2(t-j\tau)}{\gamma_2(t)} \right\} \\ &\quad \times m_1(t-j\tau) + \lambda_2 m_1(t) , \\ \frac{\Delta m_2(t)}{\Delta t} &= -\tau \Lambda_3 \sum_{j=0}^{m-1} m_3(j\tau) \left\{ \frac{\gamma_3(j\tau)\gamma_3(t-j\tau)}{\gamma_3(t)} \right\} \\ &\quad \times m_2(t-j\tau) + \lambda_3 m_2(t) . \end{aligned} \tag{40}$$

Here the relaxation parameters Λ_1, Λ_2 and Λ_3 have already been determined and the non-stationarity functions $\gamma_n(t)$ have been introduced earlier. By analogy with Eq. (6) we can introduce a set of dynamic parameters of non-stationarity (PNS) for the arbitrary nth relaxation level

$$\Gamma_n(T,t) = 1 - \gamma_n(t) = 1 - \gamma_n(T,t) . \tag{41}$$

The whole set of values of dynamic PNS $\gamma_n(t)$ determines the broad spectrum of non-stationarity effects of the considered process.

The obtained equations are similar to the well known Zwanzig–Mori's kinetic equations [10,11,12,13,14, 15,16,17,18] used in non-equilibrium statistical physics of condensed matters. Let us point out three essential distinctions of our Eqs. (40) from the results in [10,11,12]. In Zwanzig–Mori's theory the key moment in the analysis of considered physical systems is the presence of a Hamiltonian and an operation of a statistical averaging carried out with the help of quantum density operator or classic Gibbs distribution function [33]. In our examined case, both the Hamiltonian and the distribution function are absent. There are exact classic or quantum equations of motion in physics; so Liouville's equation and Liouville's operator are useful in many applications. The motion of individual particles and whole statistic system is described by variables varying in continuous time. Therefore, for physical systems it is possible to use effectively the methods of integro-differential calculus, based on the mathematically accustomed (but from the physical point of view difficult for understanding) representation of infinitesimal variations of values of coordinates and time. By nature, the monitored time evolution of most complex systems is discrete. As well known, discretization is inherent in a wide variety both of classical and quantum complex systems. This forces us to abandon the concept of an infinite small values and continuity and instead turn to discrete-difference schemes. And, at last, the third feature is connected with incorporating the issue of non-stationary processes into our theory. The Zwanzig–Mori theory is typically applied only for stationary processes. Due to the introduction of normalized vectors of states and the use of the appropriate projection technique [13] our theory allows to take into account non-stationary processes as well. The latter ones can be described by the non-Markov kinetic equations together with the introduction of the set of non-stationarity functions.

The non-stationary theory [32] put forward here differs from the stationary case [31]. The external structure of the kinetic equations remains invariant; they represent the kinetic equations with memory. However, the functions and the parameters, which are included in these equations, appreciably differ from each other. As we already remarked above, non-stationarity effects enter both, in the functions $\gamma_n(t)$ and in spectral and kinetic parameters.

Correlation and Memory in Discrete Non-Markov Stochastic Processes Generated by Random Events

Here we shall find a chain of the kinetic interconnected finite-difference equations for a discrete correlation function $a(n)$ and memory functions $M_s(n)$ in the linear scale of events $E = \{\xi_1, \xi_2, \xi_3, \dots, \xi_N\}$.

The Basic Assumptions and Concepts of the Theory of Discrete Non-Markov Stochastic Processes of the Events Correlations

As an example we shall consider the time variations of the total X-ray flux of an astrophysical object at a succession of events:

$$E = \{\xi_1, \xi_2, \xi_3, \dots, \xi_k, \dots, \xi_N\} , \tag{42}$$

where ξ_i is an event, which occurs at time instant t_i, where $i = 1, \dots, N$ counts the event number.

The average value $\langle E \rangle$, fluctuations $\delta\xi$ and dispersion σ^2 for the set of N events are obtained as:

$$\begin{aligned} \langle E \rangle &= \frac{1}{N} \sum_{i=1}^{N} \xi_i , \delta\xi_i = \xi_i - \langle E \rangle , \\ \sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} \delta\xi_i^2 = \frac{1}{N} \sum_{i=1}^{N} \{\xi_i - \langle E \rangle\}^2 . \end{aligned} \tag{43}$$

According to [35,36,37,38], for the description of the dynamical properties of the studied system we introduce the correlation dependence of the discrete set of events (see Eq. (42)) using the CF:

$$a(n) = \frac{1}{(N-m)\,\sigma^2} \sum_{i=1}^{N-m} \delta\xi_i\, \delta\xi_{i+m} . \tag{44}$$

Here $n = m\Delta n$, $\Delta n = 1$ is the discretization step. The function $a(n)$, which emerges in this way, is the "event" correlation function (ECF). The normalized ECF must obey the conditions of normalization and of the attenuation of correlation, i. e.: $\lim_{n\to 1} a(n) = 1$, $\lim_{n\to\infty} a(n) = 0$. We remark, however, that the second condition for the case the physical complex systems is typically not observed (at $N \gg 0$). It is necessary to note that in [18] the correlation function for the aftershock events has been introduced:

$$C(n + n_W, n_W) = \frac{[\langle t_{n+n_W}\, t_{n_W} \rangle - \langle t_{n+n_W} \rangle \langle t_{n_W} \rangle]}{\left(\sigma^2_{n+n_W} \sigma^2_{n_W}\right)^{1/2}} ,$$

where the averages and the variance are given by

$$\begin{aligned} \langle t_m \rangle &= \frac{1}{N} \sum_{k=o}^{N-1} t_{m+k} , \\ \langle t_m\, t'_m \rangle &= \frac{1}{N} \sum_{k=o}^{N-1} t_{m+k}\, t'_{m+k} , \quad \text{and} \\ \sigma_m^2 &= \langle t_m^2 \rangle - \langle t_m \rangle^2 , \end{aligned}$$

respectively.

By the direct analogy of [31,32,35] we use the finite-difference Liouville's equation of motion in the event scale for describing the evolution of discrete set of events Eq. (11), (13):

$$\frac{\Delta\xi_i(n)}{\Delta n} = i\,\widehat{L}(n, 1)\,\xi_i(n) . \tag{45}$$

Here $\xi_i(n+1) = U(n+1, n)\,\xi_i(n)$, $U(n+1, n)$ is the "event" evolution operator. It determines the shift in linear event scale to one step Δn. The evolution operator $U(n+1, n)$ and Liouville's quasioperator $\widehat{L}(n, 1)$ can be made explicit by writing: $\widehat{L}(n, 1) = (i\Delta n)^{-1}\,(U(n$ $+1, n) - 1)$.

Let's represent the set of values of the dynamical variable $\delta\xi_j = \delta\xi(j\Delta n)$, $j = 1, \dots, N$ as the k-component vector of system state in linear Euclidean space:

a) the vector of initial state of studied complex system:

$$\mathbf{A}_k^1 = \{\delta\xi_1, \delta\xi_2, \delta\xi_3, \dots, \delta\xi_k\} , \tag{46}$$

b) the vector of final system's state, which is shifted on the m events along the event scale:

$$\mathbf{A}_{m+k}^m = \{\delta\xi_{m+1}, \delta\xi_{m+2}, \delta\xi_{m+3}, \dots, \delta\xi_{m+k}\} , \tag{47}$$

where $1 \le k \le N$. The vectors of initial and final states, which are submitted in a similar way, are very convenient for analyzing the dynamics of the observed discrete stochastic processes with the help of discrete non-Markov processes.

To represent the ECF in a more compact form, we use the expression for the scalar product of vectors $\langle \mathbf{A}_k^1 \cdot \mathbf{A}_{m+k}^m \rangle = \sum_{j=1}^{k} A_j^1\, A_{m+j}^m$, and the Eqs. (44), (46) and (47):

$$a(n) = \frac{\langle \mathbf{A}_k^1(1)\, \mathbf{A}_{k+m}^m(n) \rangle}{\langle |\mathbf{A}_k^1(1)|^2 \rangle} . \tag{48}$$

Construction of Chain of Finite-Difference Non-Markov Kinetic Equations for the Events Correlation

Let us consider the finite-difference Liouville's equation (Eq. (44)) for the vector of final system states:

$$\frac{\Delta \mathbf{A}^{m}_{m+k}(n)}{\Delta n} = i\,\widehat{L}(n,1)\,\mathbf{A}^{m}_{m+k}(n)\,. \tag{49}$$

We introduce the projection operator Π, which projects the final vector $\mathbf{A}^{m}_{m+k}(n)$ on the direction of initial vector, and also the orthogonal operator P. The operators Π and P possess the following properties: $\Pi = |\mathbf{A}^1_k(1)\rangle\langle \mathbf{A}^1_k(1)|/\langle|\mathbf{A}^1_k(1)|^2\rangle$, $\Pi^2 = \Pi$, $\mathrm{P} = 1 - \Pi$, $\mathrm{P}^2 = \mathrm{P}$, $\Pi\mathrm{P} = \mathrm{P}\Pi = 0$. They are idempotent and mutually complementary.

The initial ECF $a(n)$ (Eq. (48)) can be derived by means of projecting the vector of final states $\mathbf{A}^{m}_{m+k}(n)$ on the vector of initial state $\mathbf{A}^1_k(1)$:

$$\Pi\,\mathbf{A}^{m}_{m+k}(n) = \frac{\mathbf{A}^1_k(1)\langle \mathbf{A}^1_k(1)\,\mathbf{A}^{m}_{m+k}(n)\rangle}{\langle|\mathbf{A}^0_k|^2\rangle} = \mathbf{A}^1_k(1)\,a(n)\,. \tag{50}$$

The operators Π and P split Euclidean vector space $A(k)$ into two mutually orthogonal subspaces:

$$\begin{aligned} &A(k) = A'(k) + A''(k)\,, \quad A'(k) = \Pi A(k)\,, \\ &A''(k) = \mathrm{P}A(k)\,, \quad \mathbf{A}^{m}_{m+k} \in A(k)\,. \end{aligned} \tag{51}$$

As a result the finite-difference Liouville's Eq. (47) can be represented as a system of 2 equations into mutually orthogonal linear subspaces:

$$\frac{\Delta \mathbf{A}'(n)}{\Delta n} = i\,\widehat{L}_{11}\,\mathbf{A}'(n) + i\,\widehat{L}_{12}\,\mathbf{A}''(n)\,, \tag{52}$$

$$\frac{\Delta \mathbf{A}''(n)}{\Delta n} = i\,\widehat{L}_{21}\,\mathbf{A}'(n) + i\,\widehat{L}_{22}\,\mathbf{A}''(n)\,. \tag{53}$$

Here $\widehat{L}_{ij} = \Pi_i\,\widehat{L}\,\Pi_j$ are the matrix elements of Liouville's quasioperator:

$$\begin{aligned} &\widehat{L} = \widehat{L}_{11} + \widehat{L}_{12} + \widehat{L}_{21} + \widehat{L}_{22}\,, \\ &\widehat{L}_{11} = \Pi\widehat{L}\Pi\,, \quad \widehat{L}_{12} = \Pi\widehat{L}\mathrm{P}\,, \\ &\widehat{L}_{21} = \mathrm{P}\widehat{L}\Pi\,, \quad \widehat{L}_{22} = \mathrm{P}\widehat{L}\mathrm{P}\,. \end{aligned} \tag{54}$$

To solve the system of Eqs. (52), (53) we eliminate the non-reducible part, which contains $\mathbf{A}''(n)$ and derive the self-contained equation for the reducible part $\mathbf{A}'(n)$. In doing so we solve the Eq. (52) step-by-step and shall substitute the obtained solution into the Eq. (53). As a result we arrive at the closed kinetic equation:

$$\begin{aligned} \frac{\Delta \mathbf{A}'(n+m\Delta n)}{\Delta n} &= i\,\widehat{L}_{11}\,\mathbf{A}'(n+m\Delta n) \\ &\quad + i\,\widehat{L}_{12}\{1 + i\Delta n\,\widehat{L}_{22}\}^m\,\mathbf{A}''(n) \\ &\quad - \widehat{L}_{12}\sum_{j=1}^{m}\{1 + i\Delta n\,\widehat{L}_{22}\}^j\,\Delta n \\ &\quad \times \widehat{L}_{21}\,\mathbf{A}'(n+[m-j]\Delta n)\,. \end{aligned} \tag{55}$$

By use of projection operators Π and P we found the closed finite-difference kinetic equation of non-Markov type for the initial ECF:

$$\frac{\Delta a(n)}{\Delta n} = i\lambda_1\,a(n) - \Delta n\,\Lambda_1\sum_{j=1}^{m} M_1(j\Delta n)\,a(n-j\Delta n)\,. \tag{56}$$

As $\Delta n = 1$, solution of the last equation must be following:

$$a(n+1) = \{i\lambda_1 + 1\}\,a(n) - \Lambda_1\sum_{j=1}^{m} M_1(j)\,a(n-j)\,. \tag{57}$$

Here λ_1 is the proper value of Liouville's quasioperator $\widehat{L}$, Λ_1 is the relaxation parameter, which dimension is square of frequency, $M_1(j\Delta n)$ is the normalized memory function of the first order:

$$\begin{aligned} \lambda_1 &= \frac{\langle A^1_k(1)\,\widehat{L}\,A^1_k(1)\rangle}{\langle|A^1_k(1)|^2\rangle}\,, \\ \Lambda_1 &= \frac{\langle A^1_k\,\widehat{L}_{12}\,\widehat{L}_{21}\,A^1_k(1)\rangle}{|A^1_k(1)|^2\rangle}\,, \\ M_1(j\Delta n) &= \frac{\langle A^1_k(1)\,\widehat{L}_{12}(1 + i\Delta n\,\widehat{L}_{22})^j\,\widehat{L}_{21}\,A^1_k(1)\rangle}{\langle A^1_k(1)\,\widehat{L}_{12}\,\widehat{L}_{21}\,A^1_k(1)\rangle}\,. \end{aligned}$$

To obtain the finite-difference kinetic equation for the normalized event memory function of first order and, further, for the higher $(s-1)$th orders as well, we have to repeat the foregoing procedure step-by-step. However, we shall make use of the Gram–Schmidt orthogonalization procedure [16]:

$$\langle \mathbf{W}_s\mathbf{W}_p\rangle = \delta_{sp}\,\langle|\mathbf{W}_s|^2\rangle\,. \tag{58}$$

Where δ_{sp} is a Kronecker's symbol. Now we shall derive the recurrence formula $\mathbf{W}_s = \mathbf{W}_s(n)$ for defining the set of the orthogonal dynamic variables:

$$\begin{aligned} &\mathbf{W}_0 = \mathbf{A}^1_k\,, \\ &\mathbf{W}_1 = \{i\widehat{L} - \lambda_1\}\mathbf{W}_0\,, \\ &\mathbf{W}_2 = \{i\widehat{L} - \lambda_2\}\mathbf{W}_1 - \Lambda_1\mathbf{W}_0, \ldots \end{aligned} \tag{59}$$

According to the foregoing formulas we can introduce the succession of projection operators $\Pi_s = \Pi_1^{(s)}$ and the set of mutually complementary projectors $\mathrm{P}_s = 1 - \Pi_s$, which possess the following properties:

$$\Pi_s = \frac{|\mathbf{W}_s\rangle\langle\mathbf{W}_s|}{\langle|\mathbf{W}_s|^2\rangle}\,, \qquad \Pi_s^2 = \Pi_s\,,$$
$$\mathrm{P}_s^2 = \mathrm{P}_s\,, \qquad \Pi_s\,\mathrm{P}_s = \mathrm{P}_s\,\Pi_s = 0\,,$$
$$\Pi_s\,\Pi_p = \delta_{sp}\,\Pi_s\,, \qquad \mathrm{P}_s\,\mathrm{P}_p = \delta_{sp}\,\mathrm{P}_s\,.$$

Each of these operators pairs Π_s, P_s splits the corresponding Euclidean vector space $\mathbf{W}_s$ into the two mutual complementary subspaces: $W_s = W_s' + W_s'$, $W_s' = \Pi_s W_s$, $W_s'' = P_s W_s$. Using the projection operator technique for the next orthogonal variables $\mathbf{W}_s$, we shall obtain the chain of interconnected kinetic finite-difference equations of the non-Markov type for the normalized correlation functions of the $(s-1)$th order:

$$\frac{\Delta M_1(n)}{\Delta n} = i\,\lambda_2\,M_1(n) - \Lambda_2 \sum_{j=1}^{m} M_2(j)\,M_1(n-j)\,,$$
$$\dots\,,$$
$$\frac{\Delta M_{s-1}(n)}{\Delta n} = i\,\lambda_s\,M_{s-1}(n) - \Lambda_s \sum_{j=1}^{m} M_{s-1}(j)\,M_s(n-j)\,. \tag{60}$$

In these equations the normalized events memory function of the first order: $M_1(n) = \langle \mathbf{W}_1(1+i\Delta n\widehat{L})^m \mathbf{W}_1\rangle / \langle|\mathbf{W}_1|^2\rangle$, memory function of the $(s-1)$th order: $M_{s-1}(n) = \langle \mathbf{W}_{s-1}(1+i\Delta n\widehat{L})^m \mathbf{W}_{s-1}\rangle / \langle|\mathbf{W}_{s-1}|^2\rangle$, the proper value of the Liouville's quasioperator $\widehat{L}$: $\lambda_s = \langle \mathbf{W}_s \widehat{L} \mathbf{W}_s\rangle / \langle|\mathbf{W}_s|^2\rangle$ and the relaxation parameter $\Lambda_s = \langle|\mathbf{W}_s|^2\rangle / \langle|\mathbf{W}_{s-1}|^2\rangle$ are introduced.

The foregoing finite-difference kinetic Eqs. (60) present the generalization of the statistical theory [31,32,35] for the case of event correlations in discrete stochastic evolution of non-Hamilton complex systems.

Information Measures of Memory in Complex Systems

As an information measures of memory it is useful to apply different dimensionless quantifiers. As a first measure we use the frequency dependence of non-Markovity parameter. This measure was introduced in [31] and it is defined as:

$$\varepsilon_i(\nu) = \left\{\frac{\mu_{i-1}(\nu)}{\mu_i(\nu)}\right\}^{1/2}\,. \tag{61}$$

Here, $\mu_i(\nu)$ denotes the frequency power spectrum of memory function of the ist order $M_i(n)$: $\mu_i(\nu) = |\Delta n \sum_{n=1}^{N} M_i(n)\cos(2\pi n\nu)|^2$. The non-Markovity parameter $\varepsilon_i(\nu)$ along with the memory functions enables us to characterize quantitatively the statistical memory effects in discrete complex systems of various nature. Because the functions $\mu_i(\nu)$ exist for each of the ith levels of relaxation, we obtain the statistical spectrum of parameters: $\varepsilon_i(\nu)$, $i = 1, 2, 3, \dots$.

Alternatively, a study of 'memory' in physiological time series for electroencephalographic (EEG) and magnetoencephalographic (MEG) signals, both of healthy subjects and patients (including epilepsy patients) has been based on the detrended-fluctuation analysis (DFA) [39,40].

The characterization of memory *per se* is based on a set of dimensionless statistical quantifiers which are capable for measuring the memory strength which is inherent to the complex dynamics.

According to [41] a second set an information memory measure can be constructed as follows:

$$\delta_i(\nu) = \left|\frac{\tilde{M}_i'(\nu)}{\tilde{M}_{i+1}'(\nu)}\right|\,.$$

Here, $\mu_i(\nu) = |\tilde{M}_i(\nu)|^2$ denotes the power spectrum of the corresponding memory function $M_i(t)$, $\tilde{M}_i'(\nu) = \mathrm{d}\tilde{M}_i(\nu)/\mathrm{d}\nu$ and $\tilde{M}_i(\nu)$ is the Fourier transform of the memory function $M_i(t)$. The measures $\varepsilon_i(\nu)$ are suitable for the quantification of the memory effects on a relative scale whereas the second set $\delta_i(\nu)$ proves to be useful for quantifying the amplification of relative memory effects occurring on different complexity levels. Both measures provide statistical criteria for comparison between the relaxation time scales and memory time scales of the process under consideration. For values obeying $\{\varepsilon, \delta\} \gg 1$ one can observe a complex dynamics characterized by the short-ranged temporal memory scales. In the memoryless limit these processes assume a δ-like memory with parameters $\varepsilon, \delta \to \infty$. When $\{\varepsilon, \delta\} > 1$ one deals with a situation with moderate memory strength, and the case where both $\varepsilon, \delta \sim 1$ typically constitutes a more regular and robust random process exhibiting strong memory features.

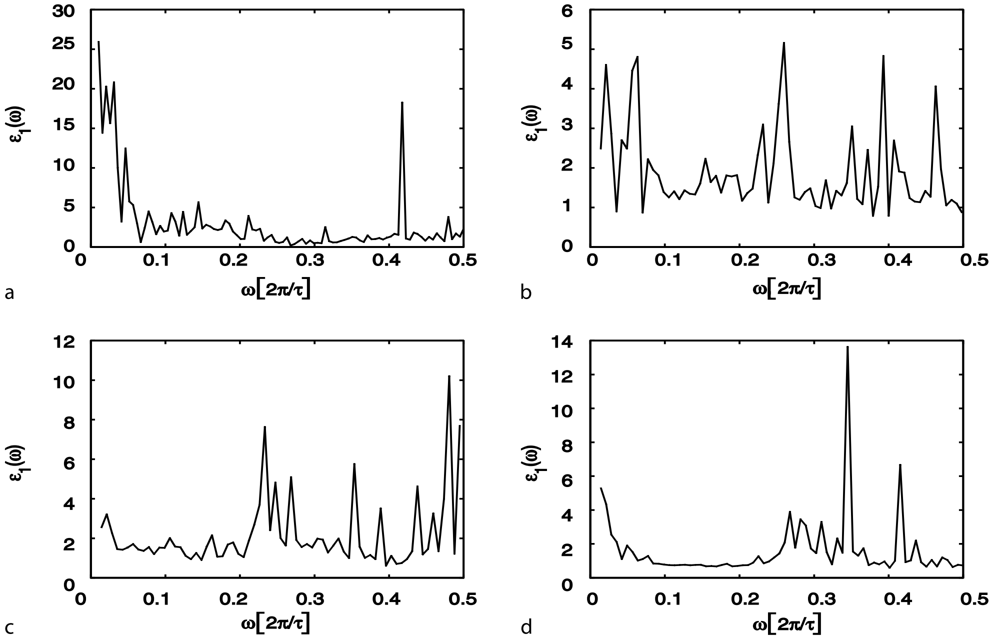
Manifestation of Strong Memory in Complex Systems

A fundamental role of the strong and weak memory in the functioning of the human organism and seismic phenomena can be illustrated by the example of some situations examined next. We will consider some examples of the time series for both living and for seismic systems. It is necessary to note that a comprehensive analysis of the experimental data includes the calculation and the presentation of corresponding phase portraits in some planes of the dynamic orthogonal variables, the autocorrelation time functions, the memory time functions and their frequency power spectra, etc. However, we start out by calculating two statistical quantifiers, characterizing two informational measures of memory: the parameters $\epsilon_1(\omega)$ and $\delta_1(\omega)$.

Figures 1 and 3 present the results of experimental data of pathological states of human cardiovascular systems (CVS). Figure 2 depicts the analysis for the seismic observation. Figures 4 and 5 indicate the memory effects for the patients with Parkinson disease (PD), and the last two Figs. 6, 7 demonstrate the key role of the strength of memory in the case of time series of patients suffering from photosensitive epilepsy which are contrasted with signals taken from healthy subjects. All these cases convincingly display the crucial role of the statistical memory in the functioning of complex (living and seismic) systems.

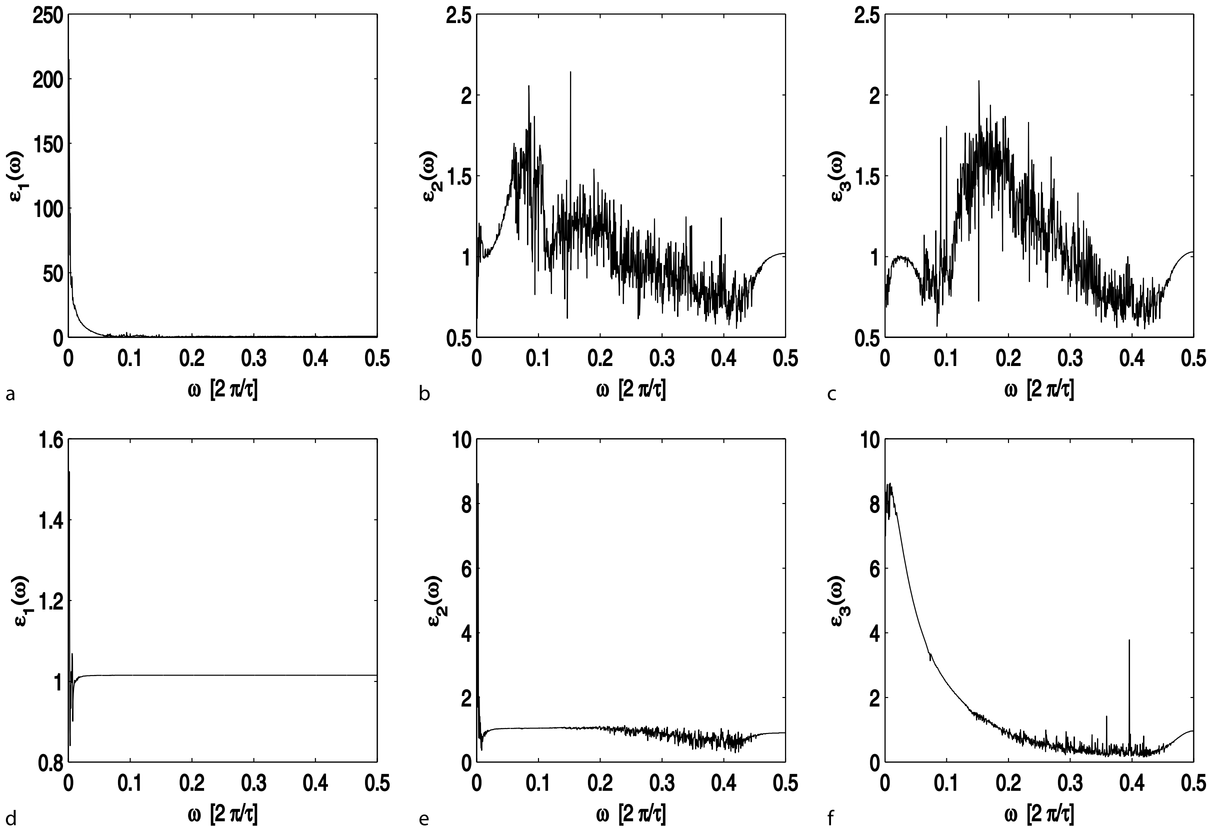
A characteristic role of the statistical memory can be detected from Fig. 1 for the typical representatives taken from patients from four different CVS-groups: (a) for healthy subject, (b) for a patient with rhythm driver migration, (c) for a patient after myocardial infarction (MI), (d) for a patient after MI with subsequent sudden cardiac death (SSCD). All these data were obtained from the short time series of the dynamics of RR-intervals from the electric signals of the human ECG's. It can be seen here that significant memory effects typically lead to the long-time correlations in the complex systems. For healthy we observe weak memory effects while and large values of the measure memory $\epsilon_1(\omega = 0) \approx 25$. The strong memory and the long memory time (approximately, 10 times more) are being observed with the help of 3 patient groups: with RDM (rhythm driver migration) (b), after MI (c) and after MI with SSCD (d).

Figure 2 depicts the strong memory effects presented in seismic phenomena. By a transition from the steady state of Earth ((a), (b) and (c)) to the state of strong earthquake (EQ) ((d), (e), and (f)) a remarkable amplification



Correlations in Complex Systems, Figure 1

Frequency spectrum of the first information measure of memory (first point in the statistical spectrum on non-Markovity parameter) $\varepsilon_1(\omega)$ for the fourth cardiac patient groups from the short time series of RR-intervals: healthy subject (a), patient with rhythm driver migration (RDM) (b), patient after myocardial infarction (MI) (c), and patient after MI with subsequent sudden cardiac death (SCD) (d). The frequency is marked in terms of units of τ^{-1}. All spectra reveal the miscellaneous faces of statistical memory's strength. For the healthy subject one can see Markov effects and weak memory. For other three cases of cardiac diseases we note the diverse displays of strong memory. The strong memory has been accompanied by the spikes of the weak memory: for RDM on the all frequency regions, for patient with MI for the middle and high frequencies and for patient after MI with SSCD only for high frequencies. From Fig. 7 in [104]



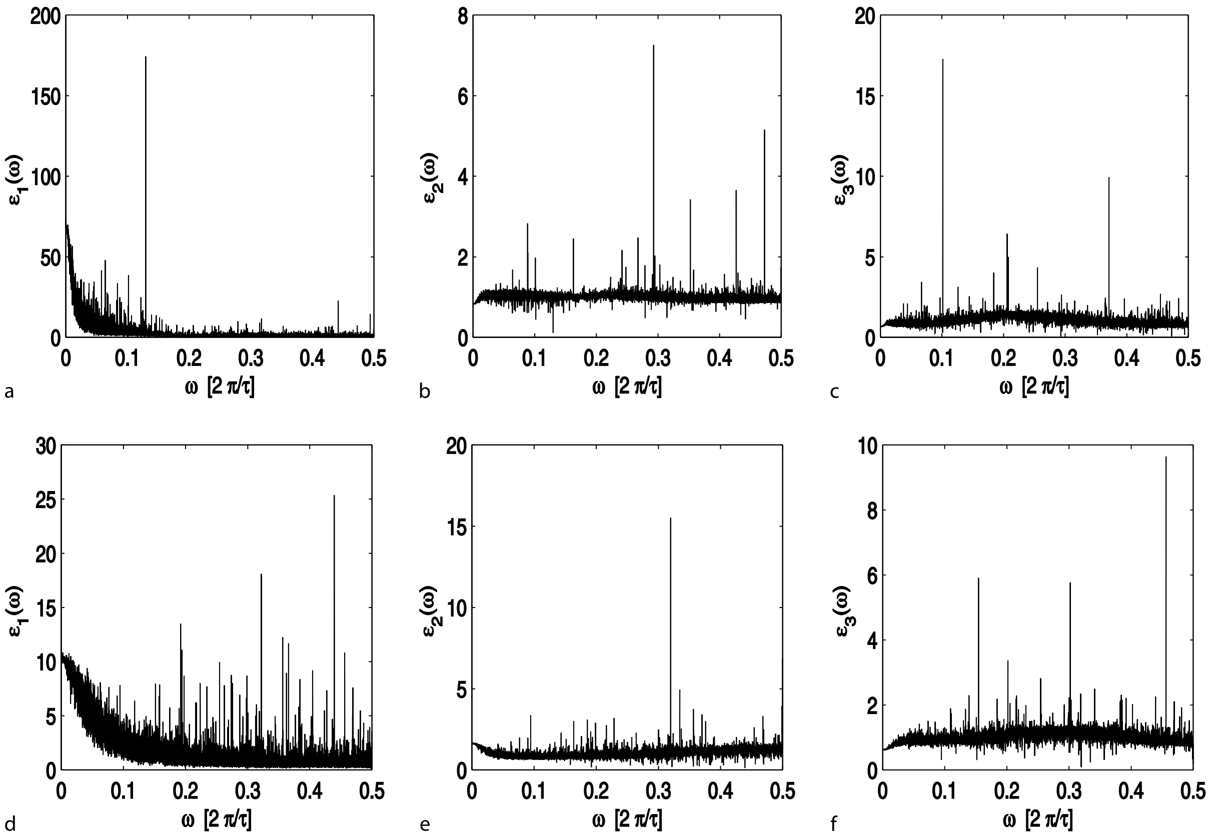
Correlations in Complex Systems, Figure 2

Frequency spectra of the first three points of the first measure of memory (non-Markovity parameters) $\varepsilon_1(\omega)$, $\varepsilon_2(\omega)$, and $\varepsilon_3(\omega)$ for the seismic phenomena: a, b, c long before the strong Earthquake (EQ) for the steady state of Earth and d, e, f during the strong EQ. Markov and quasi-Markov behavior of seismic signals with manifestation of the weak memory is observed only for ε_1 in state before the strong EQ. All remaining cases b, c, d and e relate to non-Markov processes. Strong non-Markovity and strong memory is typical for case d (state during the strong EQ). In behavior of $\varepsilon_2(\omega)$ and $\varepsilon_3(\omega)$ one can see a transition from quasi-Markovity (at low frequencies) to strong non-Markovity (at high frequencies). From Fig. 6 in [105]

of memory effects is highly visible. The term amplification refers to the appearance of strong memory and the prolongation of the memory correlation time in the seismic system. Recent study show that discrete non-Markov stochastic processes and long-range memory effects play a crucial role in the behavior of seismic systems. An approach, permitting us to obtain an algorithm of strong EQ forecasting and to differentiate technogenic explosions from weak EQs, can be developed thereupon.

Figure 3 demonstrates an intensification of memory effects of one order at the transition from healthy people ((a), (b) and (c)) to patient suffering from myocardial infarction. The figures were calculated from the long time series of the RR-intervals dynamics from the human ECG's. The zero frequency values $\epsilon_1(\omega = 0)$ at $\omega = 0$ sharply reduced, approximately of the size of one order for patient as compared to healthy subjects.

Figures 4 and 5 illustrate the behavior for patients with Parkinson's disease. Figure 4 shows time recording of the pathological tremor velocity in the left index finger of a patient with Parkinson's disease (PD) for eight diverse pathological cases (with or without medication, with or without deep brain stimulation (DBS), for various DBS, medication and time conditions). Figure 5, arranged in accordance with these conditions, displays a wide variety of the memory effects in the treatment of PD's patients. Due to the large impact of memory effects this observation permits us to develop an algorithm of exact diagnosis of Parkinson's disease and a calculation of the quantitative parameter of the quality of treatment. A physical role of the strong and long memory correlation time enables us to extract a vital information about the states of various patient on basis of notions of correlation and memory times.



Correlations in Complex Systems, Figure 3

The frequency dependence of the first three points of non-Markovity parameter (NMP) for the healthy person (a), (b), (c) and patient after myocardial infarction (MI) (d), (e), (f) from the time dynamics of RR-intervals of human ECG's for the case of the long time series. In the spectrum of the first point of NMP $\varepsilon_1(\omega)$ there is an appreciable low-frequency (long time) component, which concerns the quasi-Markov processes. Spectra NMP $\varepsilon_2(\omega)$ and NMP $\varepsilon_3(\omega)$ fully comply with non-Markov processes within the whole range of frequencies. From Fig. 6 in [106]

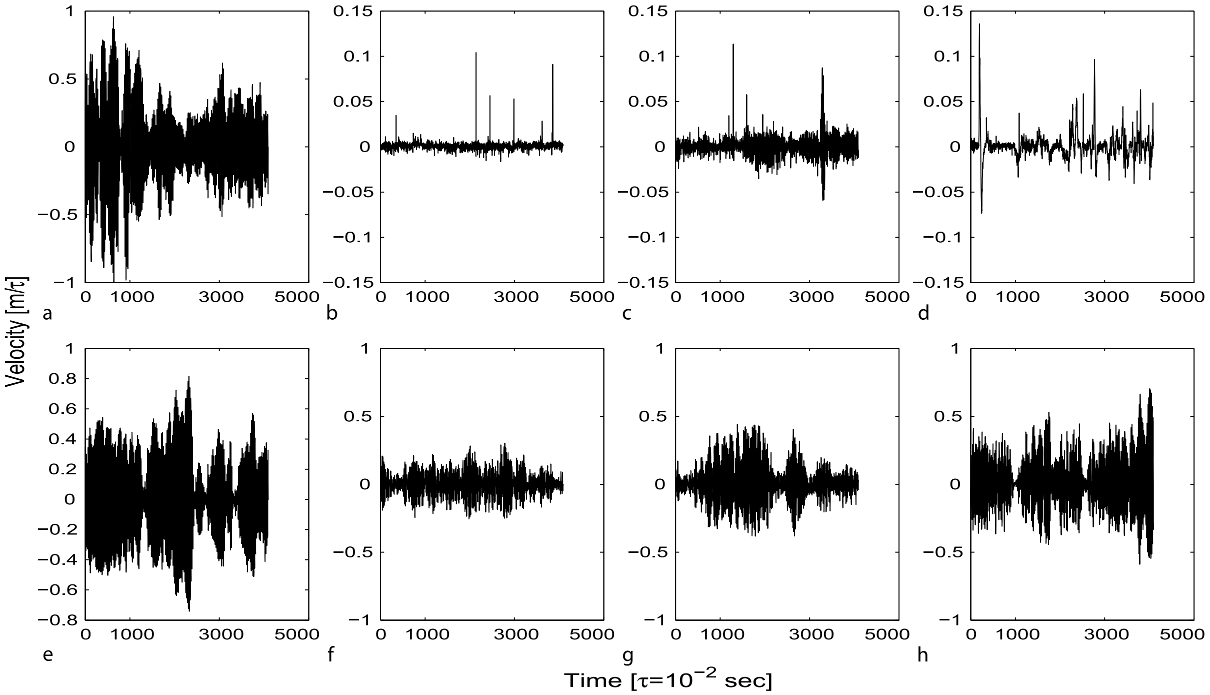
According to Figs. 6 and 7 specific information about the physiological mechanism of photosensitive epilepsy (PSE) was obtained from the analysis of the strong memory effects via the registration the neuromagnetic responses in recording of magnetoencephalogram (MEG) of the human brain core. Figure 6 presents the topographic dependence of the first level of the second memory measure $\delta_1(\omega = 0; n)$ for the healthy subjects in the whole group (upper line) vs. patients (lower line) for red/blue combination of the light stimulus. This topographic dependence of $\varepsilon_1(\omega = 0; n)$ depicted in Fig. 6 clearly demonstrates the existence of long-range time correlation. It is accompanied by a sharp increase of the role of the statistical memory effects in the all MEG's sensors with sensor numbers $n = 1, 2, \ldots, 61$ of the patient with PSE in comparison with healthy peoples. A sizable difference between the healthy subject and a subject with PSE occurs.

To emphasize the role of strong memory one can continue studying the topographic dependence in terms of the novel informational measure, the index of memory, defined as:

$$\nu(n) = \frac{\delta_1^{\text{healthy}}(0; n)}{\delta_1^{\text{patient}}(0; n)}\,, \tag{62}$$

see in Fig. 7.

This measure quantifies the detailed memory effects in the individual MEG sensors of the patient with PSE versus the healthy group. A sharp increase of the role of the memory effects in the stochastic behavior of the magnetic signals is clearly detected in sensor numbers $n = 10, 46, 51, 53$ and 59. The observed points of MEG sensors locate the regions of a protective mechanism against PSE in a human organism: frontal (sensor 10),



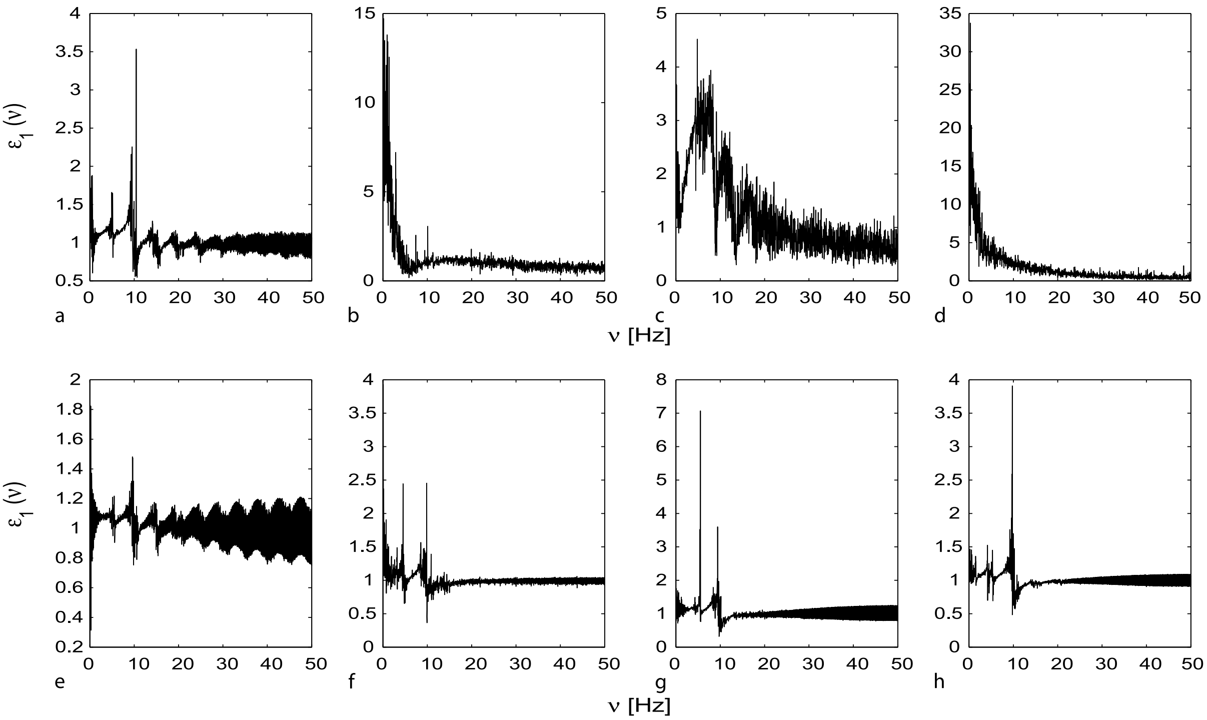
Correlations in Complex Systems, Figure 4

Pathological tremor velocity in the left index finger of the sixth patient with Parkinson's disease (PD). The registration of Parkinsonian tremor velocity is carried out for the following conditions: a "OFF-OFF" condition (no any treatment), b "ON-ON" condition (using deep brain stimulation (DBS) by electromagnetic stimulator and medicaments), c "ON-OFF" condition (DBS only), d "OFF-ON" condition (medicaments (L-Dopa) only), e–h the "15 OFF", "30 OFF", "45 OFF", "60 OFF" conditions – the patient's states 15 (30, 45, 60) minutes after the DBS is switched off, no treatment. Let's note the scale of the pathological tremor amplitude (see the vertical scale). Such representation of the time series allows us to note the increase or the decrease of pathological tremor. From Fig. 1 in [107]

occipital (sensors 46, 51 and 53) and right parietal (sensor 59) regions. The early activity in these sensors may reflect a protective mechanism suppressing the cortical hyperactivity due to the chromatic flickering.

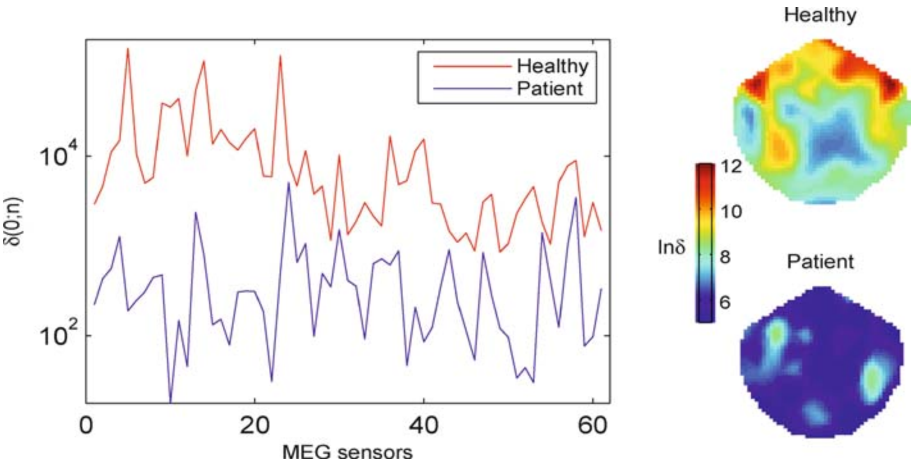
We remark that some early steps towards understanding the normal and various catastrophical states of complex systems have already been taken in many fields of science such as cardiology, physiology, medicine, neurology, clinical neurophysiology, neuroscience, seismology and so forth. With the underlying systems showing fractal and complicated spatial structures numerous studies applying the linear and nonlinear time series analysis to various complex systems have been discussed by many authors. Specifically the results obtained shows evidence of the significant nonlinear structure evident in the registered signals in the control subjects, whereas nonlinearity for the patients and catastrophical states were not detected. Moreover the couplings between distant parts and regions were found to be stronger for the control subjects. These prior findings are leading to the hypothesis that the real normal complex systems are mostly equipped with significantly nonlinear subsystems reflecting an inherent mechanism which stems against a synchronous excitation vs. outside impact or inside disturbances. Such nonlinear mechanisms are likely absent in the occurrence of catastrophical or pathological states of the complex systems.

From the physical point of view our results can be used as a toolbox for testing and identifying the presence or absence of various memory effects as they occur in complex systems. The set of our memory quantifiers is uniquely associated with the appearance of memory features in the chaotic behavior of the observed signals. The registration of the behavior belonging to these indicators, as elucidated here, is of beneficial use for detecting the catastrophical or pathological states in the complex systems. There exist alternative quantifiers of different nature as well, such as the Lyapunov's exponent, Kolmogorov–Sinai entropy, correlation dimension, etc., which are widely used in nonlinear dynamics and relevant applications. In the present context, we have found out that the employed memory measures are not only convenient for the analysis but are



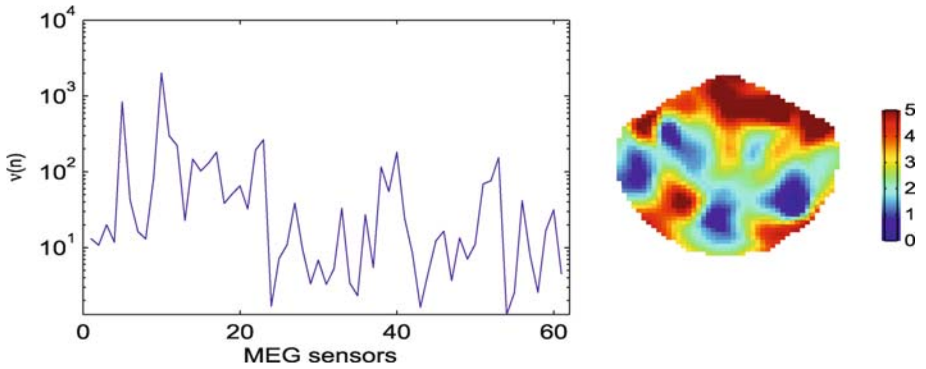
Correlations in Complex Systems, Figure 5

The frequency dependence of the first point of the non-Markovity parameter $\varepsilon_1(\nu)$ for pathological tremor velocity in the patient. As an example, the sixth patient with Parkinson's disease is chosen. The figures are submitted according to the arrangement of the initial time series. The characteristic low-frequency oscillations are observed in frequency dependence (a, e–h), which get suppressed under medical influence (b–d). The non-Markovity parameter reflects the Markov and non-Markov components of the initial time signal. The value of the parameter on zero frequency $\varepsilon_1(0)$ reflects the total dynamics of the initial time signal. The maximal values of parameter $\varepsilon_1(0)$ correspond to small amplitudes of pathological tremor velocity. The minimal values of this parameter are characteristic of significant pathological tremor velocities. The comparative analysis of frequency dependence $\varepsilon_1(\nu)$ allows us to estimate the efficiency of each method of treatment. From Fig. 5 in [107]



Correlations in Complex Systems, Figure 6

The topographic dependence of the first point of the second measure of memory $\delta_1(\omega = 0; n)$ for the healthy on average in the whole group (*upper line*) vs. patient (*lower line*) for R/B combination of the light stimulus. One can note the singular weak memory effects for the healthy on average in sensors with No. 5, 23, 14, 11 and 9



Correlations in Complex Systems, Figure 7
The topographic dependence of the memory index $\nu(n) = \nu_1(n; 0)$ for the the whole group of healthy on average vs. patient for an R/B combination of the light stimulus. Strong memory in patient vs. healthy appears clearly in sensors with No. 10, 5, 23, 40 and 53

also ideally suitable for the identification of anomalous behavior occurring in complex systems. The search for other quantifiers, and foremost, the ways of optimization of such measures when applied to the complex discrete time dynamics presents a real challenge. Especially this objective is met when attempts are made towards the identification and quantification of functioning in complex systems. This work presents initial steps towards the understanding of basic foundation of anomalous processes in complex systems on the basis of a study of the underlying memory effects and connected with this, the occurrence of long lasting correlations.

Some Perspectives on the Studies of Memory in Complex Systems

Here we present a few outlooks on the fundamental role of statistical memory in complex systems. This involves the issue of studying cross-correlations. The statistical theory of stochastic dynamics of cross-correlation can be created on the basis of the mentioned formalism of projection operators technique in the linear space of random variables. As a result we obtain the cross-correlation memory functions (MF's) revealing the statistical memory effects in complex systems. Some memory quantifiers will appear simultaneously which will reflect cross-correlation between different parts of CS. Cross-correlation MF's can be very useful for the analysis of the weak and strong interactions, signifying interrelations between the different groups of random variables in CS. Besides that the cross-correlation can be important for the problem of phase synchronization, which can find a unique way of studying of synchronization phenomena in CS that has a special importance when studying aspects of brain and living systems dynamics.

Some additional information about the strong and weak memory effects can be extracted from the observation of correlation in CS in the random event's scales. Similar effects are playing a crucial role in the differentiation between stochastic phenomena within astrophysical systems, for example, in galaxies, pulsars, quasars, microquasars, lacertides, black holes, etc. One of the most important area of application of developed approach is a bispectral and polyspectral analysis for the diverse CS. From the mathematical point of view a correct definition of the spectral properties in the functional space of random functions is quite important. A variety of MF's arises in the quantitative analysis of the fine details of memory effects in a nonlinear manner. The quantitative control of the treatment quality in the diverse areas of medicine and physiology may be one of the important biomedical application of the manifestation of the strong memory effects.

These and other features of memory effects in CS call for an advanced development of brain studies on the basis of EEG's and MEG's data, cardiovascular, locomotor and respiratory human systems, in the development of the control system of information flows in living systems. An example is the prediction of strong EQ's and the clear differentiation between the occurrence of weak EQ's and the technogenic explosions, etc.

In conclusion, we hope that the interested reader becomes invigorated by this presentation of correlation and memory analysis of the inherent nonlinear system dynamics of varying complexity. He can find further details how significant memory effects typically cause long time correlations in complex systems by inspecting more closely some of the published items in [42–103].

There are the relationships between standard fractional and polyfractal processes and long-time correlation in complex systems, which were explained in [39,40,44,45,46,49,53,54,60,62,64,76,79,83,84,94] in detail.

Example of using the Hurst exponent over time for testing the assertion that emerging markets are becoming more efficient can be found in [51].

While over 30 measures of complexity have been proposed in the research literature one can distinguish [42,55,66,81,89,99] with the specific designation of long-time correlation and memory effects.

The papers [48,57] are focused on long range correlation processes that are nonlocal in time and whence show memory effects.

The statistical characterization of the nonstationarities in real-world time series is an important topic in many fields of research and some numerous methods of characterizing nonstationary time series were offered in [59,65,84].

Long-range correlated time series have been widely used in [52,61,63,68,74] for the theoretical description of diverse phenomena.

Example of the study an anatomy of extreme events in a complex adaptive system can be found in [67].

Approaches for modeling long-time and long-range correlation in complex systems from time series are investigated and applied to different examples in [50,56,69,70,73,75,80,82,86,100,101,102].

Detecting scale invariance and its fundamental relationships with statistical structures is one of the most relevant problems among those addressed correlation analysis [47,71,72,91].

Specific long-range correlation in complex systems are the object of active research due to its implications in the technology of materials and in several fields of scientific knowledge with the use of quantified histograms [78], decrease of chaos in heart failure [85], scaling properties of ECG's signals fluctuations [87], transport properties in correlated systems [88] etc.

It is demonstrated in [43,92,93] how ubiquity of the long-range correlations is apparent in typical and exotic complex statistical systems with application to biology, medicine, economics and to time clustering properties [95,98].

The scale-dependent wavelet and spectral measures for assessing cardiac dysfunction have been used in [97].

In recent years the study of an increasing number of natural phenomena that appear to deviate from standard statistical distributions has kindled interest in alternative formulations of statistical mechanics [58,101].

At last, papers [77,90] present the samples of the deep and multiple interplay between discrete and continuous long-time correlation and memory in complex systems and the corresponding modeling the discrete time series on the basis of physical Zwanzig–Mori's kinetic equation for the Hamilton statistical systems.

Bibliography

Primary Literature

1. Markov AA (1906) Two-dimensional Brownian motion and harmonic functions. Proc Phys Math Soc Kazan Imp Univ 15(4):135–178; in Russian
2. Chapman S, Couling TG (1958) The mathematical theory of nonuniform gases. Cambridge University Press, Cambridge
3. Albeverio S, Blanchard P, Steil L (1990) Stochastic processes and their applications in mathematics and physics. Kluwer, Dordrecht
4. Rice SA, Gray P (1965) The statistical mechanics of simple liquids. Interscience, New York
5. Kubo R, Toda M, Hashitsume N, Saito N (2003) Statistical physics II: Nonequilibrium statistical mechanics. In: Fulde P (ed) Springer Series in Solid-State Sciences, vol 31. Springer, Berlin, p 279
6. Ginzburg VL, Andryushin E (2004) Superconductivity. World Scientific, Singapore
7. Sachs I, Sen S, Sexton J (2006) Elements of statistical mechanics. Cambridge University Press, Cambridge
8. Fetter AL, Walecka JD (1971) Quantum theory of many-particle physics. Mc Graw-Hill, New York
9. Chandler D (1987) Introduction to modern statistical mechanics. Oxford University Press, Oxford
10. Zwanzig R (2001) Nonequilibrium statistical mechanics. Cambridge University Press, Cambridge
11. Zwanzig R (1961) Memory effects in irreversible thermodynamics. Phys Rev 124:983–992
12. Mori H (1965) Transport, collective motion and Brownian motion. Prog Theor Phys 33:423–455; Mori H (1965) A continued fraction representation of the time correlation functions. Prog Theor Phys 34:399–416
13. Grabert H, Hänggi P, Talkner P (1980) Microdynamics and nonlinear stochastic processes of gross variables. J Stat Phys 22:537–552
14. Grabert H, Talkner P, Hänggi P (1977) Microdynamics and time-evolution of macroscopic non-Markovian systems. Z Physik B 26:389–395
15. Grabert H, Talkner P, Hänggi P, Thomas H (1978) Microdynamics and time-evolution of macroscopic non-Markovian systems II. Z Physik B 29:273–280
16. Hänggi P, Thomas H (1977) Time evolution, correlations and linear response of non-Markov processes. Z Physik B 26:85–92
17. Hänggi P, Talkner P (1983) Memory index of first-passage time: A simple measure of non-Markovian character. Phys Rev Lett 51:2242–2245
18. Hänggi P, Thomas H (1982) Stochastic processes: Time-evolution, symmetries and linear response. Phys Rep 88:207–319
19. Lee MH (1982) Orthogonalization process by recurrence relations. Phys Rev Lett 49:1072–1072; Lee MH (1983) Can the velocity autocorrelation function decay exponentially? Phys Rev Lett 51:1227–1230
20. Balucani U, Lee MH, Tognetti V (2003) Dynamic correlations. Phys Rep 373:409–492

21. Hong J, Lee MH (1985) Exact dynamically convergent calculations of the frequency-dependent density response function. Phys Rev Lett 55:2375–2378
22. Lee MH (2000) Heisenberg, Langevin, and current equations via the recurrence relations approach. Phys Rev E 61:3571–3578; Lee MH (2000) Generalized Langevin equation and recurrence relations. Phys Rev E 62:1769–1772
23. Lee MH (2001) Ergodic theory, infinite products, and long time behavior in Hermitian models. Phys Rev Lett 87(1–4):250601
24. Kubo R (1966) Fluctuation-dissipation theorem. Rep Progr Phys 29:255–284
25. Kawasaki K (1970) Kinetic equations and time correlation functions of critical fluctuations. Ann Phys 61:1–56
26. Michaels IA, Oppenheim I (1975) Long-time tails and Brownian motion. Physica A 81:221–240
27. Frank TD, Daffertshofer A, Peper CE, Beek PJ, Haken H (2001) H-theorem for a mean field model describing coupled oscillator systems under external forces. Physica D 150:219–236
28. Vogt M, Hernandez R (2005) An idealized model for nonequilibrium dynamics in molecular systems. J Chem Phys 123(1–8):144109
29. Sen S (2006) Solving the Liouville equation for conservative systems: Continued fraction formalism and a simple application. Physica A 360:304–324
30. Prokhorov YV (1999) Probability and mathematical statistics (encyclopedia). Scien Publ Bolshaya Rossiyskaya Encyclopedia, Moscow
31. Yulmetyev R et al (2000) Stochastic dynamics of time correlation in complex systems with discrete time. Phys Rev E 62:6178–6194
32. Yulmetyev R et al (2002) Quantification of heart rate variability by discrete nonstationarity non-Markov stochastic processes. Phys Rev E 65(1–15):046107
33. Reed M, Samon B (1972) Methods of mathematical physics. Academic, New York
34. Graber H (1982) Projection operator technique in nonequilibrium statistical mechanics. In: Höhler G (ed) Springer tracts in modern physics, vol 95. Springer, Berlin
35. Yulmetyev RM (2001) Possibility between earthquake and explosion seismogram differentiation by discrete stochastic non-Markov processes and local Hurst exponent analysis. Phys Rev E 64(1–14):066132
36. Abe S, Suzuki N (2004) Aging and scaling of earthquake aftershocks. Physica A 332:533–538
37. Tirnakli U, Abe S (2004) Aging in coherent noise models and natural time. Phys Rev E 70(1–4):056120
38. Abe S, Sarlis NV, Skordas ES, Tanaka HK, Varotsos PA (2005) Origin of the usefulness of the natural-time representation of complex time series. Phys Rev Lett 94(1–4):170601
39. Stanley HE, Meakin P (1988) Multifractal phenomena in physics and chemistry. Nature 335:405–409
40. Ivanov P Ch, Amaral LAN, Goldberger AL, Havlin S, Rosenblum MG, Struzik Z, Stanley HE (1999) Multifractality in human heartbeat dynamics. Nature 399:461–465
41. Mokshin AV, Yulmetyev R, Hänggi P (2005) Simple measure of memory for dynamical processes described by a generalized Langevin equation. Phys Rev Lett 95(1–4):200601
42. Allegrini P et al (2003) Compression and diffusion: A joint approach to detect complexity. Chaos Soliton Fractal 15: 517–535
43. Amaral LAN et al (2001) Application of statistical physics methods and concepts to the study of science and technology systems. Scientometrics 51:9–36
44. Arneodo A et al (1996) Wavelet based fractal analysis of DNA sequences. Physica D 96:291–320
45. Ashkenazy Y et al (2003) Magnitude and sign scaling in power-law correlated time series. Physica A Stat Mech Appl 323:19–41
46. Ashkenazy Y et al (2003) Nonlinearity and multifractality of climate change in the past 420,000 years. Geophys Res Lett 30:2146
47. Azbel MY (1995) Universality in a DNA statistical structure. Phys Rev Lett 75:168–171
48. Baldassarri A et al (2006) Brownian forces in sheared granular matter. Phys Rev Lett 96:118002
49. Baleanu D et al (2006) Fractional Hamiltonian analysis of higher order derivatives systems. J Math Phys 47:103503
50. Blesic S et al (2003) Detecting long-range correlations in time series of neuronal discharges. Physica A 330:391–399
51. Cajueiro DO, Tabak BM (2004) The Hurst exponent over time: Testing the assertion that emerging markets are becoming more efficient. Physica A 336:521–537
52. Brecht M et al (1998) Correlation analysis of corticotectal interactions in the cat visual system. J Neurophysiol 79: 2394–2407
53. Brouersa F, Sotolongo-Costab O (2006) Generalized fractal kinetics in complex systems (application to biophysics and biotechnology). Physica A 368(1):165–175
54. Coleman P, Pietronero L (1992) The fractal structure of the universe. Phys Rep 213:311–389
55. Goldberger AL et al (2002) What is physiologic complexity and how does it change with aging and disease? Neurobiol Aging 23:23–26
56. Grau-Carles P (2000) Empirical evidence of long-range correlations in stock returns. Physica A 287:396–404
57. Grigolini P et al (2001) Asymmetric anomalous diffusion: An efficient way to detect memory in time series. Fractal-Complex Geom Pattern Scaling Nat Soc 9:439–449
58. Ebeling W, Frommel C (1998) Entropy and predictability of information carriers. Biosystems 46:47–55
59. Fukuda K et al (2004) Heuristic segmentation of a nonstationary time series. Phys Rev E 69:021108
60. Hausdorff JM, Peng CK (1996) Multiscaled randomness: A possible source of 1/f noise in biology. Phys Rev E 54: 2154–2157
61. Herzel H et al (1998) Interpreting correlations in biosequences. Physica A 249:449–459
62. Hoop B, Peng CK (2000) Fluctuations and fractal noise in biological membranes. J Membrane Biol 177:177–185
63. Hoop B et al (1998) Temporal correlation in phrenic neural activity. In: Hughson RL, Cunningham DA, Duffin J (eds) Advances in modelling and control of ventilation. Plenum Press, New York, pp 111–118
64. Ivanova K, Ausloos M (1999) Application of the detrended fluctuation analysis (DFA) method for describing cloud breaking. Physica A 274:349–354
65. Ignaccolo M et al (2004) Scaling in non-stationary time series. Physica A 336:595–637
66. Imponente G (2004) Complex dynamics of the biological rhythms: Gallbladder and heart cases. Physica A 338:277–281

67. Jefferiesa P et al (2003) Anatomy of extreme events in a complex adaptive system. Physica A 318:592–600
68. Karasik R et al (2002) Correlation differences in heartbeat fluctuations during rest and exercise. Phys Rev E 66:062902
69. Kulessa B et al (2003) Long-time autocorrelation function of ECG signal for healthy versus diseased human heart. Acta Phys Pol B 34:3–15
70. Kutner R, Switala F (2003) Possible origin of the non-linear long-term autocorrelations within the Gaussian regime. Physica A 330:177–188
71. Koscielny-Bunde E et al (1998) Indication of a universal persistence law governing atmospheric variability. Phys Rev Lett 81:729–732
72. Labini F (1998) Scale invariance of galaxy clustering. Phys Rep 293:61–226
73. Linkenkaer-Hansen K et al (2001) Long-range temporal correlations and scaling behavior in human brain oscillations. J Neurosci 21:1370–1377
74. Mercik S et al (2000) What can be learnt from the analysis of short time series of ion channel recordings. Physica A 276:376–390
75. Montanari A et al (1999) Estimating long-range dependence in the presence of periodicity: An empirical study. Math Comp Model 29:217–228
76. Mark N (2004) Time fractional Schrodinger equation. J Math Phys 45:3339–3352
77. Niemann M et al (2008) Usage of the Mori–Zwanzig method in time series analysis. Phys Rev E 77:011117
78. Nigmatullin RR (2002) The quantified histograms: Detection of the hidden unsteadiness. Physica A 309:214–230
79. Nigmatullin RR (2006) Fractional kinetic equations and universal decoupling of a memory function in mesoscale region. Physica A 363:282–298
80. Ogurtsov MG (2004) New evidence for long-term persistence in the sun's activity. Solar Phys 220:93–105
81. Pavlov AN, Dumsky DV (2003) Return times dynamics: Role of the Poincare section in numerical analysis. Chaos Soliton Fractal 18:795–801
82. Paulus MP (1997) Long-range interactions in sequences of human behavior. Phys Rev E 55:3249–3256
83. Peng C-K et al (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685–1689
84. Peng C-K et al (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82–87
85. Poon CS, Merrill CK (1997) Decrease of cardiac chaos in congestive heart failure. Nature 389:492–495
86. Rangarajan G, Ding MZ (2000) Integrated approach to the assessment of long range correlation in time series data. Phys Rev E 61:4991–5001
87. Robinson PA (2003) Interpretation of scaling properties of electroencephalographic fluctuations via spectral analysis and underlying physiology. Phys Rev E 67:032902
88. Rizzo F et al (2005) Transport properties in correlated systems: An analytical model. Phys Rev B 72:155113
89. Shen Y et al (2003) Dimensional complexity and spectral properties of the human sleep EEG. Clinic Neurophysiol 114:199–209
90. Schmitt D et al (2006) Analyzing memory effects of complex systems from time series. Phys Rev E 73:056204
91. Soen Y, Braun F (2000) Scale-invariant fluctuations at different levels of organization in developing heart cell networks. Phys Rev E 61:R2216–R2219
92. Stanley HE et al (1994) Statistical-mechanics in biology – how ubiquitous are long-range correlations. Physica A 205: 214–253
93. Stanley HE (2000) Exotic statistical physics: Applications to biology, medicine, and economics. Physica A 285:1–17
94. Tarasov VE (2006) Fractional variations for dynamical systems: Hamilton and Lagrange approaches. J Phys A Math Gen 39:8409–8425
95. Telesca L et al (2003) Investigating the time-clustering properties in seismicity of Umbria-Marche region (central Italy). Chaos Soliton Fractal 18:203–217
96. Turcott RG, Teich MC (1996) Fractal character of the electrocardiogram: Distinguishing heart-failure and normal patients. Ann Biomed Engin 24:269–293
97. Thurner S et al (1998) Receiver-operating-characteristic analysis reveals superiority of scale-dependent wavelet and spectral measures for assessing cardiac dysfunction. Phys Rev Lett 81:5688–5691
98. Vandewalle N et al (1999) The moving averages demystified. Physica A 269:170–176
99. Varela M et al (2003) Complexity analysis of the temperature curve: New information from body temperature. Eur J Appl Physiol 89:230–237
100. Varotsos PA et al (2002) Long-range correlations in the electric signals that precede rupture. Phys Rev E 66:011902
101. Watters PA (2000) Time-invariant long-range correlations in electroencephalogram dynamics. Int J Syst Sci 31:819–825
102. Wilson PS et al (2003) Long-memory analysis of time series with missing values. Phys Rev E 68:017103
103. Yulmetyev RM et al (2004) Dynamical Shannon entropy and information Tsallis entropy in complex systems. Physica A 341:649–676
104. Yulmetyev R, Hänggi P, Gafarov F (2000) Stochastic dynamics of time correlation in complex systems with discrete time. Phys Rev E 62:6178
105. Yulmetyev R, Gafarov F, Hänggi P, Nigmatullin R, Kayumov S (2001) Possibility between earthquake and explosion seismogram processes and local Hurst exponent analysis. Phys Rev E 64:066132
106. Yulmetyev R, Hänggi P, Gafarov F (2002) Quantification of heart rate variability by discrete nonstationary non-Markov stochastic processes. Phys Rev E 65:046107
107. Yulmetyev R, Demin SA, Panischev OY, Hänggi P, Timashev SF, Vstovsky GV (2006) Regular and stochastic behavior of Parkinsonian pathological tremor signals. Physica A 369:655

Books and Reviews

Badii R, Politi A (1999) Complexity: Hierarchical structures and scaling in physics. Oxford University Press, New York

Elze H-T (ed) (2004) Decoherence and entropy in complex systems. In: Selected lectures from DICE 2002 series: Lecture notes in physics, vol 633. Springer, Heidelberg

Kantz H, Schreiber T (2004) Nonlinear time series analysis. Cambridge University Press, Cambridge

Mallamace F, Stanley HE (2004) The physics of complex systems (new advances and perspectives). IOS Press, Amsterdam

Parisi G, Pietronero L, Virasoro M (1992) Physics of complex systems: Fractals, spin glasses and neural networks. Physica A 185(1–4):1–482

Sprott JC (2003) Chaos and time-series analysis. Oxford University Press, New York

Zwanzig R (2001) Nonequilibrium statistical physics. Oxford University Press, New York

Cosmic Gravitational Background, Stochastic

CARLO UNGARELLI
CNR-Institute of Geosciences and Earth Resources,
Pisa, Italy

Article Outline

Glossary

Cosmic microwave background Quasi-isotropic thermal, background of electromagnetic radiation of cosmological (i. e. not astrophysical origin). The observed spectrum of this radiation is consistent with a black body with temperature $T = 2.725\,\mathrm{K}$ and represents the best experimental confirmation of the Planck spectrum. According to the standard cosmological model, the dynamics of the primordial universe is determined, via the Einstein gravitational equations, by a hot relativistic plasma of electrons, photons and baryons. The photons interact with the plasma through Thomson scattering. As the universe expanded, the plasma cools until it becomes favorable for electrons to combine with protons and form hydrogen atoms. This occurs at a temperature of about 3×10^3 K (corresponding to an average energy of about 0.3 eV). At this point, the photons scatter off the now neutral atoms and start to travel freely through space. This process is called recombination or decoupling (referring to electrons combining with nuclei and to the decoupling of matter and radiation, respectively). Since the universe is expanding the photon temperature keeps decreasing reaching the current value of $T = 2.725\,\mathrm{K}$. Accordingly, the photons observed today (i. e. the CMB photons) come from a spherical surface, called the **surface of last scattering**, corresponding to the time of decoupling.

Friedmann–Robertson–Walker model Model based upon an isotropic and homogeneous metric describing an expanding universe corresponding to a solution of the Einstein gravitational equation with a perfect fluid source (described in terms of its energy density ρ and its pressure p). The line element depends on the three-dimensional constant curvature k and the scale factor $a(t)$ describing the expansion: $\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)(\mathrm{d}r^2/(1 - kr^2) + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2))$. The expansion rate is given by $H = \dot{a}/a$. Given the equation of state $p = p(\rho)$, the solution for the scale factor is obtained from the **Friedmann equation** $H^2 = 8\pi G\rho/3$.

Inflation Primordial epoch in the evolution of the Universe characterized by an accelerating expansion ($\ddot{a} > 0$). In particular, to achieve acceleration, a nonstandard source is needed (i. e. a perfect fluid with an effective equation of state such that $\rho + 3p < 0$). The existence of such a phase is postulated in order to explain the observed quasi-isotropy and homogeneity of the Universe on large scales, as well as to provide a mechanism to generate primordial seeds for structure formation.

Definition of the Subject

As first predicted by Albert Einstein [1], propagation effects at finite speed in the dynamical gravitational equations yield to the existence of wave-like solutions of the gravitational field linearized around the flat-vacuum configuration. Such solutions correspond to the so-called gravitational waves (GWs). At the theoretical level, the work of Bondi [2] (who considered self-gravitating systems like binaries of compact objects) showed that GWs carry energy, while observationally the discovery of the 1916 + 13 binary pulsar system by Hulse and Taylor [3] (in particular, the speed-up of the binary pulsar period), provided an indirect evidence for the existence of GWs.

On the experimental side, after the pioneering work of Weber in the 1960s and the subsequent development of meter-scale, resonant bar detectors (ALLEGRO, AURIGA EXPLORER, NAUTILUS and NIOBE), in more recent years a number of new detectors have been designed and built to search for GWs. Kilometer-scale interferometers (GEO [4], LIGO [5], TAMA [6] and VIRGO [7]), sensitive to GWs in the frequency range between 10 Hz and 1 kHz, are already producing scientific runs at the design sensitivity (or close to it, as for VIRGO). The Laser Interferometric Space Antenna (LISA) [8], a space-based, three-arms interferometer will be aimed at detecting GWs in the frequency range between 0.1 mHz and 10 mHz, while a next generation of space-detectors especially designed to detect primordial GWs has already being planned [9,10].

Gravitational waves can be produced by violent dynamical phenomena involving compact astrophysical objects (like neutron stars and/or black holes), as well as by various mechanisms related to the dynamics of the primordial universe. Those primordial GWs produce a background of gravitational radiation that can be considered as a sort of "cosmological noise", i. e. the gravitational counterpart of the cosmic microwave background radiation (CMB). The detection of such primordial gravitational radiation can be exploited to investigate physical processes occurring during the early stages of the evolution of the universe, when the typical energy scales are not accessible experimentally in any other way. The reason why primordial GWs are characterized by this unique feature is the well-known fact that in an expanding universe particles decoupled from the primordial plasma at a temperature T_D carry information about the state of the Universe at T_D. In particular, for gravitons it is possible to estimate the decoupling temperature assuming that the gravitons are kept in thermal equilibrium through two-body scattering processes (see e. g. [11]). In this case, the interaction rate is[1]. $\Gamma_{\mathrm{G}} \sim T^5/M_{\mathrm{pl}}^4$, where M_{pl} is the Planck mass ($\sim 10^{19}$ GeV). Since the energy density for relativistic particles scales with the fourth power of the temperature, assuming an adiabatic expansion the expansion rate H scales as $H \sim \dot{T}/T$. Hence, thermal equilibrium is maintained provided that the typical time-scale for graviton interaction is smaller than the expansion time-scale, i. e. $\Gamma_{\mathrm{G}} > H$. Using the Friedmann equation, $H \propto T^2/M_{\mathrm{pl}}$ this yields

$$\left(\frac{\Gamma_{\mathrm{G}}}{H}\right) \sim \left(\frac{T}{M_{\mathrm{pl}}}\right)^3 . \tag{1}$$

Gravitons are therefore decoupled just below the Planck scale. This implies that any relic GWs retain in their spectrum (i. e., amplitude and frequency) unique information about the state of the primordial universe and the corresponding physical process occurring at extremely high energy scales.

The first papers addressing possible mechanisms for the production of a cosmological background of gravitational radiation date back to the 1970s. Grishchuck [12] and Starobinsky [13] showed that metric tensor perturbations in an expanding, homogeneous and isotropic universe can be parametrically amplified provided there is a stage in the evolution of the universe when the characteristic time scale during which the metric changes is much smaller than the intrinsic wave period of the perturbations. This phenomenon, known as parametric amplification of vacuum fluctuations, is characteristic of inflationary models and its phenomenological consequences have been widely investigated (see e. g. [14]). Furthermore, as Witten [15] and Hogan [16] first pointed out, cosmological first-order phase transitions (e. g. expected to occur at the QCD scale ($T \sim 100$ MeV) and/or at the electro-weak scale ($T \sim 100$ GeV)) may be associated to the production of gravitational radiation. In particular, Hogan and Witten showed that in a cosmological, first-order phase transition, both the collision of detonation waves triggered by the nucleation of the low-temperature phase and the generation of acoustic noise in the relativistic plasma could produce a relic background of gravitational waves. Since those pioneering works, detailed computations of various types of gravitational wave signals have been performed, and data analysis techniques for extracting this kind of signal have been developed.

Introduction

Without loss of generality, a cosmological gravitational background is expected to be Gaussian, isotropic, stationary and unpolarized (for a detailed discussion see e. g. [17]). All the physical properties are then encoded into the so-called *spectrum*, given by the following, dimensionless quantity

$$\Omega_{\mathrm{gw}}(f) = \frac{1}{\rho_{\mathrm{c}}} f\, \tilde{\rho}_{\mathrm{gw}}(f) , \tag{2}$$

where f is the frequency, $\tilde{\rho}_{\mathrm{gw}}(f)$ is the gravitational wave energy density per unit frequency, i. e.

$$\rho_{\mathrm{gw}} = \int_0^\infty \mathrm{d}f\, \tilde{\rho}_{\mathrm{gw}}(f) , \tag{3}$$

and $\rho_{\mathrm{c}} = 3H_0^2/(8\pi G)$ is the critical energy density ($H_0 = h_0 \times 100\,\mathrm{km/(s\,Mpc)}$ being the current value of the Hubble constant, with h_0 parameterizing the current experimental uncertainty). Hence, to characterize any stochastic background of gravitational radiation, the combination $h_0^2\, \Omega_{\mathrm{gw}}$ is adopted.

As a signal to be detected, a stochastic background of gravitational waves is described in terms of a random process that cannot be disentangled from the detector noise. Most of the theoretical predictions yield values for the GWs spectrum much lower than the noise level

[1] In the following, we will adopt units for which temperatures, masses and energies are all measured in electron-volts (eV) or multiples of it and distances are expressed as inverse of energies

of any current or planned earth-based detector; on the other hand, the instrumental noise features of GW detectors are not sufficiently well understood to search for excess noise. For these reasons, an optimal detection strategy has been developed [18] based upon the cross-correlation between two or more widely separated detectors (so as to minimize the common noise sources). In particular, the cross-correlation improves the sensitivity (in terms of the spectrum) by a factor $(\Delta f\, T_{\text{obs}})^{1/2}$, where Δf is the effective detector bandwidth and T_{obs} is the total observation time. Using the design LIGO sensitivity, assuming four months of observation time and a constant spectrum a stochastic background of GWs could be detected at 90% confidence level provided that $h_0^2\, \Omega_{\text{gw}} > 5 \times 10^{-6}$ [19] (the recent analysis of LIGO data [20] sets a 95% confidence level upper limit $h_0^2\, \Omega_{\text{gw}} < 6.5 \times 10^{-5}$), while the advanced-LIGO design sensitivity sets a lower detection limit $\Omega_{\text{gw}} \sim 8 \times 10^{-9}$ [19]. As for the planned, three-arm space-based interferometer LISA, a possible strategy for detecting a stochastic background of gravitational radiation exploits some combinations of the signals from the three spacecraft that are insensitive to any GW signal (the so-called Sagnac combination) [21]. Such combinations therefore provide a calibration channel to measure the instrumental noise power spectrum. However, beside primordial GWs, in the frequency range 0.1–10 mHz an astrophysical background is also expected due to the incoherent superposition of GWs emitted by large populations (both galactic and extra-galactic) of binary systems of compact objects. Such an astrophysical background provides an additional noise contribution for the detection of any primordial gravitational radiation in the LISA frequency band.

In general the intensity and the spectral content of a generic cosmological background of gravitational radiation depends upon the producing mechanism and the underlying theoretical model. However, it is possible to characterize primordial gravitational radiation in terms of both some observational constraints and some model-independent features.

Observational Bounds

There are three main observational bounds that a cosmological background of gravitational radiation must satisfy: the nucleosynthesis bound, the COBE/CMB bound and the pulsar-timing bound.

The *nucleosynthesis bound* [22] is related to the highly-accurate predictions of the standard cosmological model pertaining to the primordial abundances of light elements. Such abundances depend upon the ratio between the number density of neutrons n_{n} and the number density of protons n_{p}. In conditions of thermal equilibrium $n_{\text{n}}/n_{\text{p}} \sim \mathrm{e}^{-\Delta m/T}$, where $\Delta m = m_{\text{n}} - m_{\text{p}} \sim 1.3\,\text{MeV}$. In particular, the process responsible for maintaining thermal equilibrium is $p + e^- \rightleftarrows n + \nu_e$, as long as the corresponding rate Γ is greater than the expansion rate H. When this condition is no longer fulfilled, the ratio $n_{\text{n}}/n_{\text{p}}$ freezes out at a value $\mathrm{e}^{-\Delta m/T_{\text{f}}}$, and this determines the number of neutrons yielding the formation of light elements (mainly He^4). Since the Hubble constant is given by the Friedmann equation $H^2 = 8\pi\, G\, \rho/3$, where the energy density ρ incorporates all the possible contributions at the nucleosynthesis epoch, its value at the freeze-out depends also upon a possible contribution due to a gravitational radiation component of cosmological origin. Hence, a constraint on the value of the freeze-out temperature yields an upper-limit for the energy density associated to any extra-contribution other than standard model particles. Assuming that the only extra contribution is due to cosmological gravitons, the corresponding spectrum must satisfy the following bound [23]

$$\int_0^\infty \mathrm{d}f\, h_0^2 \Omega_{\text{gw}}(f) \lesssim 5 \times 10^{-6} \qquad (4)$$

COBE/CMB bound. On large scales, the spectrum amplitude of a stochastic background of cosmological origin must be compatible with the measurements of the CMB temperature fluctuations. Indeed, at large wavelengths, a stochastic background of gravitational radiation produces a dynamical red-shift on the frequencies of the CMB photons, thus generating fluctuations in their temperature (according to the so-called Sachs–Wolfe effect [24]). Those fluctuations have been measured firstly by the COBE experiment [25] and more recently by WMAP [26]. Since on large scales also density perturbations (inhomogeneities in the matter distribution) generate fluctuations in the CMB temperature, any gravitational wave spectrum must satisfy the following bound [27]

$$h_0^2\, \Omega_{\text{gw}}(f) \lesssim \left(\frac{H_0}{f}\right)^2 \left\langle\frac{\delta T}{T}\right\rangle^2 , \qquad (5)$$

where $\langle \delta T/T \rangle \sim 5 \times 10^{-6}$ is the CMB temperature rms fluctuation (taking only the quadrupole contribution). The frequency range over which the constraint (5) holds corresponds to modes that are currently inside the causal horizon today ($f > 3 \times 10^{-8}\,\text{Hz}$) and were associated to super-horizon scales at the surface of last-scattering ($f < 10^{-16}\,\text{Hz}$). At the maximum frequency ($10^{-16}\,\text{Hz}$) the bound therefore yields an upperlimit for any primordial GW spectrum $h_0^2\, \Omega_{\text{gw}} \lesssim 10^{-14}$. This upperlimit is particularly relevant for gravitational backgrounds produced by the amplification of vacuum fluctuations occurring during a primordial, inflationary era: the corresponding spectra are indeed characterized by a wide frequency window ranging from $f \sim H_0 \sim 3 \times 10^{-18}$ Hz up to GHz values.

Recently, it has been pointed out [28] that since any cosmological background of gravitational radiation behaves as a gas of free-streaming particles, in addition to its effect on the expansion rate at the nucleosynthesis epoch, also the growth of density perturbations is affected. In particular, if the primordial metric perturbations yielding to a background of gravitational radiation are not adiabatic (as occurs for the parametrically amplified perturbations during inflation), then the effects on the CMB photons are different from those associated to adiabatic massless neutrinos. Using the recent WMAP data, this feature allows us to put a constraint on the spectrum of primordial gravitational radiation down to frequencies corresponding to the comoving horizon at decoupling ($f \sim 10^{-15}$ Hz). At those frequencies, the analysis carried out in [28] establishes a current 95% confidence-level upper limit $h_0^2\, \Omega_{\mathrm{gw}} \lesssim 8.4 \times 10^{-6}$, while future CMB experiments (like Planck) could reach sensitivities such that $h_0^2\, \Omega_{\mathrm{gw}} \lesssim 1.4 \times 10^{-6}$ [28].

Pulsar timing bound. A consistent fraction of known pulsars with a period in the range [1.5–30] ms (the so-called millisecond pulsars) are characterized by an extremely stable integrated pulse profile and therefore may be considered as natural clocks. Those astrophysical systems are natural detectors of gravitational radiation: a GW traveling between a detector and a pulsar produces a fluctuation in the arrival time of the pulse proportional to the wave amplitude. For a total observation time T_{obs} and an uncertainty $\Delta\tau_{\mathrm{arr}}$ in the arrival time such detectors would be sensitive to GWs whose amplitude is $h \sim \Delta\tau_{\mathrm{arr}}/T_{\mathrm{obs}}$ at frequencies $f \sim T_{\mathrm{obs}}^{-1}$. The highest sensitivity is reached for continuous signals (like a stochastic background) with observation times of one or two years (corresponding to frequencies $10^{-9} - 10^{-8}$ Hz). The most stringent constraint on a stochastic background of gravitational radiation comes from the analysis of the data concerning PSR B1855+09 [29] which gives the following 95% confidence-level bound at $\bar{f} = 4.4 \times 10^{-9}$ Hz

$$h_0^2\, \Omega_{\mathrm{gw}}(\bar{f}) < 1.0 \times 10^{-8}\,. \tag{6}$$

Characteristic Frequency and Amplitude

Beside the observational constraints discussed above, predictions concerning relic GWs are poorly constrained both because of our lack of knowledge about physics beyond the standard model and because of large uncertainties in the understanding of the various production mechanisms concerning primordial GWs. Nevertheless, it is possible to outline some general features describing the typical frequency and amplitude spectrum of cosmological gravitational radiation (see e. g. [11]).

The typical frequency at which a gravitational signal could be detected can be estimated [11] disentangling the dynamical effects associated with the production mechanism from the kinematical contribution depending on the red-shift associated to the cosmic time at which the signal is produced. Assuming that a graviton is produced at a certain temperature T_{P} with a frequency f_{P} during the epoch where the Universe was either radiation or matter dominated, the frequency at which it could be currently detected is $f_0 = f_{\mathrm{P}}/(1 + z_{\mathrm{P}})$. To compute the red-shift z_{P} one exploits the adiabatic expansion of the universe, i. e. the fact that the entropy per comoving volume $S \propto g_{\mathrm{S}}(T)\, a^3(t)\, T^3$ is conserved (here $a(t)$ is the time-dependent scale factor of the Universe and $g_{\mathrm{S}}(T)$ accounts for the effective number of degrees of freedom at temperature T). The result is ([11,30])

$$f_0 \sim 8 \times 10^{-14}\, f_{\mathrm{P}} \left(\frac{100}{g_{\mathrm{S}}(T_{\mathrm{P}})}\right)^{1/3} \left(\frac{1\,\mathrm{GeV}}{T_{\mathrm{P}}}\right) \tag{7}$$

The dynamics of the production mechanism enters into the estimate of the frequency f_{P}. It is natural to express the wavelength λ_{P} corresponding to such frequency in terms of the horizon scale H_{P}^{-1} at the time of production, i. e. $\lambda_{\mathrm{P}} = \xi H_{\mathrm{P}}^{-1}$, with $\xi \leqslant 1$. Assuming that the gravitational radiation is produced during the era where the energy density is dominated by relativistic radiation, $H_{\mathrm{P}}^2 \propto g(T_{\mathrm{P}})\, T_{\mathrm{P}}^4/M_{\mathrm{pl}}^2$, $g(T_{\mathrm{P}}) \sim g_{\mathrm{S}}(T_{\mathrm{P}})$ (here $g(T_{\mathrm{P}})$ is the number of degrees of freedom at temperature T_{P}). This yields [11]

$$f_0 \sim 1.6 \times 10^{-7}\, \frac{1}{\xi} \left(\frac{100}{g(T_{\mathrm{P}})}\right)^{-1/6} \left(\frac{T_{\mathrm{P}}}{1 GeV}\right) Hz\,. \tag{8}$$

Hence, the physics at TeV scale could be probed by space-borne interferometers like LISA, while earth-based interferometers could investigate phenomena whose typical energy scale is in the range 10^8–10^{10} GeV.

As for the characteristic intensity of primordial gravitational radiation, despite a strong dependence upon the production mechanism, it is still possible to draw some general considerations [11]. When primordial gravitational waves are produced at the Planck scale by collisions and decay together with other relativistic particles that are currently observed (i. e. CMB photons), it is reasonable to assume $\rho_{\text{gw}}(t_{\text{P}}) \sim \rho_{\gamma}(t_{\text{P}})$. In this case, taking into account the different red-shift characterizing photons and gravitons (the latter decoupling just below the Planck scale), the present value for the GW spectrum is estimated to be [30]

$$h_0^2\,\Omega_{\text{gw}} \sim 1.7 \times 10^{-5} \left(\frac{100}{g(T_{\text{P}})}\right)^{1/3} \Omega_{\text{gw}}(t_{\text{P}})\,, \tag{9}$$

where $\Omega_{\text{gw}}(t_{\text{P}})$ is the fraction of energy density produced in GWs. Another possible scenario incorporates all possible mechanisms of primordial production of GWs different from the ones giving rise to CMB photons. Also in this case it is possible to relate the GW spectrum to the CMB spectrum. For instance, one may consider the spectrum of gravitational radiation produced by the amplification of vacuum fluctuations during an inflationary epoch; in particular, it is possible to estimate the spectrum amplitude for a frequency corresponding to the highest curvature scale H_{max} reached during the inflationary phase (see e. g. [11]). The result is

$$h_0^2\,\Omega_{\text{gw}} \sim \left(\frac{H_{\text{max}}}{M_{\text{pl}}}\right)^2 h_0^2\,\Omega_{\gamma}\,, \tag{10}$$

where $h_0^2\,\Omega_{\gamma} \sim 10^{-5}$ is the current fraction (in terms of the critical density) of radiation energy density. Slow-roll inflationary models are characterized by a typical Hubble scale during the acceleration era that cannot exceed a value of about $8 \times 10^{-6}\,M_{\text{pl}}$ (to be consistent with the CMB anisotropy measurements) [31], yielding to an almost constant spectrum in the frequency window $[10^{-16}, 10^{9}]$ Hz $h_0^2\,\Omega_{\text{gw}} \sim 10^{-15}$ [32]. While this is too low to be detected both by earth-based interferometers (even in their advanced configurations) and by LISA (because of the presence of a strong astrophysical background) [33], a space detector sensitive in the frequency band $[0.1, 1]$ Hz (which is almost free from astrophysical backgrounds) could reach the desired sensitivity [34]. Pre-big-bang models [35], string-theory-based cosmological models, are characterized by an inflationary phase where the curvature grows reaching a maximum threshold value $H_{\text{max}} \sim 10^{-1}\,M_{\text{pl}}$. Such a feature in particular produces a stochastic background characterized by a steep-blue spectrum [36] ($\Omega_{\text{gw}} \sim f^3$); the COBE bound is easily satisfied, and, for a certain range of the free parameters of the model, the spectrum can reach a maximum value $h_0^2\,\Omega_{\text{gw}} \sim 10^{-7}$ [37] (compatible with the nucleosynthesis bound) in the frequency range $[10^{-4}, 10^{3}]$ Hz. This signal could be in principle detected by Advanced LIGO.

Production Mechanisms of Relic GWs

The main mechanisms that can produce a primordial stochastic background of gravitational radiation can be divided into two broad classes:

- Parametric amplification of metric tensor perturbations occurring during an early epoch of inflationary expansion;
- Causal process (mainly phase transitions) occurring during the primordial stages of the evolution of the Universe.

Parametric Amplification of Vacuum Fluctuations

In a cosmological model characterized by different dynamical epochs, in each of them any quantum field has different normal modes corresponding to different sets of creation and annihilation operators. Those sets are connected by the so-called Bogoliubov transformations [38] that mix positive frequency modes (i. e. annihilation operators) with negative frequency modes (i. e. creation operators). As a consequence, in an expanding universe, an initial vacuum state (with respect to a certain set of annihilation/creation operators), due to the dynamical evolution of the metric, in a different epoch corresponds to a multi-particle state. The physical idea underlying quantum production of particles in an expanding universe can be qualitatively explained as follows [11]. For the sake of simplicity, let's consider a cosmological model with two different phases (A and B), where the transition between them occurs over a time scale $\Delta\tau \sim H^{-1}(\tau)$. For a mode whose physical frequency at the time of transition is f_τ there exist two alternative regimes: abrupt/sudden transition ($2\pi f_\tau \Delta\tau \ll 1$) or smooth, adiabatic transition ($2\pi f_\tau \Delta\tau \gg 1$). For modes characterized by a sudden transition, while the physical state associated to the field in the phase A has not significantly changed, the corresponding particle number must be evaluated with respect to the set of annihilation and creation operators relative to the phase B. Furthermore, due to the mixing between positive frequency and negative frequency modes, the vacuum state relative to the phase A corresponds to a multi-particle state relative to the phase B. On the other hand, for modes for which the transition is adiabatic, no particle production takes place. Applying this analysis to a cosmological scenario with a primordial phase of accelerated expansion driven by a scalar field, vacuum metric fluctuations are amplified thus producing, in the final decelerating radiation/matter dominated phase a final, multi-graviton phase that gives rise to a stochastic background of gravitational radiation. The corresponding spectrum is given by [11]

$$\Omega_{\mathrm{gw}} = \frac{16\pi^4}{\rho_{\mathrm{c}}} n_{\mathrm{f}} f^4 , \tag{11}$$

where n_{f} is the number of gravitons per cell of phase space. As mentioned before the frequency spectrum of such backgrounds extends over a huge range of frequencies, from $f_{\mathrm{min}} \sim 3 \times 10^{-18}$ Hz (corresponding to a wavelength compatible with the size of our observable universe) up to a maximum frequency in the GHz range (this ultraviolet cut-off is fixed by the maximum curvature attained during the inflationary phase). As for the strength of the spectrum, the predictions are sensitive to the mechanism that produces the inflationary stage (see previous section).

Causal Mechanisms: Phase Transitions

During the cosmological evolution of the Universe, phase transitions corresponding to symmetry breaking of particle physics fundamental interactions are expected to take place. As for the generation of relic GWs, the presence and type of topological defects and the order of the transition are of particular relevance.

Cosmic Strings Cosmic strings are topological defects forming as a consequence of the spontaneous breaking of a global $U(1)$ gauge symmetry [39,40]. They are characterized by a large mass-per-unit-length μ (related to the scale at which the symmetry is broken), a closed shape and a typical length-scale that can be larger or smaller than the Hubble radius $r_{\mathrm{H}} = 1/H$. Since their tension is equal to their mass per unit length, they have relativistic oscillations, therefore shrinking in size and emitting gravitational radiation. Assuming that a multi-scale network of such strings has been formed, loops whose radius is smaller than r_{H} oscillate, radiate away GWs, shrink and eventually disappear, replaced by loops broken off from loops whose radius is larger than r_{H}. The typical scale for the GW spectrum is determined by the loop radius: since in the network loops with different sizes are present, the corresponding (almost Gaussian) spectrum is nearly flat over a frequency range $[10^{-8}, 10^{10}]$ Hz. A simple dimensional analysis shows that the spectrum can be estimated as $\Omega_{\mathrm{gw}} \sim \gamma\mu/M_{\mathrm{pl}}^2$, where γ is a dimensionless parameter encoding the radiation efficiency. Numerical simulations of cosmic strings predict $h_0^2\, \Omega_{\mathrm{gw}} \sim 10^{-9} \div 10^{-8}$ for $\mu/M_{\mathrm{pl}}^2 < 10^{-6}$ [41]. Cosmic strings may also produce intense bursts of gravitational radiation [42]: in particular, the strongest bursts are produced at sections of strings characterized by large Lorentz boosts. Though the analysis of [42] suggests that such bursts could be detected even with the current LIGO design sensitivity, a more recent analysis [43] points out that large-scale earth-based interferometers at current sensitivity (such as LIGO) are unlikely to detect such GW radiation when the string network is characterized by a classical distribution of loop sizes. On the other hand, if the size of the string loops is much smaller than its classical value ($l_{\mathrm{classical}} \sim \gamma\mu t/M_{\mathrm{pl}}^2$) and it is determined by gravitational back-reaction, the burst amplitude is enhanced and those signals could be detectable with the Advanced LIGO design sensitivity [43].

Cosmic strings can also be produced in some classes of inflationary models based upon string theory [44]. Those strings, dubbed cosmic superstrings, have two key features distinguishing them from the usual field-theoretic strings: (a) Due to their probabilistic interaction and to the higher dimensionality of the underlying theory, their reconnection probability p is less than 1 ($10^{-3} < p < 1$ [45], while for field theoretic strings $p = 1$); (b) different kinds of strings can be formed. As well as strong bursts of gravitational radiation, cosmic superstrings can also generate a stochastic background of GWs, as a result of the incoherent superposition of cusp bursts from a strings network. Taking into account various effects, Siemens and collaborators [46] have investigated the detectability of such a background considering a wide range of experiments. In particular, they have found that interferometric GW detectors are sensitive to portions of the parameter space $[p, \mu/M_{\mathrm{pl}}]$, complementary to ones constrained by the various experimental bounds.

First-Order Phase Transitions In a first-order phase transition, the universe is locally trapped into a meta-stable, high-temperature unbroken symmetry phase. Such a meta-stable configuration is separated from the stable vacuum phase by a barrier in the (temperature-dependent) potential describing the order parameter of the transition. At the quantum level, a transition from the meta-stable to the stable vacuum state takes place through tunneling. As a consequence, random nucleation of bubbles occurs, where the configuration inside a generic bubble corresponds to the true vacuum, while outside it is represented by the meta-stable phase. When the temperature drops below a critical threshold, bubbles have enough volume to expand (the latent heat released during the transition is converted into kinetic energy for the bubbles); they approach a velocity close to the speed of light, collide and eventually yield to a universe in the unbroken-symmetry phase. The occurrence of such collisions breaks the bubbles spherical symmetry and GWs (as well as other kinds of radiation) are produced ([15,16,47]). The corresponding GW spectrum is described by two parameters, the bubble nucleation rate per unit volume β and the ratio α between the false vacuum energy density and the radiation energy density released at the critical, transition temperature T_{tr}. Since the gravitational radiation is produced by collision of bubbles, the spectrum is strongly peaked at the frequency characteristic of the collision rate ($2\pi f_{\mathrm{P}} \sim \beta$). A detailed calculation yields [47]

$$f_{\mathrm{P}} \sim 5 \times 10^{-8} \left(\frac{\beta}{H_{\mathrm{tr}}}\right) \left(\frac{T_{\mathrm{tr}}}{1\,\mathrm{GeV}}\right) \left(\frac{g_{\mathrm{tr}}}{100}\right)^{1/6} \mathrm{Hz} \qquad (12)$$

where H_{tr} and g_{tr} are the Hubble parameter and the number of degrees of freedom at the transition temperature T_{tr}. For the electroweak phase transition (EWPT) $\beta/H_{\mathrm{tr}} \sim 10^2 \div 10^3$, $T_{\mathrm{tr}} \sim 100\,\mathrm{GeV}$ and the peak frequency is therefore in the range $10^{-4} \div 5 \times 10^{-3}$ Hz. A calculation of the peak of the spectrum gives [47]

$$h_0^2\, \Omega_{\mathrm{gw}}(f_{\mathrm{P}}) \sim 10^{-6} \xi^2 \frac{\alpha^2}{1+\alpha^2} \frac{v_b^3}{0.24+v_b^3} \left(\frac{H_{\mathrm{tr}}}{\beta}\right)^2 \left(\frac{100}{g_{\mathrm{tr}}}\right)^{1/3} \qquad (13)$$

where ξ is the fraction of latent heat converted into bubble-wall kinetic energy and v_b is the Hubble expansion velocity. If the Higgs mass is larger than the W-masses (as required by current phenomenological constraints), nonperturbative analysis based upon lattice simulations establishes that within the Standard Model the EWPT cannot be first-order. In the class of the minimal supersymmetric extensions of the standard model, provided that the Higgs mass lies in the range 110–115 GeV and the right-handed stop (i. e. the supersymmetric partner of the top quark) has a mass between 105 and 165 GeV, a first-order EWPT can take place. In this case the GWs spectrum is too weak. Within the larger class of non-minimal supersymmetric extensions of the standard model the strength of the transition can be enhanced: in particular, considering a class of models where an additional gauge singlet in the Higgs sector (thus also obtaining a model explaining the observed baryon asymmetry), for the GW spectrum peak values of $h_0^2\, \Omega_{\mathrm{gw}}(f_{\mathrm{p}}) \sim 10^{-15} \div 10^{-10}$ can be reached [48], corresponding to peak frequencies $f_{\mathrm{p}} \sim 10\,\mathrm{mHz}$.

During a first-order phase transition, the onset of turbulent anisotropic eddies induced in the background fluid by the rapid expansion and collision of bubbles may also produce a GW background ([47,48,49,50]). Again within the non-minimal supersymmetric extensions of the standard model there are regions of the parameter space [48] where the GW spectrum can reach peak values $h_0^2\, \Omega_{\mathrm{gw}} \sim 10^{-10}$ for frequencies $f_{\mathrm{P}} \sim 10^{-3}$ Hz. Cosmic turbulence with additional magnetic fields (affecting the turbulent energy spectrum) can also generate a GW background at the end of a first-order phase transition. In [51] a scenario is analyzed where a first-order phase transition occurs before neutrino decoupling (i. e. much later than EWPT); the corresponding GW spectrum presents a peak in the frequency window relevant for the LISA experiment.

Finally, phenomenologically interesting GW spectra can be produced in non-minimal inflationary models (the so-called extended inflation models) where the accelerating phase ends with bubble collision [52]: the peak frequency of the spectrum lies in the frequency range of ground-based detectors. A detailed analysis [53] has shown that in two-field models where one field performs the first-order phase transition and the other drives the accelerating phase by slowly rolling towards the minimum of the potential, if the phase transition takes place well before the end of inflation, the spectrum has a peak in the frequency range $[10, 10^3]$ Hz and can be detectable by large-scale, earth-based interferometers.

Future Directions

The search for primordial gravitational waves is rather challenging both experimentally and theoretically. While so little is known about the earliest evolution of the Universe (thus rendering the predictions for a cosmological background of gravitational waves as mere indications), the experimental effort to detect such a signal is rather challenging. Nevertheless, the development of space-based detectors, combined with the possibility that future CMB experiments could detect the tensor contributions of primordial metric perturbations hopefully will provide some concrete chances to detect primordial gravitational waves.

Bibliography

Primary Literature

1. Einstein A (1916) Sitzungsber Preuss Akad Wiss Berlin (Math Phys) 1916:688–696
2. Bondi H (1957) Nature 179:1072; (1960) ibid 186:535
3. Hulse RA, Taylor JH (1975) Astrophys J 95:L51–L53
4. http://www.geo600.uni-hannover.de
5. http://www.ligo.org
6. http://www.tama.mtk.nao.ac.jp
7. http://www.virgo.pi.infn.it
8. http://lisa.jpl.nasa.gov
9. http://science.hq.nasa.gov/universe/science/bang.html
10. Seto N, Kawamura S, Nakamura T (2001) Phys Rev Lett 87:221103
11. Maggiore M (2000) Phys Rep 331:283
12. Grishchuk LP (1975) Sov Phys JEPT 40:409
13. Starobinski AA (1979) JEPT Lett 30:682

14. Rubakov VA, Sazhin M, Veryaskin A (1982) Phys Lett B 115:189; Fabbri R, Pollock M (1983) Phys Lett B 125:445; Abbott LF, Harari DD (1986) Nucl Phys B 264:487; Allen B (1988) Phys Rev D 37:2078; Sahni V (1990) Phys Rev D 42:453
15. Witten E (1984) Phys Rev D 30:272
16. Hogan CJ (1986) Mon Not R Astr Soc 218:629
17. Allen B (1995) Les Houches 1995, Relativistic gravitation and gravitational radiation. In: Proceedings of Astrophysical Sources of Gravitational Radiation, 26 Sep–6 Oct, Les Houches, France, e-Print: gr-qc/9604033, p. 373–417
18. Christensen N (1992) Phys Rev D 46:5250; Flanagan EE (1993) Phys Rev D 48:2389; Allen B, Romano JD (1999) Phys Rev D 59:102001
19. Mandic V, Buonanno A (2006) Phys Rev D 73:063008
20. Abbott B et al (2007) Astrophys J 659:918
21. Armstrong JW, Estabrook FB, Tinto M (2001) Class Quantum Grav 18:4059; Tinto M, Armstrong JW, Estabrook FB (2001) Phys Rev D 63(R):021101
22. Schwartzmann VF (1969) JETP Lett 9:184
23. Copi CJ, Schramm DN, Turner MS (1997) Phys Rev D 55:3389
24. Sachs R, Wolfe A (1967) Astrophys J 147:73
25. Bennett C et al (1996) Astrophys J 464:L1
26. Hinshaw G et al (2007) Astrophys J Suppl 170:288
27. Allen B, Koranda S (1994) Phys Rev D 50:3713
28. Smith T, Pierpaoli E, Kamionkowski M (2006) Phys Rev Lett 97:021301
29. Thorsett S, Dewey R (1996) Phys Rev D 53:3468
30. Kamionkowski M, Kosowsky A, Turner M (1994) Phys Rev D 49:2837
31. Kinney WH, Kolb EW, Melchiorri A, Riotto A (2006) Phys Rev D 74:023502
32. Friedman BC, Cooray A, Melchiorri A (2006) Phys Rev D 74:123509
33. Ungarelli C, Vecchio A (2001) Phys Rev D 63:064030
34. Ungarelli C, Corasaniti PS, Mercer RA, Vecchio A (2005) Class Quantum Grav 22:S955
35. Gasperini M, Veneziano G (1993) Astropart Phys 1:317; (1993) Mod Phys Lett A 8:3701; (1994) Phys Rev D 50:2519; (2003) Phys Rep 373:1
36. Brustein R, Gasperini M, Giovannini M, Veneziano G (1995) Phys Lett B 361:45
37. Buonanno A, Maggiore M, Ungarelli C (1997) Phys Rev D 55:3330
38. Birrell ND (1982) Davis PCW Quantum fields in curved space. Cambridge University Press, Cambridge
39. Vilenkin A (1981) Phys Lett B 107; (1985) Phys Rep 121:263
40. Vilenkin A, Shellard EPS (1994) Cosmic strings and other Topological Defects. Cambridge University Press, Cambridge
41. Caldwell RR, Allen B (1992) Phys Rev D 45:3447; Caldwell RR, Battye RA, Shellard EPS (1996) Phys Rev D 54:7146
42. Damour T, Vilenkin A (2000) Phys Rev Lett 85:3761; Phys Rev D 64:064008 (2001); Phys Rev D 71:063510 (2005)
43. Siemens X et al (2006) Phys Rev D 73:105001
44. Jones N et al (2002) High J Energy Phys 07:051; Sarangi S, Henry SH Tye (2002) Phys Lett B 536:185; Dvali G, Vilenkin A (2004) Cosmol J Astropart Phys 03:010; Jones N et al (2003) Phys Lett B 563:6; Copeland EJ et al (2004) J High Energy Phys 06:013
45. Jackson MG et al (2005) J High Energy Phys 10:013
46. Siemens X, Mandic V, Creighton J (2007) Phys Rev Lett 98:111101
47. Kosowsky A, Turner MS, Watkins R (1992) Phys Rev D 45:4514; Phys Rev Lett 69:2026 (1992); Kosowsky A, Turner MS (1993) Phys Rev D 47:4372; Kamionkowski M, Kosowsky A, Turner MS (1994) Phys Rev D 49:2837
48. Apreda R, Maggiore M, Nicolis A, Riotto A (2001) Class Quant Grav 18:L155; Nucl Phys B 631:342 (2002)
49. Kosowsky A, Mack A, Kahniashvili T (2002) Phys Rev D 66:024030
50. Gogoberidze G, Kahniashvili T, Kosowsky A (2007) Phys Rev D 76:083002
51. Dolgov AD, Grasso D (2002) Phys Rev Lett 88:011301; Dolgov AD, Grasso D, Nicolis A (2002) Phys Rev D 66:103505
52. Turner MS, Wilczek F (1990) Phys Rev Lett 65:3080
53. Baccigalupi C, Amendola L, Fortini P, Occhionero F (1997) Phys Rev D 56:4610

Books and Reviews

Buonanno A (2006) Gravitational Waves. In: Proceedings of Les Houches Summer School-Session 86: Particle Physics and Cosmology: The Fabric of Spacetime. Les Houches, France (to appear). e-Print: arXiv:0709.4682 [gr-qc]

Gasperini M (2007) Elements of String Cosmology. Cambridge University Press, Cambridge

Maggiore M (2007) Gravitational Waves. Vol 1: Theory and Experiments. Oxford University Press, Oxford

Mukhanov V (2005) Physical foundations of cosmology. Cambridge University Press, Cambridge

Schutz BF (1985) A First Course in General Relativity. Cambridge University Press, Cambridge

Cosmic Strings

ANA ACHÚCARRO[1,2], CARLOS J. A. P. MARTINS[3,4,5]

[1] Institute Lorentz of Theoretical Physics, University of Leiden, Leiden, The Netherlands
[2] Department of Theoretical Physics, University of the Basque Country UPV-EHU, Leioa, Spain
[3] Centro de Astrofísica, Universidade do Porto, Porto, Portugal
[4] Centro de Física do Porto, Porto, Portugal
[5] DAMTP, CMS, University of Cambridge, Cambridge, UK

Article Outline

Glossary

Branes, D-branes Brane – after membrane – is the name given generically to the "solitons" (the non-perturbative states) of Superstring theory. The most important class is that of Dirichlet- or D-branes, which are dynamical hypersurfaces in which the fundamental strings can end. The number of spatial dimensions of the brane is sometimes indicated explicitly (e. g. D3-branes, D7-branes). A D1-brane is also known as a D-string.

Cosmic strings Cosmic strings are non-dissipative linear concentrations of energy that form at phase transitions in the early universe at which axial symmetries are broken. They are extremely long and thin, their length could be many times the size of the observable universe while their thickness is usually subatomic.

Cusps Closed loops of string undergo periodic oscillations, with a period related to the size of the loop. The dynamical equations predict that during each oscillation there may be a few points at which the string instantaneously doubles back on itself, dubbed cusps. In the neighborhood of the cusp, the string velocity approaches the speed of light. Such an event generates an intense pulse of gravitational and other forms of radiation, strongly beamed in the direction of motion of the cusp (like the sound produced by the cracking of a whip).

Deficit angle The spacetime metric around a straight static cosmic string looks like Minkowski space in cylindrical coordinates, except that the azimuthal coordinate has a range smaller than the usual 2π. This means that the spacetime is actually conical with a global deficit angle $\alpha = 8\pi G\mu/c^2$ if μ is the mass per unit length of the string. Pictorially, an angular wedge of width α is removed from the plane orthogonal to the string and the remaining edges identified.

Horizon, cosmological Roughly speaking, the size of the observable universe at any given time. It is the maximum distance light could have traveled since the Big Bang, and therefore the maximum distance over which physical phenomena can be causally correlated, since information cannot travel faster than light.

Goto–Nambu action An Action principle – in the sense of Lagrange – that describes the relativistic motion of an idealized, infinitely thin (Goto–Nambu) cosmic string with no internal structure and whose energy per unit length equals its tension. It is a generalization of the Action for a relativistic point particle, being proportional to the invariant area of the surface swept out in spacetime by the string motion. It provides a very good approximation to the dynamics of a relativistic magnetic flux line whenever the effects of the string core and of massive excitations transverse to the string can be ignored.

Inflation, cosmological A near-exponential expansion of the (extremely smooth) early universe in which quantum fluctuations would have been frozen and blown up to cosmological size, providing the initial inhomogeneities that would later grow, under the effect of gravity, into large scale structures such as galaxies and clusters visible in the universe today.

Kibble mechanism The inevitable formation of cosmic strings, if they are stable, in rapid phase transitions. After a symmetry-breaking phase transition the choice of a minimum depends on random fluctuations, and will be different in different regions. Neighboring regions in different ground states will have defects separating them. Below a critical temperature the field fluctuations are not enough to change between ground states and the domains effectively freeze. In a cosmological context, the typical scale of defects at formation must be smaller than the causal horizon.

Scaling The notion that a single lengthscale characterizes the evolution of a string network, so that it always looks the same (in a statistical sense) when relevant quantities are re-scaled by that lengthscale. A particular case relevant for cosmology is linear scaling, where this scale is the horizon size. A structureless (Goto–Nambu) cosmic string network usually evolves in such a scale-invariant manner, with about 40 strings crossing any horizon volume throughout the history of the universe. The long string network is diluted by the creation of small closed loops which oscillate and decay into a stochastic background of radiation, cosmic rays and gravitational waves.

Superstrings, cosmic Superstring theory is a candidate model for a quantum theory of gravity unified with all other interactions. It is based on the idea that the fundamental constituents of nature are one-dimensional "strings" whose vibrational modes produce all known particles and interactions (the prefix "super" refers to a symmetry between bosonic and fermionic excitations, known as supersymmetry). These fundamental superstrings are not to be confused with cosmic strings, but superstrings can also be cosmic in some particular models, in which case they are known as cosmic superstrings.

VOS model Shorthand for Velocity-dependent One-Scale model, a quantitative and physically simple analytic model for the evolution of Goto–Nambu string networks. Comparison with numerical simulations shows that the model provides an accurate description of the large-scale features of the network through two averaged (macroscopic) quantities, a characteristic length scale (or correlation length) and the root-mean square string velocity.

Worldsheet A point particle's position as a function of time traces a line in spacetime known as a worldline. A one-dimensional object such as a string traces a two-dimensional worldsheet.

Definition of the Subject

Cosmic strings are linear concentrations of energy that form whenever phase transitions in the early universe break axial symmetries, as originally shown by Kibble [31]. They are the result of frustrated order in the quantum fields responsible for elementary particles and their interactions. For about two decades, motivation for their study was provided by the possibility that they could be behind the density inhomogeneities that led to the observed large-scale structures in the universe. Precision observations, particularly of the cosmic microwave background radiation, have limited strings to a sub-dominant role in structure formation. Instead, the inhomogeneities appear to be consistent with a period of cosmological inflation, but it turns out that particle-physics models of the early universe that predict a period of inflation very often also predict the generation of cosmic strings at the end of it [27,28].

More recently, interest has been revived with the realization that there may be strong links between field theory cosmic strings and fundamental strings. The latter are the supposed ultimate building blocks of matter, and in their original context of superstring theory were thought to be microscopic. However, in its modern version – sometimes referred to as M-theory – it is possible and perhaps even mandatory to have macroscopic (cosmological-sized) fundamental strings [37,53]. Their behavior is expected to be quite similar to that of field theory cosmic strings, although there are some important differences so they may in principle be observationally distinguishable. Being relics of the phase transitions that produced them, cosmic strings provide us with a unique window into the early universe. If they are stable and survive for a significant amount of time (possibly even up to the present day), they may leave an imprint in many astrophysical and cosmological observables, and provide us with information on fundamental physics and the very early universe that would otherwise be inaccessible to us. On the other hand, gaining a quantitative understanding of their properties, interactions, evolution and consequences represents a significant challenge because of their intrinsic complexity. Their non-linearity is particularly noteworthy, with highly non-trivial feedback mechanisms between large (cosmological) and small (microscopic) scales affecting the network dynamics. Considerable reliance, therefore, must be placed on numerical simulations, which are technically difficult and computationally costly. A complementary approach is the use of analytic or semi-analytic models, usually to describe the large-scale features of the networks.

The basic picture of the cosmological evolution of string networks that has emerged for the simplest, Goto–Nambu, networks is of a scaling solution with about 40 long strings always stretching across each horizon volume plus a population of loops (other string types can lead to a different behavior). It is then possible to estimate their cosmological implications quantitatively. For example, these strings continuously source gravitational perturbations on sub-horizon scales. The one parameter in these models is the energy scale of the phase transition at which the strings are created. The astrophysical consequences of strings stem from the non-trivial gravitational field around a string [58]. Particles in the vicinity of a static straight string feel no gravitational acceleration, because in general relativity tension is a negative source of gravity and, since tension equals energy per unit length, their effects cancel. The space-time around the string is locally, but not globally, flat. In fact the space is conical, with a deficit angle

$$\alpha = 8\pi \frac{G\mu}{c^2} \,, \tag{1}$$

where μ is the mass per unit length; the simple way to picture this is to imagine a plane in which an angular wedge α has been removed and the edges glued together. For cosmologically interesting strings the deficit angle ranges from a few seconds of arc to a few millionths of a second of arc.

Introduction

To understand the cosmological evolution and effects of cosmic strings we start in this section with a quick summary of basic cosmology concepts that will be needed later and a description of the simplest type of cosmic strings in which the main features are already apparent. But, first, a warning about units: from now on we will set the speed of light to unity, $c = 1$, so we measure distances in light-travel time, and masses in units of Energy (and viceversa, using $E = mc^2$). Boltzmann's constant is set to unity, so temperature is measured in units of mass/energy (using $E = K_B T$). Finally, we set Planck's constant to unity, $\hbar = 1$, and measure all lengths and masses in units of Planck's length and mass/energy: $l_P = 1.62 \times 10^{-35}$ m,

$M_\mathrm{P} = 2.18 \times 10^{-8}\,\mathrm{kg} = 1.22 \times 10^{19}\ \mathrm{GeV/c^2}$. In these units, Newton's constant is given by $G = M_\mathrm{P}^{-2}$.

The early universe is very smooth. To a very good approximation it is a homogeneous and isotropic spacetime described by a single variable: the rate of expansion of its three-dimensional spatial sections. In Einstein's general relativity this spacetime is described by the flat Friedmann–Robertson–Walker (FRW) metric

$$ds^2 = dt^2 - a^2(t)[d\vec{x} \cdot d\vec{x}]\,, \tag{2}$$

where $\vec{x}$ are fixed (comoving) spatial coordinates and $a(t)$ is the scale factor that determines the fractional or *Hubble expansion rate*

$$H(t) = \frac{1}{a(t)} \frac{da}{dt}\,. \tag{3}$$

The time coordinate t is known as cosmological time; to analyze cosmic string evolution we will also need a different time parametrization known as conformal time, τ. They are related by $d\tau = dt/a(t)$, leading to the metric

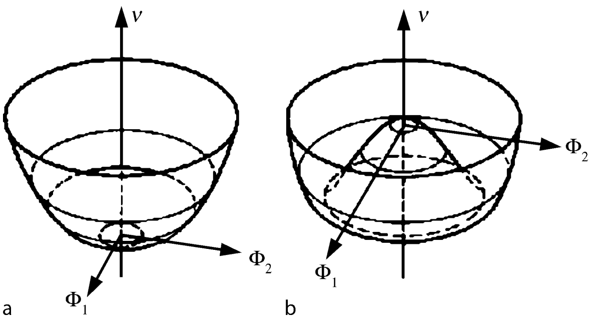
$$ds^2 = a^2(\tau)[d\tau^2 - d\vec{x} \cdot d\vec{x}]\,. \tag{4}$$

The age of the universe is currently estimated to be about 13.7 billion years [55]. The universe starts very hot and dense and is cooled by the expansion, with the temperature decreasing as $T(t) \sim a(t)^{-1}$. The Hubble expansion rate is determined by the energy contents of the universe. In a universe dominated by radiation or very relativistic matter (the hottest, earliest stages), the scale factor evolves as $a(t) \sim t^{1/2}$ and the energy density in radiation as $\rho_\mathrm{radiation} \sim a(t)^{-4} \sim t^{-2}$. The energy density of non-relativistic matter is inversely proportional to volume $\rho_\mathrm{matter} \sim a(t)^{-3}$ and eventually takes over (after about 4000 years), leading to a period of matter domination, during which $a(t) = t^{2/3}$ and therefore $\rho_\mathrm{matter} \sim t^{-2}$. More recently – about five billion years ago – we have entered an epoch of accelerated expansion due possibly to a cosmological constant or some unknown form of *dark energy* whose energy density is constant in time $\rho_\mathrm{dark\ energy} \sim$ const. Dark energy should not be confused with dark matter, an unknown form of matter whose presence we can detect through its gravitational effects but that does not interact with electromagnetic fields and so in particular does not emit light – hence the adjective "dark"–. In the currently accepted cosmological model the energy density in the universe today would be dominated by dark energy (about 74%), followed by about 22% dark matter and only about 4% of regular (baryonic) matter [55]. Dark matter is widely believed to be a particle still to be discovered.

The universe today is far from smooth, but the structure we observe on the scale of clusters of galaxies is consistent with the gravitational collapse of tiny primordial density inhomogeneities $\delta\rho/\rho \sim 10^{-5}$ at the time the cosmic microwave background (CMB) radiation was emitted. The CMB is the oldest radiation we observe, dating back to the time when the universe was only about 380 000 years old. At this epoch the primordial plasma cooled enough to allow the formation of the first atoms (a process known as recombination), and it became transparent to photons (which is referred to as decoupling). Before that moment, the photons behave like a fluid that is strongly coupled to the protons and electrons. An overdense region in the baryon fluid would like to contract but the photon pressure pushes it back, causing both fluids to oscillate. These oscillations are imprinted in the cosmic microwave background and can be detected today in the form of Doppler peaks in its power spectrum.

The spectrum of density inhomogeneities has been accurately measured in the CMB and found to be near scale-invariant and of the right magnitude to produce the structure we observe. The perturbations to the FRW metric can be classified as *scalar* (overall changes to the Newtonian gravitational potential), *vector* (associated with velocity and/or rotational effects) and *tensor* (transverse traceless perturbations to the spatial metric, such as gravitational waves). Each of these affects the CMB in different ways, so their relative contributions can in principle be observationally distinguished. Finally, Thomson scattering of the anisotropic distribution of the CMB photons is particularly important during decoupling and recombination, and induces a partial linear polarization of the scattered radiation, at a level that is around ten percent of the anisotropy. Detection of this polarization signal is at the borderline of the sensitivity of ongoing experiments at the time of writing, but is expected to become standard with forthcoming experiments.

The energy density of a network of cosmic strings in the linear scaling regime is $\rho_\mathrm{strings} \sim t^{-2}$ and therefore it remains a constant fraction of the dominant form of energy during matter or radiation domination. Numerical estimates for the simplest, Goto–Nambu, networks suggest the fraction is around $100G\mu$. Provided the string mass is not close to the Planck scale, this is small enough not to disturb the cosmological evolution; at the same time, for a broad range of values of $G\mu$ this is large enough to be detectable in precision experiments today. Other string types (see Sect. "Field Theory Strings with More Degrees of Freedom") may have larger or smaller fractions or qualitatively different signatures. In particular, networks that do not reach linear scaling may come to dominate



Cosmic Strings, Figure 1
The effective potential energy *V* for a simple string-forming field theory model. The a and b plots correspond to the high and low temperature configurations, respectively. For simplicity the complex field $\mathbf{\Phi}$ has been split into two real scalar fields, $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$

the energy density (which rules them out) or to disappear completely.

The simplest field theory model that produces cosmic strings has a single complex scalar field $\mathbf{\Phi}$ (this is shorthand for a function $\mathbf{\Phi}(t,\vec{x})$ with complex values that do not change under coordinate transformations). Let us assume that the Hamiltonian determining the field dynamics is invariant under an axial symmetry such as a phase rotation, $\mathbf{\Phi} \to \mathbf{\Phi}\mathrm{e}^{i\theta}$. For example, take the potential energy

$$\int \mathrm{d}^3x\, V = \int \mathrm{d}^3x \frac{\lambda}{2}\left(|\mathbf{\Phi}|^2 - \eta^2\right)^2 \tag{5}$$

where λ is a dimensionless coupling constant and η is an energy scale related to the temperature of the symmetry-breaking transition. This has a set of degenerate ground states: the minimum of the potential in field space is the circle $|\mathbf{\Phi}| = \eta$, known as the *vacuum manifold*. Any configuration $\mathbf{\Phi}(t,\vec{x}) = \text{const.} = \eta\mathrm{e}^{i\chi}$ with χ real and constant is a possible ground state or *vacuum*, irrespective of the value of the phase χ.

Figure 1 illustrates what happens. At high temperature the field fluctuations are large enough to make the central peak around $|\mathbf{\Phi}| = 0$ irrelevant, and the effective potential is symmetric and has a minimum there. As the temperature falls the energy will eventually be too low to permit fluctuations over the peak, at which point the field will tend to settle towards one of the ground states. The random choice of minimum in this condensation process then breaks the original axial symmetry. This is the case, for instance, in superfluid ^{4}He.

When a large system goes through a phase transition like this, each part of it has to make this random choice, which need not be the same everywhere. The minimization of gradient terms in the energy of the system tends to make it evolve towards increasingly more uniform configurations, but causality (the principle that no information can travel faster than light) imposes that this evolution can only happen at a limited rate. As a result one expects many domains, each with an uncorrelated choice of ground state. Where these domains meet there is some probability of forming linear defects – cosmic strings – around which the phase angle varies by 2π (or possibly multiples thereof). This is the Kibble mechanism. Notice that the field vanishes at the string's core, so there is trapped potential energy (as well as gradient energy). These strings are known as *global* strings because the axial symmetry that is broken below the phase transition is "global", that is, the transformation $\mathbf{\Phi} \to \mathbf{\Phi}\mathrm{e}^{i\theta}$ is independent of position.

The next step is to consider charged scalar fields interacting with an electromagnetic field. The best known example of a symmetry–breaking transition of this kind is the condensation of Cooper pairs in a superconductor, that has the effect of making photons massive below the critical temperature (in this case the axial symmetry is of the "local" of "gauge" type). The cosmic strings that result are magnetic flux tubes that do not dissipate because the magnetic field is massive outside the string core.

This type of vortex was first discussed by Abrikosov [1] in the context of type II superconductors. Nielsen and Olesen [45] generalized these ideas to the relativistic quantum field theory models used in particle physics, in particular the Abelian Higgs model which is a relativistic version of the Landau–Ginzburg model of superconductivity, governed by the action

$$S = \int \mathrm{d}^4x \left[|\partial_\mu \mathbf{\Phi} - iqA_\mu \mathbf{\Phi}|^2 - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \right. \\ \left. - \frac{\lambda}{2}(|\mathbf{\Phi}|^2 - \eta^2)^2 \right] . \tag{6}$$

A_μ is the gauge field and $\mathbf{\Phi}$ is a complex scalar of charge $q\,(q = 2e$ in superconductors, where $\mathbf{\Phi}$ is the Cooper pair wavefunction). The second term is the usual Maxwell action for the electromagnetic field, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The energy per unit length of a straight, static string lying on the z-axis is

$$E = \int \mathrm{d}^2x \left[|\partial_x \mathbf{\Phi} - iqA_x \mathbf{\Phi}|^2 + |\partial_y \mathbf{\Phi} - iqA_y \mathbf{\Phi}|^2 \right. \\ \left. + \frac{1}{2}B^2 + \frac{\lambda}{2}(|\mathbf{\Phi}|^2 - \eta^2)^2 \right] \tag{7}$$

where $B = \partial_x A_y - \partial_y A_x$ is the z-component of the magnetic field. Finite energy configurations must have $|\mathbf{\Phi}| = \eta$

(the *vacuum manifold* is still a circle) but the phase of $\mathbf{\Phi}$ is undetermined provided the gradient terms and the magnetic field go to zero fast enough. This condition allows for finite energy solutions $A_t = A_r = A_z = 0$, $\mathbf{\Phi}(r,\theta) \sim \eta e^{in\theta}$, $A_\theta(r,\theta) \sim n/(qr)$, as $r \to \infty$, in which the total magnetic flux in the plane perpendicular to the string is quantized,

$$\int \mathrm{d}^2 x B = \oint \vec{A} \cdot \vec{\mathrm{d}} l = \frac{2\pi n}{q}$$

and n is the winding number of the string. If the constants λ and q are such that fluctuations in the scalar field $\mathbf{\Phi}$ and the gauge field A_μ have equal masses, it is possible to show that the string saturates an inequality of the form

$$\text{Energy per unit length} \geq \text{constant x} \mid \text{magnetic flux}|$$

known as the Bogomolnyi bound [7]. In this case, parallel strings at close range exert no force on each other and there are static multivortex solutions [56]. If the mass of the scalar excitations is lower (higher) than that of the gauge excitations, parallel strings will attract (repel).

More complicated particle physics models – in particular those describing the early universe – involve gauge symmetries that generalize the electromagnetic interaction, mediated by photons, to more complicated interactions such as the electroweak or Grand Unified (GUT) interactions. The messenger fields that play the role of the photons may be massless in the early universe and become massive following a symmetry-breaking transition, and cosmic strings carry the magnetic flux of these other massive gauge fields (not the electromagnetic field).

From a cosmological point of view, the gauge field has the important effect of making the gradient terms decay exponentially fast away from the string so the energy per unit length of these strings is finite. Abrikosov–Nielsen–Olesen strings have no long-range interactions, so their evolution is dominated by their tension and is well described in the thin string or Goto–Nambu approximation.

Field continuity implies that a string of this kind cannot simply come to an end: it must form a closed loop or extend to infinity, and it cannot break into segments. For this reason, strings, once formed, are hard to eliminate. In the absence of energy loss mechanisms, the strings would eventually dominate the energy density of the universe. On the other hand, the strings can decay into radiation, they may cross and exchange partners, and they may also cross themselves, forming a closed loop which may shrink and eventually disappear. The outcome of these competing mechanisms is that the network is expected to reach a scale-invariant (or *scaling*) regime, where the network's characteristic length scale is proportional to the size of the horizon. We will discuss string evolution in more detail in Sect. "String Evolution". If a random tangle of strings was formed in the early universe, there would always be some strings longer than the horizon, so a few would remain even today. Because cosmological phase transitions typically happen in the very early universe, cosmic strings contain a lot of trapped energy, and can therefore significantly perturb the matter distribution. To first order there is a single parameter quantifying the effects of strings, its energy per unit length. In the simple relativistic strings, the mass per unit length and the string tension are equal, because of Lorentz invariance under boosts along the direction of the string (but this need not be true for more elaborate models, see Sect. "Field Theory Strings with More Degrees of Freedom"). Cosmic strings are exceedingly thin, but very massive. Typically, for strings produced around the epoch of grand unification, the mass per unit length would be of order $\mu \sim 10^{21}$ kg m^{-1} and their thickness 10^{-24} m. The gravitational effects of strings are effectively governed by the dimensionless parameter $G\mu$, where G is Newton's constant. For GUT-scale strings, this is 10^{-6}, while for electroweak-scale strings it is 10^{-34}.

String Formation

Spontaneous symmetry breaking is a ubiquitous feature of our theories of fundamental particle interactions. Cosmic strings are formed in many symmetry-breaking phase transitions. If the symmetry is broken from a group G down to a subgroup H, the set of degenerate vacuum or ground states is the manifold $M = G/H$, and the topology of this manifold determines the types of defect that can form. In our previous examples M was a circle; in general, strings can form if M contains closed curves that cannot be contracted within M (the technical term is that M is not simply connected, or that its first homotopy group is nontrivial) [31].

The Kibble mechanism described in Sect. "Introduction" relates the initial density of strings to the size ξ of the domains over which the field is correlated,

$$\rho_{\text{string}} \sim C \frac{1}{\xi^2}$$

with C a constant of order one reflecting the probability that a string forms when three or more domains meet. The correlation length cannot grow faster than the speed of light so in the early universe an obvious upper bound on ξ is the size of the horizon at the time of the phase transition. If the dynamics of the phase transition is known, ξ can be estimated more accurately. In a first order phase transition ξ is given by the typical distance between bubble nucleation sites, which depends on the nucleation rate. In second order phase transitions ξ depends on the critical exponents and the rate of cooling through the critical temperature, T_c, as shown by Zurek [65].

Vortex lines or topological strings can therefore appear in a wide range of physical contexts, from cosmic strings in the early universe through disclinations in room-temperature nematic liquid crystals, to magnetic flux tubes in some superconductors and vortex lines in low-temperature superfluid helium. These systems provide us with a range of opportunities to test aspects of the cosmic string formation and evolution scenario experimentally.

The Kibble mechanism in first order transitions was confirmed in experiments on nematic liquid crystals [15,16]. The Kibble–Zurek scenario for second order transitions has been experimentally verified in Superfluid ^{3}He [11,50] and in Josephson Tunneling Junction arrays [41].

The ^{3}He experiments in a rotating cryostat in Helsinki also confirmed the scale invariance of the initial distribution of loops, $n(R) \sim R^{-4}$, where $n(R)\mathrm{d}R$ is the number density of loops with radii between R and $R + \mathrm{d}R$, as predicted by Vachaspati and Vilenkin [61].

More recently, the formation of a defect network following the annihilation of ^{3}He–A / ^{3}He–B boundary layers has been observed [14]. The precise type of defects is still under investigation but this system constitutes an interesting analogue to the formation of strings from the annihilation of branes in brane inflation scenarios.

There are also a few systems where the string density disagrees with the Kibble–Zurek predictions. In ^{4}He, [21] the reasons are understood: the strings are fuzzy and the network does not survive long enough to be detected [36]. In the case of superconducting films the results are somewhat inconclusive [40] and also it is not completely clear what the expected density of flux quanta should be after a temperature quench; an alternative formation mechanism with different vortex clustering properties has been proposed in [24]. In fact the formation of defects in systems with gauge fields is clearly very relevant to cosmology but is still not completely understood (see Kibble [33] for a recent discussion).

String Evolution

The motion of a cosmic string with worldsheet coordinates σ^a and background space-time coordinates x^μ in a metric $g_{\mu\nu}$ is obtainable from a variational principle applied to the Goto–Nambu action [22,43] ($a = 0, 1; \mu = 0, 1, 2, 3$)

$$S = \mu \times \text{Area} = \mu \int \mathrm{d}\tau\, \mathrm{d}\sigma |\det g_{ab}|^2$$
$$= \mu \int \mathrm{d}\tau\, \mathrm{d}\sigma \left| \det \begin{pmatrix} \dot{x}^\rho \dot{x}^\nu g_{\rho\nu} & \dot{x}^\rho x^{\nu\prime} g_{\rho\nu} \\ x^{\rho\prime} \dot{x}^\nu g_{\rho\nu} & x^{\rho\prime} x^{\nu\prime} g_{\rho\nu} \end{pmatrix} \right|^{\frac{1}{2}} \qquad (8)$$

where μ is again the string mass per unit length, and with dots and primes respectively denoting derivatives with respect to the time-like (σ^0) and space-like ($\sigma^1 = \sigma$) coordinates on the worldsheet. g_{ab} is called the induced metric. We are interested in strings in a FRW background spacetime (see Eq. (4)) and can choose worldsheet coordinates that make the induced metric diagonal

$$\sigma^0 = \tau\,, \quad \dot{\mathbf{x}} \cdot \mathbf{x}' = 0\,, \qquad (9)$$

The choice of conformal time coordinate simplifies the microscopic evolution equations, although as we shall see later on physical time is a more natural choice for the macroscopic evolution (see Sect. "Introduction" for definitions of the two time choices). It is also useful to define the coordinate energy per unit σ,

$$\epsilon^2 = \frac{\mathbf{x}'^2}{1 - \dot{\mathbf{x}}^2}\,. \qquad (10)$$

Then the usual variational techniques can be used to show that the microscopic string equations of motion are

$$\ddot{\mathbf{x}} + 2\frac{\dot{a}}{a}\dot{\mathbf{x}}(1 - \dot{\mathbf{x}}^2) = \frac{1}{\epsilon}\left(\frac{\mathbf{x}'}{\epsilon}\right)' \qquad (11)$$

and

$$\dot{\epsilon} + 2\epsilon\frac{\dot{a}}{a}\dot{\mathbf{x}}^2 = 0\,. \qquad (12)$$

For simplicity we are neglecting effects such as cusps and a frictional force due to particle scattering (which for heavy strings is only relevant during a transient period very early in the network's evolution). The first one is just a wave equation with a particular damping term (provided by the expansion of the universe). The damping also has the effect of reducing the coordinate energy per unit σ.

As was mentioned earlier the expansion of the universe stretches the strings, so in the absence of energy loss mechanisms their energy would grow with the scale factor and the string network would eventually become the dominant component of the universe's energy density – which would be in conflict with observational results. Such decay mechanisms do exist (at least for the simplest models), being ultimately due to radiation losses and to the fact that whenever strings interact they will reconnect [42,51]. In particular closed loops may be formed, and these subsequently oscillate and eventually decay. This decay is thought to be mainly into gravitational radiation, but other forms of radiation are also produced very efficiently.

Provided the decay rate is high enough, the network will not have pathological consequences, but will instead reach a linear scaling solution, where the string density is a constant fraction of the background density and on large scales the network looks the same (in a statistical sense) at all times. Scaling is in fact an attractor solution, as has been shown both using numerical simulations and analytic models. Physically, the reason for this is that if one has a high density of strings then the number of string interactions increases and therefore loop production becomes more efficient and the decay rate increases. Conversely if the density is too low then there are few interactions and the decay rate is correspondingly lower. Numerical simulations confirm this broad picture, but also reveal that string evolution is a complex non-linear process, involving non-trivial interactions between various different scales.

There have been thus far two generations of numerical simulations of Goto–Nambu cosmic string networks in expanding universes. The first (Albrecht and Turok [2], Bennett and Bouchet [9], Allen and Shellard [3]) dates from around 1990, at the peak of the interest in cosmic strings as possible seeds for the large-scale structures we observe today. In the last few years, the renewed interest in strings in the context of models with extra dimensions led to a second generation of simulations (Martins and Shellard [39], Ringeval et al. [49], Olum and Vanchurin [46]), which build upon previous knowledge and exploit the dramatic improvements in hardware and software in the intervening decade and a half to achieve a much higher resolution.

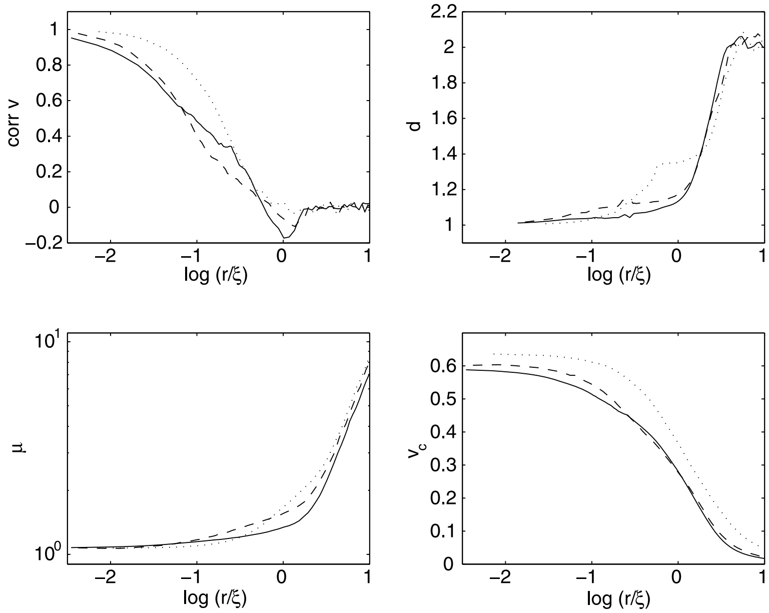
A different approach is provided by full field theory simulations [62]. These are closer to the microphysics of the defects and provide important information on the interactions of the defects and their energy loss mechanisms, but their shorter dynamic range means that they are not optimal for understanding the non-linear feedback mechanisms between widely different scales which affect the dynamics of the network. From this point of view they play a very important role as calibrators, both for Goto–Nambu simulations and for analytic models. One can also carry out Minkowski space simulations (either of Goto–Nambu or field theory type). Neglecting the expansion of the universe is numerically desirable, since such simulations are much easier to implement and evolve much faster. However, the expansion plays a non-trivial role in the network dynamics, so these results should not be naively extrapolated to realistic cosmological scenarios.

Initial conditions for the numerical simulations are usually set up using the Vachaspati–Vilenkin algorithm [61]. One often adds to it random initial velocities, since these tend to enhance the rate of relaxation. All simulations agree on the broad, large-scale features of string networks, and in particular on the fact that the linear scaling solution is an attractor for the evolution. In Goto–Nambu simulations, the initial fraction of the total energy in the form of closed loops is around 20%, but in the linear scaling regime this fraction is around 50% or even slightly more. On the other hand, in field theory simulations this fraction tends to be somewhat smaller.

The first-generation simulations suggested a dynamical picture where the long-string network lost energy to large, long-lived loops, with sizes of order the correlation length. Refinements had each loop self-intersecting into around 10 daughter loops, but loop production from the long strings was essentially monochromatic. The second-generation simulations, however, reveal a quite different picture. Large loops do self-intersect (and indeed the number of daughter loops produced by each one seems to be around 20), but there is also a direct production of large quantities of small loops from slowly moving long-strings with fractal-like substructure. In other words, the loop production is in fact bi-modal. All three second-generation simulations agree on this broad picture, though not on which of the two loop production scales is dominant.

The second-generation simulations present some tentative evidence for the scaling of small-scale features of the network. An open question is whether or not this is expected to happen, given that gravitational backreaction (which would provide a characteristic scale) is not included in any of the network simulations carried out to date. One possible explanation stems from the fact that large loops are not scaled up versions of small loops. Indeed, small loops tend to be nearly circular, whereas large loops are not only far from circular but even far from planar. In other words, the self-intersection probability for a given loop depends on its size, and this may be sufficient to dynamically select a preferred scale. Incidentally, the loop fragmentation processes in these networks highlight the fact that there is a steady flow of energy from large to small scales which is entirely analogous to a Richardson cascade in turbulence. (In this case energy enters via long strings falling inside the horizon, and leaves via radiative decays.)

Figure 2 shows some relevant quantities characterizing Goto–Nambu string networks in the linear scaling regime. These are always plotted as a function of relative scale, to the network's correlation length. The top left panel shows the correlation function for the velocity vectors. A first striking feature is that in the expanding universe cosmic string velocities are anti-correlated on scales around the



Cosmic Strings, Figure 2
Characteristic small-scale properties of cosmic string networks in the linear scaling regime, for matter (*solid lines*), radiation (*dashed*) and Minkowski spacetime (*dotted*) runs. In all plots the horizontal axis represents the logarithm of the physical lengthscale relative to the correlation length of the network. The simulations leading to these results are described in [39]. Top panels show the correlation function for the velocity vectors and the fractal dimension, *bottom panels* show coarse-grained mass per unit length and coherent velocity. The plotted quantities are described in the main text

correlation length (which is smaller than but comparable to the causal horizon), but such a feature is not present in Minkowski space. This anti-correlation is the result of a 'memory' of the network for recent reconnection events, and it is ultimately due to the damping effect of expansion. The top right panel depicts the fractal dimension of the network: this interpolates between $d = 1$ (straight segments) on small scales and $d = 2$ (Brownian network) on large scales, but it does so in a non-trivial way (which is again different depending on whether or not there is expansion) and over a wide range of scales. The fractal dimension evolves with time, decaying on any given physical scale: the strings continually become smoother on any scale, so as to minimize energy. Finally, the bottom panels show the renormalized (or 'coarse-grained') string mass per unit length on the left, and the corresponding coherent velocity on the right panel – notice that the effect of expansion is to reduce the velocities on any given scale.

The technical difficulty and computational cost of numerical simulation provide strong motivation for alternative analytic approaches, which essentially abandon the detailed statistical physics of the string network to concentrate on its thermodynamics. The best example is the velocity-dependent one-scale (VOS) model [38], which builds on previous work by Kibble and Bennett [8,32] and has demonstrated quantitative success when compared with both field theory and Goto–Nambu numerical simulations. The 'one-scale' assumption is that the network has a single characteristic lengthscale, which coincides with the string correlation length and the string curvature radius. This is an approximation which can be tested numerically.

The first assumption in this analysis is to localize the string so that we can treat it as a one-dimensional line-like object. This is clearly a good assumption for gauged strings, such as magnetic flux lines, but may seem more questionable for strings possessing long-range interactions, such as global strings or superfluid vortex lines. However, good agreement between the VOS model and simulations has been found in both cases. The second step is to average the microscopic string equations of motion to derive the key evolution equations for suitable macroscopic quantities, specifically its energy E and RMS velocity v defined by

$$E = \mu a(\tau) \int \epsilon \, d\sigma \, , \quad v^2 = \langle \dot{\mathbf{x}}^2 \rangle = \frac{\int \dot{\mathbf{x}}^2 \epsilon \, d\sigma}{\int \epsilon \, d\sigma} \, . \tag{13}$$

Notice that the energy is an 'extensive' quantity but the RMS velocity is an averaged quantity (and the averaging is weighted by the coordinate energy ϵ). In keeping with the above coordinate choices, the microscopic quantities (on the right-hand side of both equations) are defined in terms of conformal time, but it turns out that the macroscopic evolution that we are now considering is best described in terms of physical time – please refer to the cosmology review for the explicit relation between the two.

Any string network divides fairly neatly into two distinct populations, *viz.* long (or 'infinite') strings and small closed loops. In the following we will focus on the long strings. The long string network is a Brownian random walk on large scales and can be characterized by a correlation length L, which can be used to replace the energy $E = \rho V$ in long strings in our averaged description, that is,

$$\rho \equiv \frac{\mu}{L^2} \,. \tag{14}$$

A phenomenological term must then be included to account for the loss of energy from long strings by the production of loops, which are much smaller than L. A *loop chopping efficiency* parameter $\tilde{c}$ is introduced to characterize this loop production as

$$\left(\frac{\mathrm{d}\rho}{\mathrm{d}t}\right)_{\text{to loops}} = \tilde{c}v\frac{\rho}{L} \,. \tag{15}$$

In this approximation, we would expect the loop parameter $\tilde{c}$ to be a constant; comparison with numerical simulations suggests $\tilde{c} \sim 0.23$.

From the microscopic string equations of motion, one can then average to derive the evolution equation for the correlation length L,

$$2\frac{\mathrm{d}L}{\mathrm{d}t} = 2HL(1+v^2) + \tilde{c}v \,, \tag{16}$$

where H is the Hubble expansion rate defined in Eq. (3). The first term in (16) is due to the stretching of the network by the Hubble expansion which is modulated by the redshifting of the string velocity, while the second is the loop production term. One can also derive an evolution equation for the long string velocity with only a little more than Newton's second law

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \left(1-v^2\right)\left(\frac{k(v)}{L} - 2Hv\right) \,. \tag{17}$$

The first term is the acceleration due to the curvature of the strings and the second is the damping term from the Hubble expansion. Note that strictly speaking it is the curvature radius R which should appear in the denominator of the first term. In the present context we are identifying $R = L$. The function $k(v)$ is the *momentum parameter*, defined by

$$k(v) \equiv \frac{\langle(1-\dot{\mathbf{x}}^2)(\dot{\mathbf{x}}\cdot\mathbf{u})\rangle}{v(1-v^2)} \,, \tag{18}$$

with $\dot{\mathbf{x}}$ the microscopic string velocity and $\mathbf{u}$ a unit vector parallel to the curvature radius vector. For most relativistic regimes relevant to cosmic strings it is sufficient to define it as follows:

$$k_{\mathrm{r}}(v) = \frac{2\sqrt{2}}{\pi}\frac{1-8v^6}{1+8v^6} \,, \tag{19}$$

while in the opposite case ($v \to 0$), we have the non-relativistic limit $k_0 = 2\sqrt{2}/\pi$.

Scale-invariant attractor solutions of the form $L \propto t$ (or $L \propto H^{-1}$) together with $v =$ const., only appear to exist when the scale factor is a power law of the form

$$a(t) \propto t^{\beta} \,, \quad 0 < \beta = \text{const.} < 1 \,. \tag{20}$$

This condition implies that

$$L \propto t \propto H^{-1} \,, \tag{21}$$

with the proportionality factors dependent on the expansion rate β. It is useful to introduce the following parameters to describe the relative correlation length and density, defining them respectively as

$$L = \gamma t \,, \quad \zeta \equiv \gamma^{-2} = \rho t^2/\mu \,. \tag{22}$$

By looking for stable fixed points in the VOS equations, we can express the actual scaling solutions in the following implicit form:

$$\gamma^2 = \frac{k(k+\tilde{c})}{4\beta(1-\beta)} \,, \quad v^2 = \frac{k(1-\beta)}{\beta(k+\tilde{c})} \,, \tag{23}$$

where k is the constant value of $k(v)$ given by solving the second (implicit) equation for the velocity. It is easy to verify numerically that this solution is well-behaved and stable for all realistic parameter values.

If the scale factor is not a power law, then simple scale-invariant solutions like (23) do not exist. Physically this happens because the network dynamics are unable to adapt rapidly enough to the changes in the background cosmology. An example of this is the transition between the radiation and matter-dominated eras. Indeed, since this relaxation to a changing expansion rate is rather slow, realistic cosmic string networks are strictly speaking *never* in scaling during the matter-dominated era. Another example is the onset of dark energy domination around the present day. In this case, the network is gradually slowed down by the accelerated expansion, and asymptotically it becomes frozen in comoving coordinates. The corresponding scaling laws for the correlation length and velocity are $L \propto a$ and $v \propto a^{-1}$.

Despite its success in describing the large-scale features of string networks, the VOS model has the shortcoming of not being able to account for the small-scale features developing on the strings as the network evolves, as clearly shown by numerical simulations. This small-scale structure is in the form of wiggles and kinks, and can be phenomenologically characterized by its fractal properties, as we have sketched above. As a first analytic simplification, the string wiggles can be characterized through a renormalized string mass per unit length that is larger than the bare (Goto–Nambu) mass. This effectively corresponds to considering a model with a non-trivial equation of state (the relation between the string tension and the mass per unit length), which turns out to be one among a larger class of models known as elastic string models. This kind of description has interesting parallels with the coarse-graining approaches that are typical of condensed matter.

A more radical approach is to explicitly abandon the one-scale assumption. This is done in the three-scale model [6], which distinguishes between the characteristic lengthscale (which is simply a measure of the total string energy in a given volume) and the persistence length (which is defined in terms of the invariant length along the string and corresponds to the correlation length or inter-string distance). Additionally there is a third lengthscale which approximately describes a typical scale of the small-scale wiggles. This kind of description is in principle highly flexible, though this can be considered a blessing and a curse. The downside is that one is forced to introduce a large number of (almost free) phenomenological parameters over which one has limited control even when comparing the model with simulations.

Having said that, the three-scale model does confirm, at least qualitatively, the expectations for the behavior of string networks. Scaling of the large scales (in this case the characteristic and persistence lengths) is found to be an attractor, just as in the VOS model. Depending on the behavior of small-scale structures, the two large length scales may reach scaling simultaneously or the former may do so before the latter – a behavior that has been seen in numerical simulations. As for the behavior of the small-scale structures, their evolution timescale is typically slower, and generically they only reach scaling due to the effects of gravitational backreaction (not included in numerical simulations). In the absence of gravitational backreaction, scaling of the small-scale characteristic length is contingent on the removal of a sufficiently large amount of small-scale structure from the long strings by radiation and loop production, which in the model is controlled by a parameter whose detailed behavior is not known.

Finally, an interesting and rather different approach starts out with the assumption that there is a range of scales where stretching due to the expansion is the dominant dynamical effect, even on scales well below the cosmological horizon. A sufficient condition for this is that one is assuming that the rate of string intercommutations is fixed in horizon units. This turns out to be sufficient to allow the construction of a statistical-type description based on two-point correlation functions [48]. Their results are to a first approximation dependent on a critical exponent which physically is related to the coherent string velocity on a given scale. Comparison with numerical simulations shows, as expected, that the best agreement is found around and just below the horizon scale.

A second assumption is that loop production at those scales is sufficiently localized to be describable as a perturbation. When loop production is thus folded into the analysis, the picture that ultimately emerges is of a complicated fragmentation cascade. In particular, this model provides supporting evidence for the two-population loop distribution picture outlined above and clearly seen in high-resolution simulations. There is a population of correlation-length sized loops, produced by direct long-string intercommutation, and a second population with sizes a few orders of magnitude below (quite possibly near the gravitational backreaction scale) and due to loop fragmentation. Whether or not the smoothing provided by gravitational radiation is necessary to yield scaling of the loop sizes is again not entirely clear at the moment, but it is in principle a question for which this formalism could provide an answer.

Astrophysical and Cosmological Consequences

As was mentioned in Sect. "Definition of the Subject", the spacetime around a straight cosmic string is flat. A string lying along the z-direction has an equation of state $p_z = -\rho$, $p_x = p_y = 0$ and therefore there is no source term in the relativistic version of the Poisson equation for the Newtonian gravitational potential

$$\nabla^2 \phi = 4\pi G(\rho + p_x + p_y + p_z) = 0 \,. \qquad (24)$$

A straight string exhibits no analogue of the Newtonian pull of gravity on any surrounding matter. However, this does not mean the string has no gravitational impact at all. On the contrary, we will see that a moving string has dramatic effects on nearby matter or propagating microwave background photons. It is not difficult to derive the spacetime metric about such a straight static string [58]. It has the simple form

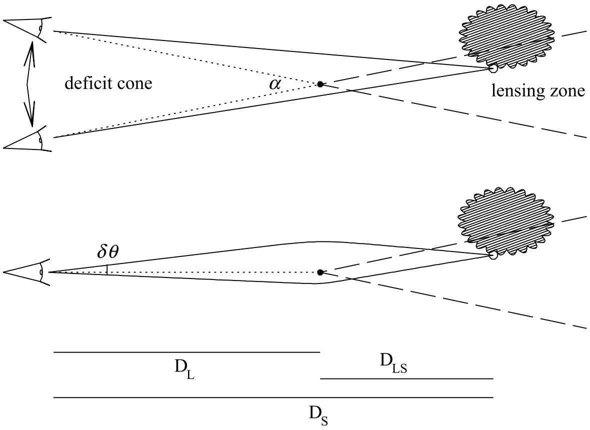
$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}z^2 - \mathrm{d}r^2 - r^2\,\mathrm{d}\theta^2 \,, \tag{25}$$

which looks like Minkowski space in cylindrical coordinates, except for the fact that the azimuthal coordinate θ has a restricted range $0 \le \theta \le 2\pi(1 - 4G\mu)$. That is, the spacetime is actually conical with a global deficit angle

$$\alpha = 8\pi G\mu \,, \tag{26}$$

where an angular wedge of width α is removed and the remaining edges identified.

This deficit angle implies that the string acts as a cylindrical gravitational lens, creating double images of sources behind the string (such as distant galaxies), with a typical angular separation $\delta\theta$ of order α and no distortion [60]. This is illustrated in Fig. 3. A long string would yield a distinctive lensing pattern. We should expect to see an approximately linear array of lensed pairs, each separated in the transverse direction. In each lensing event the two images would be identical and have essentially the same magnitude, i. e. brightness. (Except if we happen to see only part of one of the images.) This is a very unusual signature, because most ordinary gravitational lenses produce an odd number of images of substantially different magnitudes. A number of string lensing event candidates have been discussed in the past, but no confirmed one is currently known.



Cosmic Strings, Figure 3
An illustration of the mechanism behind lensing by cosmic strings. The *thick black dot* represents a cosmic string perpendicular to the page. The spacetime metric around the string can be obtained by removing the angular wedge of width α and identifying the edges. An observer can thus see double images of objects located on a certain zone behind the string. This zone is enclosed by the *dashed lines*, while the *solid lines* depict light rays and the angular separation of the two images, $\delta\theta$, will depend on the distances of the source and the observer to the string as well as on the deficit angle. Reprinted, with permission, from [35]

However, the above simple picture is complicated in practice by the fact that cosmic strings are not generally straight or static. Whenever strings exchange partners kinks are created that straighten out only very slowly, so we expect a lot of small-scale structure on the strings. Viewed from a large scale, the effective tension and energy per unit length will no longer be equal. Since the total length of a wiggly string between two points is greater, it will have a larger effective energy per unit length, U, while the effective tension T, the average longitudinal component of the tension force, is reduced, so $T < \mu < U$. This means that there is a non-zero gravitational acceleration towards the string, proportional to $U - T$. Moreover, the strings acquire large velocities, generally a significant fraction of the speed of light, which introduces further corrections to the deficit angle.

Another effect is the formation of over-dense wakes behind a moving cosmic string [52]. When a string passes between two objects, these are accelerated towards each other to a velocity

$$u_\perp = 4\pi G\mu v \,, \tag{27}$$

where v is the string velocity. Matter therefore collides in a sheet-like structure, leaving a wake behind the moving string. This was the basic mechanism underlying the formation of large-scale structures in cosmic string models. This model has significant attractions, such as the early formation of nonlinear structures, and one can get a good match to the observed galaxy power spectrum in models with a large cosmological constant. However, as we shall discuss, it fails to reproduce the power spectrum of CMB anisotropies observed by COBE, WMAP and other experiments; cosmic strings, therefore, can only play a subdominant role in structure formation (albeit still significant, at the ten to twenty percent level). Cosmic strings create line-like discontinuities in the cosmic microwave background signal [23,34]. For the same reason that wakes form behind a cosmic string, the CMB source on the surface of last scattering is boosted towards the observer, so there is a relative CMB temperature shift across a moving string (a red-shift of the radiation ahead of it, and a blue-shift of that behind), given by

$$\frac{\delta T}{T} \sim 8\pi G\mu v_\perp \,. \tag{28}$$

where $v_\perp$ is the component of the string velocity normal to the plane containing the string and the line of sight. This is known as the Kaiser–Stebbins effect. This simple picture is again complicated in an expanding universe with a wiggly string network and relativistic matter and radiation components. The energy-momentum tensor of the string acts as a source for the metric fluctuations, which in turn create the temperature anisotropies. The problem can be recast using Green's (or transfer) functions which project forward the contributions of strings at early times to today. The actual quantitative solution of this problem entails a sophisticated formalism to solve the Boltzmann equation and then to follow photon propagation along the observer's line of sight. At the time of writing, the most recent comparisons [12] between full-sky maps of cosmic string-induced anisotropies and WMAP data yield a cosmological constraint on the models with

$$G\mu < \text{few} \times 10^{-7} \,, \tag{29}$$

with only a weak dependence on the background cosmology – in particular, on the magnitude of the cosmological constant.

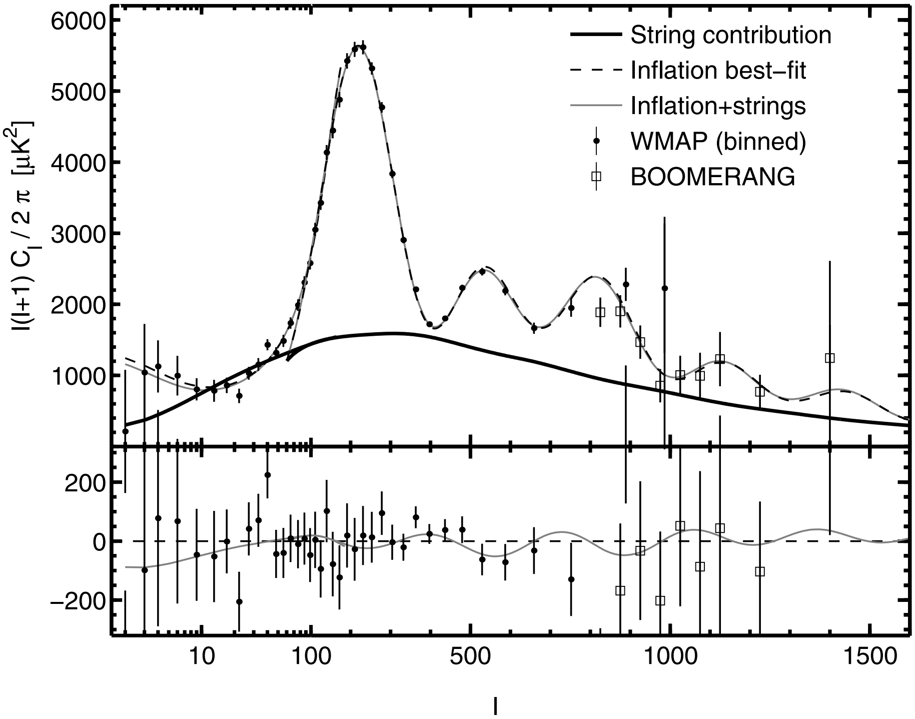
Apart from their scale-invariance (which follows from the network's attractor scaling solution discussed in the previous section), cosmic defect-induced fluctuations appear to be the antithesis of the standard inflation paradigm, because they are causal or active (they are generated inside the horizon and over an extended period), there are also large vector and tensor contributions, and they are distinctly non-Gaussian. All these characteristics leave clear signatures in the cosmic microwave background angular power spectrum, chief of which is a much broader primary Doppler peak and little evidence of secondary oscillations. Unlike inflation, active defect sources act incoherently with extra large-scale power from the vectors and tensors. Moreover, their isocurvature nature provides a partial explanation for why the broad primary peak ends up at larger multipoles (typically $\ell \sim 300$ as opposed to $\ell \sim 200$ in a flat cosmology). Isocurvature perturbations correspond to equal and opposite perturbations in the radiation and matter densities – as opposed to equal fractional perturbations in the number densities of the two components for adiabatic perturbations. On the basis of knowledge from present simulations, therefore, cosmic defects alone are extremely unlikely to have been the seeds for large-scale structure formation. However, they cannot be ruled out entirely. For example, admixtures of inflationary power spectra with significant cosmic defect contributions (at a level around 20%, see Fig. 4) do provide a satisfactory fit to present data. This is interesting among other reasons because it is the sort of level at which the non-Gaussian signatures of cosmic strings should still be discernible, although their distinct line-like discontinuities are only clearly identifiable on small angular scales around a few arc minutes.

Accelerated cosmic strings are sources of gravitational radiation [59]. Consequently, a network of long strings and closed loops produces a stochastic gravitational wave background [25] over a wide range of frequencies and with a spectrum which (at least to a first approximation) has equal power on all logarithmic frequency bins. Another distinctive signal would come from the cusps, the points at which the string instantaneously doubles back on itself, approaching the speed of light. Such an event generates an intense pulse of gravitational and other types of radiation, strongly beamed in the direction of motion of the cusp [18]. If massive cosmic strings do indeed exist, both these pulses and the stochastic background are likely to be among the most prominent signals seen by the gravitational-wave detectors now in operation or planned, in particular LIGO and LISA.

A stringent, though indirect, limit on the string energy per unit length comes from observations of the timing of millisecond pulsars. Gravitational waves between us and a pulsar would distort the intervening space-time, and so cause random fluctuations in the pulsar timing. The fact that pulsar timing is extremely regular places an upper limit on the energy density in gravitational waves, and hence on the string scale. The upper limit [29] is of order $G\mu < 10^{-7}$, though there is still considerable uncertainty because this depends on assumptions about the evolution of small-scale structure.

Although gravitational waves are thought to be the main decay byproduct of the evolution of the simplest cosmic string networks, direct decay into particle radiation is extremely efficient and there are claims that it could be the dominant energy-loss mechanism responsible for scaling [62]. In more complicated models, there are certainly other decay channels. If the strings are global (rather than local), then they will preferentially produce Goldstone bosons instead. In axion models, these Goldstone bosons acquire a small mass and become the axions (a prime dark matter candidate). One can estimate the number density of axions using analytic models of cosmic string evolution such as the VOS model. In another class of string models known as superconducting (since they have additional current-like degrees of freedom), then the decay products can include electromagnetic radiation.

Finally, we should mention the claims that cosmic strings could be responsible for a number of high energy astrophysical and cosmological enigmas, including ultra-high energy cosmic rays, gamma ray bursts, and baryo-



Cosmic Strings, Figure 4
The CMB temperature power spectrum contribution from cosmic strings, normalized to match the WMAP data at $\ell = 10$, as well as the best-fit cases from inflation only (model PL) and inflation plus strings (PL+S). These are compared to the WMAP and BOOMERANG data. The *lower plot* is a repeat but with the best-fit inflation case subtracted, highlighting the deviations between the predictions and the data. Reprinted, with permission, from [13]

genesis (the creation of the matter-antimatter asymmetry of the universe). Since cosmic defects can produce high energy particles, they could contribute to the observed cosmic ray spectrum, notably at ultra-high energies $E \geq 10^{11}$ GeV where the usual acceleration mechanisms seem inadequate [26]. Many ideas have been explored, such as particle emission from cosmic string cusps, but most have been found to produce a particle flux well below current observational limits. However, among the interesting scenarios deserving further study are those with hybrid defects (such as monopoles connected by strings) or vortons.

Field Theory Strings with More Degrees of Freedom

Kibble's original idea was to consider strings in grand unification scenarios, in which the strong and electroweak forces become unified at an energy scale of around 10^{15-16} GeV. Other studies have shown that practically any viable Supersymmetric Grand Unified Theory has a pattern of symmetry breaking transitions that leads to the possibility of cosmic string formation at some point in its history [28]. In particular, in these models the inflationary phase generically ends with a phase transition at which strings are produced. Other studies suggest that this is also a feature of brane inflation models.

So far we have discussed the field theory realization of the simplest model of cosmic string, the Abrikosov–Nielsen–Olesen string, in which the mass per unit length equals the tension and there is no internal structure apart from the magnetic field. In fact, the situation can be much more complex in the early Universe, and realistic particle physics models lead to networks with a much richer phenomenology. The added complexity makes these networks much harder to study, whether by analytic or numerical methods, and consequently they are not as well understood as the simplest case.

We can only give here a brief description of the possible complications. The list below is not complete and furthermore there are strings that fit more than one category. For a more detailed discussion we refer the reader to the reviews by Vilenkin and Shellard, Hindmarsh and Kibble, Carter, and Achúcarro and Vachaspati, where references to the original literature can also be found.

Wiggly Strings; Varying Tension String Networks; Cycloops

The name wiggly strings is sometimes used to refer to any type of string whose mass per unit length is different from its tension. We already mentioned in Sect. "Astrophysical and Cosmological Consequences" that small structure (wiggles) on the string produces a renormalized effective mass per unit length $U > \mu$ and an effective tension $T < \mu$. There are other effects that can affect the mass and tension, for instance the presence of currents along the strings.

A particular kind of small structure is found in extra dimensional models. If spacetime has more than the three spatial dimensions we observe, strings may be able to wrap around the extra dimensions in different ways leading to a renormalized four-dimensional tension and mass per unit length. The effective tension can of course vary along the strings. In extreme cases, the extra dimensional wrapping effects concentrate around certain points along the string which behave like 'beads' (see hybrid networks below) and are called cycloops.

As discussed towards the end of Sect. "String Evolution", the additional degrees of freedom (which can be thought of as a mass current) make the evolution of the networks highly non-trivial. The one-scale assumption is no longer justified: the correlation length, inter-string distance and string curvature radius become distinct lengthscales. Depending on the exact interplay between the bare strings and the mass current wiggles, these lengthscales can evolve differently, and some of them might be scaling while the others are not. The presence of extra dimensions provides a further energy flux mechanism (as energy may be lost into or gained from the extra dimensions) which will affect the string dynamics, but at the time of writing its exact effects have not been studied in detail.

Non-topological/Embedded/Electroweak/ Semilocal Strings

In the Abrikosov–Nielsen–Olesen case, the scalar field is zero at the core of the string, and the symmetry is unbroken there. The zero field is protected by the topological properties of the vacuum manifold (the non-contractible circle) and the string is called topological. In more realistic models, the criterion for topological string production is a non-simply connected vacuum manifold, however complicated. These strings are unbreakable and stable.

On the other hand there are examples in which there is no topological protection but the strings are nevertheless stable. The scalar field configuration at the core can be deformed continuously into a ground state, so these non-topological strings can break, their magnetic flux can spread out, or be converted to a different type of flux. But whether this happens is a dynamical question that depends on the detailed masses and couplings of the particles present, on the temperature, etc.

The best studied examples of non-topological strings look like Abrikosov–Nielsen–Olesen strings 'embedded' in a larger model such as the Glashow–Salam–Weinberg model of electroweak interactions [44,57]. These *electroweak* strings carry magnetic flux of the Z boson, but the strings would only be stable for unphysical values of the Z boson mass. They are closely related to *semilocal* strings, another example of embedded strings where the symmetry breaking involves both local and global symmetries intertwined in a particular way. For low scalar mass these can be remarkably stable.

In general, non-topological strings are not resilient enough for the networks to survive cosmological evolution. If the strings are unstable to spreading their magnetic flux, the network will not form. If the strings are breakable the network may form initially but it will quickly disappear (see hybrid networks below for a concrete example). A remarkable exception to this rule are semilocal strings with very low scalar mass: the network forms as a collection of segments which then grow and reconnect to form longer strings or loops. These evolve like a network of Abrikosov–Nielsen–Olesen strings plus a small population of segments and there is some evidence of scaling [5].

Dressed/Superconducting Strings/Vortons

In realistic particle physics models, a stable string will trap in its core any particles or excitations whose mass is lower inside due to the interactions with the scalar field. These *dressed* strings have a more complicated core structure. In extreme cases, the mass of these trapped particles is zero in the core and they lead to persistent currents along the strings, which are then known as *superconducting* [63]. In some cases, the decay of a loop of superconducting string can be stopped by these currents, leading to long-lived remnants called *vortons* [20] that destroy scaling; typical vortons will either dominate the energy density of the universe (contrary to observations) or contribute to the dark matter if they are sufficiently light.

Hybrid Networks

Hybrid networks contain more than one type of defect, such as for instance strings of different kinds or composite defects combining strings, monopoles and/or domain walls.

Composite Defects The production of strings may be accompanied by the production of other defects such as monopoles or domain walls, before or afterwards, that change the behavior of the network as a whole. These networks can have radically different scaling properties – in particular, linear scaling may not exist at all. Consider for instance a sequence of breakings of the form $G \to H \to K$ in which a symmetry group G first breaks down to a subgroup H which subsequently breaks to an even smaller subgroup K at a lower temperature. Two cases are particularly relevant for strings:

- The first breaking produces stable magnetic monopoles, the second confines – totally or partly – the magnetic field to flux tubes (strings) leading to a network of monopoles connected by strings. This can happen either as string segments, with monopoles at the ends, which eventually contract and disappear or as a network of strings carrying heavy "beads" (the monopoles) which can lead to a scaling solution.
- The first breaking produces stable strings, the second makes domain walls attached to the strings (e. g. in *axion* models). The network is made of pancake-like structures that contract under the wall tension and eventually disappear, although in some cases there may be long-lived remnants.

Non-Abelian/(p,q) Strings Another type of hybrid network contains different types of strings whose intercommutation leads to three-point junctions and bridges. These networks are also very different from the simplest ones but the current consensus is that they also seem to reach a scaling solution during cosmological evolution.

In the Non-abelian case, the magnetic flux carried by the string is not just a number but can have different internal "orientations". These become relevant when the strings cross, limiting the ways in which they can reconnect.

Hybrid networks containing several interacting string types are also found in superstring models (see next section). The most interesting type, usually referred to as (p,q) strings, contains two types of string each carrying different type of flux that is separately conserved: fundamental and solitonic or D-strings, roughly corresponding to electric and magnetic flux tubes. The numbers p and q refer to the units of each kind of flux carried by the strings. Since the mass per unit length depends on these fluxes, (p,q) networks are expected to have a hierarchy of different tensions, as well as junctions and bridges. In fact, junctions and bridges will also form in any model in which parallel strings have an attractive interaction, such as Abrikosov–Nielsen–Olesen strings with extremely low scalar to vector mass ratios.

The existence of string junctions and the hierarchy of string tensions make the evolution of these networks considerably more complicated than that of the simple Goto–Nambu strings. Relatively simple analyses suggest that the heavier strings will gradually decay into the lighter ones, and scaling is eventually reached for the strings at the low end of the spectrum (the heavier ones eventually disappear), although this is still under discussion. Naive expectations that the network might be slowed down to non-relativistic speeds and eventually freeze have so far not been supported by the (admittedly simplistic) numerical simulations performed so far. Further work is needed to understand the general conditions under which scaling is (or is not) an attractor.

Cosmic Superstrings

Superstring theory is to date the only candidate model for a consistent quantum theory of gravity that includes all other known interactions. In string theory, the fundamental constituents of nature are not point-like particles but one-dimensional "strings" whose vibrational modes produce all elementary particles and their interactions. Two important features of the theory are supersymmetry (a symmetry between bosonic and fermionic excitations that keeps quantum effects under control) and the presence of extra dimensions above the four spacetime dimensions that we observe.

It is not yet known how to formulate the theory in its full generality but some weak-coupling regimes are well understood. In these, the fundamental strings live in a 10-dimensional spacetime, of which 6 dimensions are "compactified", resulting in an effective 4-dimensional spacetime we live in. There is another regime, M-theory, in which the fundamental objects are two-dimensional "membranes" and the background spacetime is 11-dimensional. These regimes are related to one another by duality transformations that interchange the role of fluctuation quanta and non-perturbative, soliton-like states (branes), so the expectation is that all regimes are different limits of a unique, underlying theory usually referred to as superstring/M-theory, or just M-theory for short.

Before the discovery of D-branes, the "solitons" of superstring theory, the question of whether fundamental superstrings could ever reach cosmological sizes was analyzed and the possibility discarded [64]. There were two main problems. First, the natural mass per unit length of fundamental strings is close to the Planck scale and would correspond to deficit angles of order 2π, which would have been observed. Second, the strings were inherently unstable to either breaking or – depending on the type of string – becoming the boundary of domain walls that would quickly contract and disappear. The discovery of branes and their role in more exotic compactifications where the six compact dimensions have strong gravitational potentials (and redshifts) has changed this picture. It is now believed that networks of cosmic superstrings could be a natural outcome of brane-antibrane annihilation, especially if the branes are responsible for a period of cosmic inflation [17,30,53].

An important difference with previous scenarios is that these strings are located in regions of the compactified dimensions with very strong gravitational redshift effects ("warping") that reduce the effective mass per unit length of the strings to a level with deficit angles in the region of 10^{-12} to 10^{-7}, compatible with current observations. Another important difference is a much lower probability that the strings intercommute when they cross, estimated to be 10^{-3} to 10^{-1}, depending on the type of strings. The lower intercommutation rates lead to much denser networks. Estimates of the corresponding enhancement in the emission of gravitational radiation by cusps puts these strings in a potentially observable window by future gravitational wave detectors [19,54].

The networks are hybrid, consisting of fundamental strings and D-strings, the latter being either one-dimensional D-branes or perhaps the result of a higher dimensional D-brane where all but one dimension are wrapped around some "holes" (cycles) in the compactified space. There may also be cycloops.

As in the case of hybrid field theory strings, whether or not superstring networks eventually reach a scaling regime is an open question. Analytic studies and numerical simulations of simplified cases suggest that scaling is certainly possible, though contingent on model parameters that at the time of writing are not well understood. In this case, in addition to the presence of junctions and a non-trivial spectrum of string tensions, a third factor can affect the evolution of these networks. If the strings are actually higher-dimensional branes partially wrapped around some extra dimensions, then energy and momentum can in principle leak into or out of these extra dimensions [4]. Since the effective damping force affecting the ordinary and extra dimensions is different, one might generically expect that this will be the case. Depending on its sign and magnitude, such an energy flow can in principle prevent scaling, either by freezing the network (if too much energy leaks out) or by making the strings dominate the universe's energy density (if too much energy leaks in, though this is less likely than the opposite case). In this sense, a somewhat delicate balance may be needed to ensure scaling. At a phenomenological level, further work will be required in order to understand the precise conditions under which each of these scenarios occurs. At a more fundamental level, it is quite likely that which of the scenarios is realized will depend on the underlying compactifications and/or brane inflation models, and that may eventually be used as a discriminating test between string theory realizations.

Future Directions

One of the most exciting prospects is the discovery of magnetic-type CMB polarization (usually referred to as B-modes) as this would reveal the presence of vector and/or tensor modes. Cosmic string models may be further constrained in the near future because B-modes are predicted to have amplitudes comparable to the electric-type E-modes (at large angular scales). At high resolution, one could also hope to observe defects directly through the B-mode signal, against a relatively unperturbed background. Conversely, the detection of vector modes would provide strong evidence against inflation without cosmic defects. Polarization data will also strongly constrain a significant isocurvature contribution to the mainly adiabatic density fluctuations. Isocurvature perturbations can be a signature of more complicated physics during inflation, such as the effects of two or more scalar fields, or the formation of defects at the end of inflation.

Ongoing and future CMB experiments, especially at high resolution, will be probing the degree of Gaussianity of the primordial fluctuations. The detection of significant and unambiguous non-Gaussianity in the primary CMB signal would be inconsistent with simple (so called single field slow-roll) inflation. More general inflationary models can accommodate certain types of non-Gaussianity, and one can also envisage non-Gaussianity from excited initial states for inflation. It is interesting to note that given the existing bounds on $G\mu$, current CMB experiments do not have the sensitivity or resolution to detect cosmic string signatures directly, in particular the Kaiser–Stebbins effect in CMB maps. However, with high-resolution sensitivities becoming available in the near future, direct constraints (or detections) will be possible. This is just one example of the interesting new science that future high-resolution CMB experiments might uncover in the years ahead. In particular, ESA'a Planck Surveyor [47], scheduled for launch in late 2009, may be able to provide significant breakthroughs.

A deeper understanding of the evolution and consequences of string networks will require progress on both numerical simulations and analytic modelings. At the time of writing there is still no numerical code that includes all the relevant physics, even for the simplest (Goto–Nambu) strings. Inclusion of gravitational backreaction is particularly subtle, and may require completely new approaches. The expected improvements in the available hardware and software will allow for simulations with much longer evolution timespan and spatial resolution, which are needed in order to understand the non-linear interactions between large and small scales all the way down to the level of the constituent quantum fields. This in turn will be a valuable input for more detailed analytic modeling, that must accurately describe the non-trivial small-scale properties of the string networks as well as the detailed features of the loop populations. Better modeling is also needed to describe more general networks – three crucial mechanisms for which at present there is only a fairly simplistic description are the presence of junctions, a non-trivial spectrum of string tensions, and the flow of energy-momentum into extra dimensions.

At a more fundamental level, a better understanding of the energy loss mechanisms and their roles in the evolution of the networks is still missing [10] and it will require new developments in the theory of quantum fields out of equilibrium. Such theoretical developments are also needed to understand defect formation in systems with gauge fields, and could be tested experimentally in superconductors.

The early universe is a unique laboratory, where the fundamental building blocks of nature can be probed under the most extreme conditions, that would otherwise be beyond the reach of any human-made laboratory. Cosmic strings are particularly interesting for this endeavor: they are effectively living fossils of earlier cosmological phases, where physical conditions may have been completely different. The serendipitous discovery of cosmic defects or other exotic phenomena in forthcoming cosmological surveys would have profound implications for our understanding of cosmological evolution and of the physical processes that drove it. The search continues while, in the meantime, the absence of cosmic string signatures will remain a powerful theoretical tool to discriminate between fundamental theories. The possibility that something as fundamental as superstring theory may one day be validated in the sky, using tools as mundane as spectroscopy or photometry, is an opportunity than neither astrophysicists nor particle physicists can afford to miss.

Bibliography

Primary Literature

1. Abrikosov AA (1957) On the magnetic properties of superconductors of the second group. J E T P 32:1442–1452
2. Albrecht A, Turok N (1989) Evolution of cosmic string networks. Phys Rev D40:973
3. Allen B, Shellard EPS (1990) Cosmic string evolution: a numerical simulation. Phys Rev Lett 64:119–122
4. Avgoustidis A, Shellard EPS (2005) Cosmic string evolution in higher dimensions. Phys Rev D71:123513
5. Achucarro A, Salmi P, Urrestilla J (2007) Semilocal cosmic string networks. Phys Rev D75:121703
6. Austin D, Copeland EJ, Kibble TWB (1993) Evolution of cosmic string configurations. Phys Rev D48:5594–5627
7. Bogomol'nyi EB (1976) The stability of classical solutions. Yad Fiz 24:861–870
8. Bennett DP (1986) The Evolution Of Cosmic Strings, Phys Rev D33:873 (Erratum-ibid.D34:1235, 1986, Erratum-ibid.D34:3932, 1986)
9. Bennett DP, Bouchet FR (1990) High resolution simulations of cosmic string evolution. Phys Rev D41:2408
10. Borsanyi S, Hindmarsh M (2007) Semiclassical decay of topological defects. e-Print arXiv:0712.0300
11. Bauerle C, Bunkov YM, Fisher SN, Godfrin H, Pickett GR (1996) Laboratory simulation of cosmic string formation in the early universe using superfluid He-3. Nature 382:332–334
12. Bevis N, Hindmarsh M, Kunz M, Urrestilla J (2007) CMB power spectrum contribution from cosmic strings using field-evolution simulations of the Abelian Higgs model. Phys Rev D75:065015
13. Bevis N, Hindmarsh M, Kunz M, Urrestilla J (2008) Fitting CMB data with cosmic strings and inflation. Phys Rev Lett 100:021301
14. Bradley DI, Fisher SN, Guénault AM, Haley RP, Kopu J, Martin H, Pickett GR, Roberts JE, Tsepelin V (2008) Relic topological defects from brane annihilation simulated in superfluid ^{3}He. Nat Phys 4:46
15. Bowick MJ, Chandar L, Schiff EA, Srivastava AM (1994) The cosmological Kibble mechanism in the laboratory – string formation in liquid crystals. Science 263:943–945
16. Chuang I, Durrer R, Turok N, Yurke B (1991) Cosmology in the laboratory – defect dynamics in liquid crystals. Science 251:1336–1367
17. Copeland EJ, Myers RC, Polchinski J (2004) Cosmic F- and D-strings. JHEP 0406:013
18. Damour T, Vilenkin A (2000) Gravitational wave bursts from cosmic strings. Phys Rev Lett 85:3761–3764
19. Damour T, Vilenkin A (2005) Gravitational radiation from cosmic (super)strings: bursts, stochastic background, and observational windows. Phys Rev D71:063510
20. Davis RL, Shellard EPS (1989) Cosmic vortons. Nucl Phys B323:209
21. Dodd ME, Hendry PC, Lawson NS, McClintock PVE, Williams CDH (2004) Non-appearance of vortices in fast mechanical expansions of liquid He-4 through the lambda transition. Phys Rev Lett 81:3703–3706
22. Goto T (1971) Relativistic quantum mechanics of one-dimensional mechanical continuum and subsidiary condition of dual resonance model. Prog Theor Phys 46:1560–1569
23. Gott JR (1985) Gravitational lensing effects of vacuum string: exact results. Ap J 288:422
24. Hindmarsh M, Rajantie A (2000) Defect formation and local gauge invariance. Phys Rev Lett 85:4660
25. Hogan CJ, Rees M (1984) Gravitational interactions of cosmic strings. Nature 311:109

26. Hill CT, Schramm DN, Walker TP (1987) Ultrahigh-energy cosmic rays from superconducting strings. Phys Rev D36:1007
27. Jeannerot R (1997) Inflation in supersymmetric unified theories. Phys Rev D56:6205–6216
28. Jeannerot R, Rocher J, Sakellariadou M (2003) How generic is cosmic string formation in SUSY GUTs. Phys Rev D68: 103514
29. Jenet FA, Hobbs GB, van Straten W, Manchester RN, Bailes m, Verbiest JPW, Edwards RT, Hotan AW, Sarkissian JM, Ord SM (2006) Upper bounds on the low-frequency stochastic gravitational wave background from pulsar timing observations: current limits and future prospects. Astrophys J 653:1571–1576
30. Jones NT, Stoica H, Tye S-H (2003) The production, spectrum and evolution of cosmic strings in brane inflation. Phys Lett B563:6–14
31. Kibble TWB (1976) Topology of cosmic domains and strings. J Phys A9:1387
32. Kibble TWB (1985) Evolution Of A System Of Cosmic Strings, Nucl Phys B 252:227 (Erratum-ibid. B 261:750,1985)
33. Kibble TWB (2007) Phase-transition dynamics in the lab and the universe. Phys Today 60:57
34. Kaiser N, Stebbins A (1984) Microwave anisotropy due to cosmic strings. Nature 310:391
35. Kuijken K, Siemens X, Vachaspati T (2008) Microlensing by cosmic strings. arXiv:0707.2971. preprint, MNRAS 384:161
36. Karra G, Rivers RJ (1998) A reexamination of quenches in ^{4}He. Phys Rev Lett 98:3707
37. Majumdar M, Davis A-C (2002) Cosmological creation of D-branes and anti-D-branes. JHEP 0203:056
38. Martins CJAP, Shellard EPS (2002) Extending the velocity-dependent one-scale string evolution model. Phys Rev D65:043514
39. Martins CJAP, Shellard EPS (2006) Fractal properties and small-scale structure of cosmic string networks. Phys Rev D73:043515
40. Maniv A, Polturak E, Koren G (2003) Phys Rev Lett 91:197001
41. Monaco R, Aaroe M, Mygind J, Rivers RJ, Koshelets VP (2006) Spontaneous fluxon production in annular josephson tunnel junctions in the presence of a magnetic field. Phys Rev B74:144513
42. Moriarty K, Myers E, Rebbi C (1988) Dynamical interactions of cosmic strings and flux vortices in superconductors. Phys Lett B207:411
43. Nambu Y (1970) Lectures at the Copenhagen Summer School. (in press)
44. Nambu Y (1977) String-like configurations in the Weinberg–Salam theory. Nucl Phys B130:505
45. Nielsen HB, Olesen P (1973) Vortex line models for dual strings. Nucl Phys B61:45–61
46. Olum K, Vanchurin V (2007) Cosmic string loops in the expanding universe. Phys Rev D75:063521
47. ESA Planck. http://www.rssd.esa.int/Planck/
48. Polchinski J, Rocha JV (2006) Analytic study of small-scale structure on cosmic strings. Phys Rev D74:083504
49. Ringeval C, Sakellariadou M, Bouchet F (2007) Cosmological evolution of cosmic string loops. JCAP 0702:023
50. Ruutu VMH, Eltsov VB, Gill AJ, Kibble TWB, Krusius M, Makhlin YG, Placais B, Volovik GE, Xu W (1996) Vortex formation in neutron irradiated superfluid He-3 as an analogue of cosmological defect formation. Nature 382:334–336
51. Shellard EPS (1987) Cosmic string interactions. Nucl Phys B283:624
52. Silk J, Vilenkin A (1984) Cosmic strings and galaxy formation. Phys Rev Lett 53:1700
53. Sarangi S, Tye S-H (2002) Cosmic string production towards the end of brane inflation. Phys Lett B536:185–192
54. Siemens X, Mandic V, Creighton J (2007) Gravitational wave stochastic background from cosmic (super)strings. Phys Rev Lett 98:111101
55. Spergel DN, Bean R, Dore O, Nolta MR, Bennet CL, Dunkley J, Hinshaw G, Jarosik N, Komatsu E, Page L, Peiris HV, Verde L, Halpern M, Hill RS, Kogut A, Limon M, Meyer SS, Odegard N, Tucker GS, Wieland JL, Wollack E, Wright EL (2007) Wilkinson microwave anisotropy probe (WMAP) three year results: implications for cosmology. Astrophys J Suppl 170:377
56. Taubes CH (1981) The existence of multi-monopole solutions to the non-abelian, Yang–Mills–Higgs equations for arbitrary simple gauge groups. Comm Math Phys 80:343
57. Vachaspati T (1992) Vortex solutions in the Weinberg–Salam model. Phys Rev Lett 68:1977–1980 (Erratum-ibid. B 26:750,1985)
58. Vilenkin A (1981) Gravitational field of vacuum domain walls and strings. Phys Rev D23:852
59. Vilenkin A (1981) Gravitational radiation from cosmic strings. Phys Lett B107:47
60. Vilenkin A (1984) Cosmic strings as gravitational lenses. Ap J L51:282
61. Vachaspati T, Vilenkin A (1984) Formation and evolution of cosmic strings. Phys Rev D30:2036
62. Vincent G, Antunes ND, Hindmarsh M (1998) Numerical simulations of string networks in the Abelian–Higgs model. Phys Rev Lett 80:2277
63. Witten E (1985) Superconducting strings. Nucl Phys B249:557–592
64. Witten E (1985) Cosmic superstrings. Phys Lett B153:243–246
65. Zurek WH (1985) Cosmological experiments in superfluid helium. Nature 317:505

Books and Reviews

Achúcarro A, Vachaspati T (2000) Semilocal and electroweak strings. Phys Rept 327:347–436

Carter B (1997) Brane dynamics for treatment of cosmic strings and vortons. arXiv:hep-th/9705172

Davis A-C, Kibble TWB (2005) Fundamental cosmic strings. Contemp Phys 46:313–322

Durrer R, Kunz M, Melchiorri A (2002) Cosmic structure formation with topological defects. Phys Rept 364:1–81

Hindmarsh M, Kibble TWB (1995) Cosmic strings. Rept Prog Phys 58:477–562

Kibble TWB (1980) Some implications of a cosmological phase transition. Phys Rept 67:183

Polchinski J (2004) Introduction to cosmic F- and D-strings. In: Lectures given at NATO advanced study institute and EC summer school on string theory: from gauge interactions to cosmology. Cargese, 7–19 Jun 2004. Cargese, pp 229–253

Vilenkin A, Shellard EPS (2000) Cosmic strings and other topological defects. Cambridge University Press, Cambridge

Zurek WH (1996) Cosmological experiments in condensed matter systems. Phys Rep 276:177–221

Cost Sharing

Maurice Koster
University of Amsterdam,
Amsterdam, Netherlands

Article Outline

Glossary

Core The *core* of a cooperative cost game $\langle N, c\rangle$ is the set of all coalitionally stable vectors of cost shares.

Cost function A *cost function* relates to each level of output of a given production technology the minimal necessary units of input to generate it. It is non-decreasing function $c\colon X \to \mathbb{R}_+$, where X is the (ordered) space of outputs.

Cost sharing problem A cost sharing problem is an ordered pair (q, c), where $q \in \mathbb{R}^N_+$ is a profile of individual demands of a fixed and finite group of agents $N = \{1, 2, \dots, n\}$ and c is a cost function.

Game theory The branch of applied mathematics and economics that studies situations where players make decisions in an attempt to maximize their returns. The essential feature is that it provides a formal modeling approach to social situations in which decision makers interact.

Cost sharing rule A *cost sharing rule* is a mapping that assigns to each cost sharing problem under consideration a vector of non-negative cost shares.

Demand game Strategic game where agents place demands for output strategically.

Demand revelation game Strategic game where agents announce their maximal contribution strategically.

Strategic game An ordered triple $G = \langle N, (A_i)_{i\in N}, (\precsim_i)_{i\in N}\rangle$, where

- $N = \{1, 2, \dots, n\}$ is the set of *players*,
- A_i is the set of available *actions* for player i,
- $\precsim_i$ is a *preference relation* over the set of possible consequences C of action.

Definition of the Subject

Throughout we will use a fixed set of agents $N = \{1, 2, \dots, n\}$, where n is a given natural number. For subsets S, T of N, we write $S \subset T$ if each element of S is contained in T; $T\backslash S$ denotes the set of agents in T except those in S. The *power set* of N is the set of all subsets of N; each coalition $S \subset N$ will be identified with the element $\mathbf{1}_S \in \{0, 1\}^N$, the vector with ith coordinate equal to 1 precisely when $i \in S$. Fix a vector $x \in \mathbb{R}^N$ and $S \subset N$. The projection of x on $\mathbb{R}^S$ is denoted by x_S, and $x_{N\backslash S}$ is sometimes more conveniently denoted by x_{-S}. For any $y \in \mathbb{R}^S$, (x_{-S}, y) stands for the vector $z \in \mathbb{R}^N$ such that $z_i = x_i$ if $i \in N\backslash S$ and $z_i = y_i$ if $i \in S$. We denote $x(S) = \sum_{i\in S} x_i$. The vector in $\mathbb{R}^S$ with all coordinates equal to zero is denoted by $\mathbf{0}_S$. Other notation will be introduced when necessary.

This article focuses on different approaches in the literature through a discussion of a couple of basic and illustrative models, each involving a single facility for the production of a finite set M of outputs, commonly shared by a fixed set $N := \{1, 2, \dots, n\}$ of agents. The feasible set of outputs for the technology is identified with a set $X \subset \mathbb{R}^M_+$. It is assumed that the users of the technology may freely dispose over any desired quantity or level of the outputs; each agent i has some demand $x_i \in X$ for output. Each profile of demands $x \in X^N$ is associated to its *cost* $c(x)$, i. e. the minimal amount of the idiosyncratic input commodity needed to fulfill the individual demands. This defines the *cost function* $c\colon X^N \to \mathbb{R}_+$ for the technology, comprising all the production externalities. A *cost sharing problem* is an ordered pair (x, c) of a demand profile x and a cost function c. The interpretation is that x is produced and the resulting cost $c(x)$ has to be shared by the collective N. Numerous practical applications fit this general description of a cost sharing problem.

In mathematical terms a cost sharing problem is equivalent to a *production sharing* problem where output is shared based on the profile of inputs. However, although many concepts are just as meaningful as they are in the cost sharing context, results are not at all easily established using this mathematical duality. In this sense consider [68] as a warning to the reader, showing that the strategic analysis of cost sharing solutions is quite different from surplus sharing solutions. This monograph will center on cost sharing problems. For further reference on production sharing see [51,67,68,91].

Introduction

In many practical situations managers or policy-makers deal with private or public enterprises with multiple users.

A production technology facilitates its users, causing externalities that have to be shared. Applications are numerous, ranging from environmental issues like pollution, fishing grounds, to sharing multipurpose reservoirs, road systems, communication networks, and the Internet. The essence in all these examples is that a manager cannot directly influence the behavior of the users, but only indirectly by addressing the externalities through some decentralization device. By choosing the right instrument the manager may help to shape and control the nature of the resulting individual and aggregate behavior. This is what is usually understood as the mechanism design or implementation paradigm. The state-of-the-art literature shows for a couple of simple but illustrative cost sharing models that one cannot push these principles too far, as there is often a trade-off between the degree of distributive justice and economic efficiency. Then this is what makes choosing 'the' right solution an ambiguous task, certainly without a profound understanding of the basic allocation principles. Now first some examples will be discussed.

Example 1 The water-resource management problem of the Tennessee Valley Authority (TVA) in the 1930s is a classic in the cost-sharing literature. It concerns the construction of a dam in a river to create a reservoir, which can be used for different purposes like flood control, hydro-electric power, irrigation, and municipal supply. Each combination of purposes requires a certain dam height and accompanying construction costs have to be shared by the purposes. Typical for the type of problem is that up to a certain critical height there are economies of scale as marginal costs of extra height are decreasing. Afterwards, marginal costs increase due to technological constraints. The problem here is to allocate the construction costs of a specific dam among the relevant purposes.

Example 2 Another illustrative cost sharing problem dating back from the early days in the cost sharing literature [69,70] deals with landing fee schedules at airports, so-called *airport problems*. These were often established to cover the costs of building and maintaining the runways. The cost of a runway essentially depends on the size of the largest type of airplane that has to be accommodated – a long runway can be used by smaller types as well. Suppose there are m types of airplanes and that c_i is the cost of constructing a landing strip suitable for type i. Moreover, index the types from small to large so that $0 = c_0 < c_1 < c_2 < \cdots < c_m$. In the above terminology the technology can be described by $X = \{0, 1, 2, \ldots, m\}$, and the cost function $c\colon X^N \to \mathbb{R}_+$ is defined by $c(x) = c_k$ where $k = \max\{x_i \mid i \in N\}$ is the maximal service level required in x. Suppose that in a given year, N_k is the set landings of type k airplanes, then the set of users of the runway is $N = \cup_k N_k$. The problem is now to apportion the full cost $c(x)$ of the runway to the users in N, where x is the demand vector given by $x_i = \ell$ if $i \in N_\ell$.

Airport problems describe a wide range of cost sharing problems, ranging from sharing the maintenance cost of a ditch system for irrigation projects [1], to sharing the dredging costs in harbors [14].

Example 3 A joint project involves a number of activities for which the estimated durations and precedence relations are known. Delay in each of these components affects the period in which the project can be realized. Then a cost sharing problem arises when the joint costs due to the accumulated delay are shared among the individuals causing the delays. See [21].

Example 4 In many applications the production technology is given by a network $G = (V, E)$ with nodes V and set of costly edges $E \subset V \times V$, and cost function $c\colon E \to \mathbb{R}_+$. The demands of the agents are now parts of the infrastructure, i. e. subsets of E. Examples include the sharing the cost of infrastructure for supply of energy and water, or transport systems.

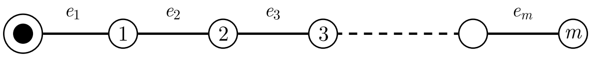
For example, the above airport problem can be modeled as such with

$$\begin{aligned} V &= \{1, 2, \ldots, m\} \cup \{\odot\} \ , \\ E &= \{(\odot, 1), (1, 2), \ldots, (m-1, m)\} \ . \end{aligned}$$

Graphically, the situation is depicted by the line graph in Fig. 1.

Imagine that the runway starts at the special node $\odot$, and that the edges depict the different pieces of runway served to the players. An airplane of type k is situated at node k, and needs all edges towards $\odot$. The edge left to the kth node is called $e_k = (k-1, k)$, and the corresponding cost is $c(e_k) = c_k - c_{k-1}$. The demand of an airplane at node k is now described by the edges on the path from node k to $\odot$.

Example 5 In more general network design problems, a link facilitates a *flow*; for instance, in telecommunication it is data flowing through the network, in road systems it is traffic. [49] discusses a model where a network planner allocates the fixed cost of a network based on the individual



Cost Sharing, Figure 1
Graphical representation of an airport problem

demands being flows. [77,124] discuss *congested* telecommunication networks, where the cost of a link depends on the size of the corresponding flow. Then these positive network externalities lead to a concentration of flow, and thus to *hub-like* networks. Economies of scale require cooperation of the users, and the problem now is to share the cost of these so-called *hub-like* networks.

Example 6 As an insurance against the uncertainty of the future net worths of its constituents, firms are often regulated to hold an amount of riskless investments, i. e. its *risk capital*. Given that returns of normal investments are higher, the difference with the riskless investments is considered as a cost. The sum of the risk capitals of each constituent is usually larger than the risk capital of the firm as a whole, and the allocation problem is to apportion this diversification effect observed in risk measurements of financial portfolios. See Denault [28].

Solving Cost Sharing Problems: Cost Sharing Rules

A *vector of cost shares* for the cost sharing problem (x, c) is an element $y \in \mathbb{R}^N$ with the property that $\sum_{i \in N} y_i = c(x)$. This equality is also called *budget-balancing* condition. Central issue addressed in the cost sharing literature is how to determine the appropriate y. The vast majority of the cost sharing literature is devoted to a mechanistic way of sharing joint costs; given a class of cost sharing problems $\mathcal{P}$, a (simple) formula computes the vector of cost shares for each of its elements. This yields a *cost sharing rule* $\mu \colon \mathcal{P} \to \mathbb{R}^N$ where $\mu(P)$ is the vector of cost shares for each $P \in \mathcal{P}$.

At this point it should be clear to the reader that many formula's will see to a split of joint cost and heading for *the* solution to cost sharing problems is therefore an ambiguous task. The least we want from a solution is that it is consistent with some basic principles of fairness or justice and, moreover, that it creates the right incentives. Clearly, the desirability of solution varies with the context in which it is used, and so will the sense of appropriateness. Moreover, the different parties involved in the decision making process will typically hold different opinions; accountants, economists, production managers, regulators and others all are looking at the same institutional entity from different perspectives. The existing cost sharing literature is about exploring boundaries of what can be thought of desirable features of cost sharing rules. More important than the rules themselves, are the properties that each of them is consistent with. Instead of building a theory on single instances of cost sharing problems, the cost sharing literature discusses structural invariance properties over *classes* of problems. Here the main distinction is made on the basis of the topological properties of the technology, whether the cost sharing problem allows for a *discrete* or *continuous* formulation. For each type of models, divisible or indivisible goods, the state-of-the-art cost sharing literature has developed into two main directions, based on the way individual preferences over combinations of cost shares and (levels of) service are treated. On the one hand, there is a stream of research in which individual preferences are not explicitly modeled and demands are considered *inelastic*. Roughly, it accommodates the large and vast growing *axiomatic* literature (see e. g. [88,129]) and the theory on cooperative cost games [105,130,136,145]. Secondly, there is the literature on cost sharing models where individual preferences are explicitly modeled and demands are elastic. The focus is on non-cooperative *demand games* in which the agents are assumed to choose their demands strategically, see e. g. [56,91,141].

As an interested reader will soon find out, in the literature there is no shortage of plausible cost sharing techniques. Instead of presenting a kind of summary, this article focuses on the most basic and most interesting ones, and in particular their properties with respect to strategic interplay of the agents.

Outline

The article is organized as follows. Section "Cooperative Cost Games" discusses cost sharing problems from the perspective of cooperative game theory. Basic concepts like core, Shapley value, nucleolus and egalitarian solution are treated. Section "Non-cooperative Cost Games" introduces the basic concepts of non-cooperative game theory including dominance relations, preferences, and Nash-equilibrium. Demand games and demand revelation games are introduced for discrete technologies with concave cost function. This part is concluded with two theorems, the strategic characterization of the Shapley value and constrained egalitarian solution as cost sharing solution, respectively. Section "Continuous Cost Sharing Models" introduces the continuous production model and it consists of two parts. First the simple case of a production technology with homogeneous and perfectly divisible private goods is treated. Prevailing cost sharing rules like proportional, serial, and Shapley–Shubik are shortly introduced. We then give a well-known characterization of additive cost sharing rules in terms of corresponding rationing methods, discuss the related cooperative and strategic games. The second part is devoted to the heterogeneous output model and famous solutions like Aumann–Shapley, Shapley–Shubik, and serial rules. We finalize with Sect. "Future Directions" where some future directions of research are spelled out.

Cooperative Cost Games

A discussion of cost sharing solutions and incentives needs a proper framework wherein the incentives are formalized. In the seminal work of von Neumann and Morgenstern [140] the notion of a *cooperative game* was introduced as to model the interaction between actors/players who coordinate their strategies in order to maximize joint profits. Shubik [122] was one of the first to apply this theory in the cost sharing context.

Cooperative Cost Game

A cooperative cost game among players in N is a function $c\colon 2^N \to \mathbb{R}$ with the property that $c(\varnothing) = 0$; for non-empty sets $S \subset N$ the value $c(S)$ is interpreted as the cost that would arise should the individuals in S work together and serve only their own purposes. The class of all cooperative cost games for N will be denoted by CG.

Any general class $\mathcal{P}$ of cost sharing problems can be embedded in CG as follows. For the cost sharing problem $(x, c) \in \mathcal{P}$ among agents in N define the *stand-alone* cost game $c_x \in CG$ by

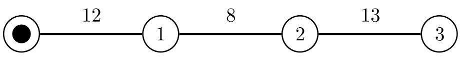
$$c_x(S) := \begin{cases} c(x_S, 0_{N\backslash S}) & \text{if } S \subset N, S \neq \varnothing \\ 0 & \text{if } S = \varnothing \,. \end{cases} \tag{1}$$

So $c_x(S)$ can be interpreted as the cost of serving only the agents in S.

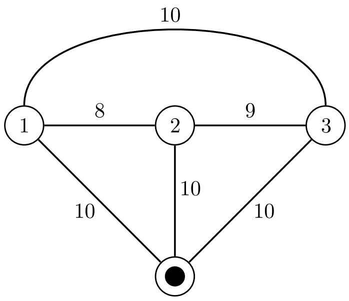
Example 7 The following numerical example will be frequently referred to. An airport is visited by three airplanes in the set $N = \{1, 2, 3\}$, which can be accommodated at cost $c_1 = 12$, $c_2 = 20$, and $c_3 = 33$, respectively. The situation is depicted in Fig. 2.

The corresponding cost game c is determined by associating each coalition S of airplanes to the minimum cost of the runway needed to accommodate each of its members. Then the corresponding cost game c is given by the table below. Slightly abusing notation we denote $c(i)$ to indicate $c(\{i\})$, $c(ij)$ for $c(\{i, j\})$ and so forth.

S	$\varnothing$	1	2	3	12	13	23	123
$c(S)$	0	12	20	33	20	33	33	33



Cost Sharing, Figure 2
Airport game



Cost Sharing, Figure 3
Minimum cost spanning tree problem

Note that, since we identified coalitions of players in N with elements in 2^N, we may write c to denote the cooperative cost game. By the binary nature of the demands the cost function for the technology formally *is* a cooperative cost game. For $x = (1, 0, 1)$ the corresponding cost game c_x is specified by

S	$\varnothing$	1	2	3	12	13	23	123
$c_x(S)$	0	12	0	33	12	33	33	33

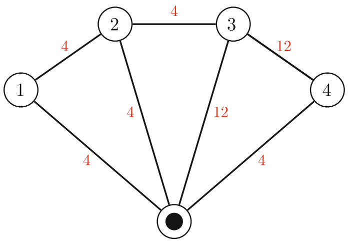
Player 2 is a *dummy player* in this game, for all $S \subset N \backslash \{2\}$ it holds $c_x(S) = c_x(S \cup \{2\})$.

Example 8 Consider the situation as depicted in the Fig. 3 below, where three players, each situated at a different node, want to be connected to the special node $\odot$ using the indicated costly links. In order to connect themselves to $\odot$ a coalition S may use only links with $\odot$ and the direct links between its members, and then only if the costs are paid for. For instance, the minimum cost of connecting player 1 in the left node to $\odot$ is 10, and the cost of connecting players 1 and 2 to $\odot$ are 18 – the cost of the direct link from 2 and the indirect link between 1 and 2. Then the associated cost game is given by

S	$\varnothing$	1	2	3	12	13	23	123
$c(S)$	0	10	10	10	18	20	19	27

Notice that in this case the network technology exhibits positive externalities. The more players want to be connected, the lower the per capita cost.

For those applications where the cost $c(S)$ can be determined irrespective of the actions taken by its complement $N \backslash S$ the interpretation of c implies *sub-additivity*, i. e. the property that for all $S, T \subset N$ with $S \cap T = \varnothing$ implies $c(S \cup T) \leq c(S) + c(T)$. This is for instance an essential feature of the technology underlying natural monopolies (see, e. g., [13,120]). Note that the cost games in Example 7 and 8 are sub-additive. This is a general property for airport games as well as minimum cost spanning tree games.



Cost Sharing, Figure 4
Non-concave MCST game

Sometimes the benefits of cooperation are even stronger. A game is called *concave* (or *sub-modular*) if for all $S, T \subset N$ we have

$$(S \cup T) + c(S \cap T) \leq c(S) + c(T) \,. \tag{2}$$

At first this seems a very abstract property, but one may show that it is equivalent with the following that

$$c(S \cup \{i\}) - c(S) \geq c(T \cup \{i\}) - c(T) \tag{3}$$

for all coalitions $S \subset T \subset N \backslash \{i\}$. This means that the marginal cost of a player i with respect to larger coalitions is non-increasing, i. e. the technology exhibits positive externalities. Concave games are also frequently found in the network literature, see [63,75,93,121].

Example 9 Although sub-additive, minimum cost spanning tree games are not always concave. Consider the following example due to [17]. The numbers next to the edges indicate the corresponding cost. We assume a complete graph and that the invisible edges cost 4. Note that in this game every three-player coalition is connected at cost 12, whereas $c(34) = 16$. Then $c(1234) - c(234) = 16 - 12 = 4$ whereas $c(134) - c(34) = -4$. So the marginal cost of player 1 is not decreasing with respect to larger coalitions.

Incentives in Cooperative Cost Games

The objective in cooperative games is to share the profits or costs savings of cooperation. Similar to the general framework, a *vector of cost shares* for a cost game $c \in CG$ is a vector $x \in \mathbb{R}^N$ such that $x(N) = c(N)$. The question is what cost share vectors make sense if (coalitions of) players have the possibility to opt out thereby destroying cooperation on a larger scale.

In order to ensure that individual players join, a proposed allocation x should at least be *individual rational* so that $x_i \leq c(i)$ for all $i \in N$. In that case no player has a justified claim to reject x as proposal, since going alone yields a higher cost. The set of all such elements is called the *imputation set*. If, in a similar fashion, $x(S) \leq c(S)$ for all $S \subset N$ then x is called *stable*; under proposal x no coalition S has a strong incentive to go alone, as it is not possible to redistribute the cost shares afterwards and make every defector better of. The *core* of a cost game c, notation core(c), consists of all stable vectors of cost shares for c. If cooperation on a voluntary basis by the grand coalition N is conceived as a desirable feature then the core and certainly the imputation set impose reasonable conditions for reaching it. Nevertheless, the core of a game can be empty.

Call a collection $\mathcal{B}$ of coalitions *balanced*, if there is a vector of positive weights $(\lambda_S)_{S \in \mathcal{B}}$ such that for all $i \in N$

$$\sum_{S \in \mathcal{B}, S \ni i} \lambda_S = 1 \,.$$

A cost game c is *balanced* if it holds for each balanced collection $\mathcal{B}$ of coalition that

$$\sum_{S \in \mathcal{B}} \lambda_S c(S) \geq c(N) \,.$$

It is the celebrated theorem below which characterizes all games with non-empty cores.

Theorem 1 (Bondareva–Shapley [20,117]**)** *The cost game c is balanced if and only if the core of c is non-empty.*

Concave cost games are balanced, see [119]. Concavity is not a necessary condition for non-emptiness of the core, since minimum cost spanning tree games are balanced as well.

Example 10 Consider the two-player game c defined by $c(12) = 10, c(1) = 3, c(2) = 8$. Then $\text{core}(c) = \{(x, 10 - x) | 2 \leq x \leq 3\}$. Note that, opposed to the general case, for two-player games sub-additivity is equivalent with non-emptiness of the core.

Cooperative Solutions

A solution on a subclass A of CG is a mapping $\mu\colon A \rightarrow \mathbb{R}^N$ that assigns to each $c \in A$ a vector of cost shares $\mu(c)$; $\mu_i(c)$ stands for the charge to player i.

The Separable Cost Remaining Benefit Solution Common practice among civil engineers to allocate costs of multipurpose reservoirs is the following solution. The *separable cost* for each player (read purpose) $i \in N$ is given by $s_i = c(N) - c(N \backslash \{i\})$, and the *remaining benefit* by $r_i = c(i) - s_i$. The separable cost remaining benefit solution charges each player i for the separable cost s_i and the non-separable costs $c(N) - \sum_{j \in N} s_j$ are then allocated in proportion to the remaining benefits r_i, leading to the formula

$$SCRB_i(c) = s_i + \frac{r_i}{\sum_{j\in N} r_j}\left[c(N) - \sum_{j\in N} s_j\right] . \qquad (4)$$

In this formula it is assumed that c is at least sub-additive to ensure that the r_i's are all positive. For the two-player game c in Example 10 the solution is given by $SCRB(c) = (2 + \frac{1}{2}(10-9), 7 + \frac{1}{2}(10-9)) = (2\frac{1}{2}, 7\frac{1}{2})$. In earlier days the solution was as well known as 'the alternate cost avoided method' or 'alternative justifiable expenditure method'. For references see [144].

Shapley Value One of the most popular and oldest solution concepts in the literature on cooperative games is due to Shapley [116], and named *Shapley-value*. Roughly it measures the average marginal impact of players. Consider an ordering of the players $\sigma\colon N \to N$ so that $\sigma(i)$ indicates the ith player in the order. Let $\sigma^*(i)$ be the set of the first i players according to σ; so $\sigma^*(1) = \{\sigma(1)\}, \sigma^*(2) = \{\sigma(1), \sigma(2)\}$, etc. The *marginal cost share vector* $m^\sigma(c) \in \mathbb{R}^N$ is defined by $m^\sigma_{\sigma(1)}(1) = c(\sigma(1))$ and for $i = 2, 3, \ldots, n$

$$m^\sigma_{\sigma(i)}(c) = c(\sigma^*(i)) - c(\sigma^*(i-1)) . \qquad (5)$$

So according to m^σ each player is charged with the increase in costs when joining the coalition of players before her. Then the *Shapley-value* for c is defined as the average of all $n!$ marginal vectors, i. e.

$$\Phi(c) = \frac{1}{n!}\sum_\sigma m^\sigma(c) . \qquad (6)$$

Example 11 Consider the airport game in Example 7. Then the marginal vectors are given by

σ	(123)	(132)	(213)	(231)	(312)	(321)
$\mu^\sigma(c)$	(12,8,13)	(12,0,21)	(0,20,13)	(0,20,13)	(0,0,33)	(0,0,33)

Hence the Shapley value of the corresponding game is $\Phi(c) = (4, 8, 21)$. Following [69,107], for airport games this allocation is easily interpreted as the allocation according to which each player pays an equal share of the cost of only those parts of the runway she uses. Then $c(e_1)$ is shared by all three players, $c(e_2)$ only by players 2 and 3, and, finally, $c(e_3)$ is paid in full by player 3. This interpretation extends to the class of standard fixed tree games, where instead of the *lattice* structure of the runway, there is a cost of a tree network to be shared, see [63].

If cost game is concave then the Shapley-value is in the core. Since then each marginal vector specifies a core-element, and in particular the Shapley-value as a convex combination of these. Reconsider the minimum cost spanning tree game c in Example 9, a non-concave game with non-empty core and $\Phi(c) = (2\frac{2}{3}, 2\frac{2}{3}, 6\frac{2}{3}, 4)$. Note that this is not a stable cost allocation since the coalition $\{2, 3\}$ would profit by defecting, $c(23) = 8 < 9\frac{1}{3} = \Phi_2(c) + \Phi_3(c)$. [50] show that in general games $\Phi(c) \in \text{core}(c)$ precisely when c is *average concave*. Although not credited as a core-selector, the classic way to defend the Shapley-value is by the following properties.

Symmetry Two players i, j are called *symmetric* in the cost game c if for all coalitions S not containing i, j it holds $c(S \cup \{i\}) = c(S \cup \{j\})$. A solution μ is *symmetric* if symmetric players in a cost game c get the same cost shares. If the cost game does not provide any evidence to distinguish between two players, symmetry is the property endorsing equal cost shares.

Dummy A player i in a cost game c is *dummy* if $c(S \cup \{i\}) = c(S)$ for all coalitions S. A solution μ satisfies *dummy* if $\mu_i(c) = 0$ for all dummy players i in c. So when a player has no impact on costs whatsoever, she can not be held responsible.

Additivity A solution is *additive* if for all cost games c_1, c_2 it holds that

$$\mu(c_1) + \mu(c_2) = \mu(c_1 + c_2) . \qquad (7)$$

For accounting reasons, in multipurpose projects it is a common procedure to subdivide the costs related to the different activities (players) into cost categories, like salaries, maintenance costs, marketing, et cetera. Each category ℓ is associated with a cost game c_ℓ where $c_\ell(S)$ is the total of category ℓ cost made for the different activities in S; then $c(S) = \sum_\ell c_\ell(S)$ is the joint cost for S. Suppose a solution is applied to each of the cost categories separately, then under an additive solution the aggregate cost share of an activity is independent from the particular cross-section in categories.

Theorem 2 (Shapley [116]) *Φ is the unique solution on CG which satisfies all three properties dummy, symmetry, additivity.*

Note that *SCRB* satisfies dummy and symmetry, but that it does not satisfy additivity. The Shapley-value is credited with other virtues, like the following due to [144].

Consider the practical situation that several division managers simultaneously take steps to increase efficiency by decreasing joint costs, but one division manager establishes a greater relative improvement in the sense that its marginal contribution to the cost associated with all possible coalitions increases. Then it is more than reasonable that this division should not be penalized. In a broader context this envisions the idea that each player in the cost game should be credited with the merits of 'uniform' technological advances.

Strong Monotonicity Solution μ is *strongly monotonic* if for any two cost games $c, \bar{c}$ it holds for all $i \in N$ that $c(S \cup \{i\}) - c(S) \geq \bar{c}(S \cup \{i\}) - \bar{c}(S)$ for all $S \subseteq N \setminus \{i\}$ implies $\mu_i(c) \geq \mu_i(\bar{c})$.

Anonymity is the classic property for solutions declaring independence of solution with respect to the name of the actors in the cost sharing problem. See e. g., [3,91,106]. Formally, the definition is as follows. For a given permutation $\pi: N \to N$ and $c \in CG$ define $\pi c \in CG$ by $\pi c(S) = c(\pi(S))$ for all $S \subseteq N$.

Anonymity Solution μ is *anonymous* if for all permutations π of N, and all $i \in N$, $\mu_{\pi(i)}(\pi c) = \mu_i(c)$ for all cost games c.

Theorem 3 (Young [144]) *The Shapley-value is the unique anonymous and strongly monotonic solution.*

[99] introduced the *balanced contributions axiom* for the model of *non-transferable utility games*, or games without side-payments, see [118]. Within the present context of *CG*, a solution μ satisfies the *balanced contributions axiom* if for any cost game c and for any non-empty subset $S \subseteq N$, $\{i, j\} \subseteq S \in N$ it holds that

$$\mu_i(S, c) - \mu_i(S \setminus \{j\}, c) = \mu_j(S, c) - \mu_j(S \setminus \{i\}, c) \,. \quad (8)$$

The underlying idea is the following. Suppose that players agree on using solution μ and that coalition S forms. Then $\mu_i(S, c) - \mu_i(S \setminus \{j\}, c)$ is the amount player i gains or loses when S is already formed and player j resigns. The balanced contributions axiom states that the gains and/or losses by other player's withdrawal from the coalition should be the same.

Theorem 4 (Myerson [99]) *There is a unique solution on CG that satisfies the balanced contributions axiom, and that is Φ.*

The balanced contribution property can be interpreted in a bargaining context as well. In the game c and with solution μ a player i can object against player j to the solution $\mu(c)$ when the cost share for j increases when i steps out of the cooperation, i. e. $\mu_j(N, c) \geq \mu_j(N \setminus \{i\})$. In turn, a *counter objection* by player j to this objection is an assertion that player i would suffer more when j ends cooperation, i. e. $\mu_j(N, c) - \mu_j(N \setminus \{i\}) \leq \mu_i(N, c) - \mu_i(N \setminus \{j\})$. The balanced contribution property is equivalent to the requirement that each objection is balanced by a counter objection. For an excellent overview of ideas developed in this spirit, see [74].

Another *marginalistic* approach is by [44]. Denote for $c \in CG$ the game restricted to the players in $S \subseteq N$ by (S, c). Given a function $P\colon CG \to \mathbb{R}$ which associates a real number $P(N, c)$ to each cost game c with player set N, the *marginal cost* of a player i is defined to be $D^i P(c) = P(N, c) - P(N \setminus \{i\}, c)$. Such a function P with $P(\varnothing, c) = 0$ is called *potential* if $\sum_{i \in N} D^i P(N, c) = c(N)$.

Theorem 5 (Hart & Mas-Colell [44]) *There exists a unique potential function P, and for every $c \in CG$ the resulting payoff vector $DP(N, c)$ coincides with $\Phi(c)$.*

Egalitarian Solution The Shapley-value is one of the first solution concepts proposed within the framework of cooperative cost games, but not the most trivial. This would be to neglect all asymmetries between the players and split total costs equally between them. But as one can expect egalitarianism in this pure form will not lead to a stable allocation. Just consider the two-player game in Example 10 where pure egalitarianism would dictate the allocation $(5, 5)$, which violates individual rationality for player 1.

In order to avoid these problems of course we can propose to look for the most egalitarian allocation *within* the core (see [7,31]). Then in this line of thinking what is needed in Example 10 is a minimal transfer of cost 2 to player 2, leading to the final allocation $(3, 7)$ – the *constrained egalitarian solution*. Although in the former example is was clear what allocation to chose, in general we need a tool to evaluate allocations for the degree of egalitarianism. The earlier mentioned papers all suggest the use of Lorenz-order (see, e. g. [8]). More precisely, consider two vectors of cost shares x and x' such that $x(N) = x'(N)$. Assume that these vectors are ordered in decreasing order so that $x_1 \geq x_2 \geq \cdots \geq x_n$ and $x'_1 \geq x'_2 \geq \cdots \geq x'_n$. Then x *Lorenz-dominates* x' – read x is more egalitarian than x' – if for all $k = 1, \ldots, n-1$ it holds that

$$\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} x'_i \,, \quad (9)$$

with at least one strict inequality. That is, x is better for those paying the most.

Example 12 Consider the three allocations of cost 15 among three players $x = (6, 5, 4)$, $x' = (6, 6, 3)$, and $x'' = (7, 4, 4)$. Firstly, x Lorenz-dominates x'' since $x_1 =$

$6 < 7 = x''_1$, and $x_1 + x_2 = x'_1 + x'_2$. Secondly, x Lorenz-dominates x' since $x_1 = x'_1, x_1 + x_2 < x'_1 + x'_2$. Notice, however, that on the basis of only Eq. (9) we can not make any judgment what is the more egalitarian of the allocations x' and x''. Since $x'_1 = 6 < 7 = x''_1$ but $x'_1 + x'_2 = 6 + 6 > 7 + 4 = x''_1 + x''_2$. The Lorenz-order is only a partial order.

The constrained egalitarian solution is the set of Lorenz-undominated allocations in the core of a game. Due to the partial nature of the Lorenz-order there may be more than one Lorenz-undominated elements in the core. And what if the core is empty? The constrained egalitarian solution is obviously not a straight-forward solution. The original idea of constrained egalitarianism as in [30], focuses on the Lorenz-core instead of the core. It is shown that there is at most one such allocation, that may exist even when the core of the underlying game is empty.

For concave cost games c, the allocation is well defined and denoted by $\mu^{\mathrm{E}}(c)$. In particular this holds for airport games. Intriguingly, empirical studies [1,2] show there is a tradition in using the solution for this type of problems.

For concave cost games c, there exists an algorithm to compute $\mu^{\mathrm{E}}(c)$. This method, due to [30], performs the following consecutive steps. First determine the maximal set S_1 of players minimizing the per capita cost $c(S)/|S|$, where $|S|$ is the size of the coalition S. Then each of these players in S_1 pays $c(S_1)/|S_1|$. In the next step determine the maximal set S_2 of players in $N \backslash S_1$ minimizing $c_2(S)/|S|$, where c_2 is the cost game defined by $c_2(S) = c(S_1 \cup S) - c(S_1)$. The players in S_2 pay $c_2(S_2)/|S_2|$ each. Continue in this way just as long as not everybody is allocated a cost share. Then in at most n steps this procedure results in an allocation of total cost, the constrained egalitarian solution. In short the algorithm is as follows

- **Stage 0:** Initialization, put $S^*_0 = \varnothing, x^* = 0_N$, go to stage $t = 1$.
- **Stage *t*:** Determine

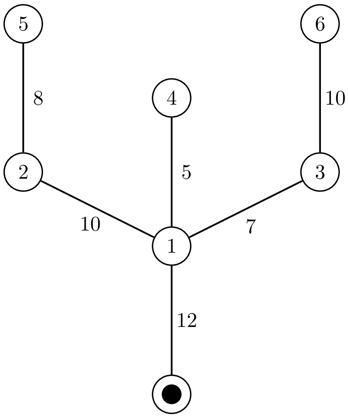
$$S_t \in \arg\max_{S \neq \varnothing} \frac{c(S \cup S^*_{t-1}) - c(S^*_{t-1})}{|S|} .$$

Put $S^*_t = S^*_{t-1} \cup S_t$ and for $i \in S_t$,

$$x^*_i := \frac{c(S^*_t) - c(S^*_{t-1})}{|S_t|} .$$

If $S^*_t = N$, we are finished, put $\mu^{\mathrm{E}}(c) = x^*$. Else repeat the stage with $t := t + 1$.

For example, this algorithm can be used to calculate the constrained egalitarian solution for the airport game in Example 7. In the first step we determine $S_1 = \{1, 2\}$, together with cost shares 10 for the players 1 and 2. Player 3 is allocated the remaining cost in the next step; hence the corresponding final allocation is $\mu^{\mathrm{E}}(c) = (10, 10, 13)$.



Cost Sharing, Figure 5
Standard fixed tree

Example 13 Consider the case where six players share the cost of the following *tree* network that connects them to ⊙. The *standard fixed tree game c* for this network associates to each coalition of players the minimum cost of connecting cost of connecting each member to ⊙, where it may use all the links of the tree. This type of games is known to be concave we can use the above algorithm to calculate $\mu^{\mathrm{E}}(c)$. In the first step we determine $S_1 = \{1, 3, 4\}$ and each herein pays 8. Then in the second step the game remains where the edges e_1, e_3, e_4 connecting S_1 have been paid for. Then it easily follows that $S_2 = \{2, 5\}$, so that players 2 and 5 pay 9 each, leaving 10 as cost share for player 6. Thus, we find $\mu^{\mathrm{E}}(c) = (8, 9, 8, 8, 9, 10)$.

Nucleolus Given a cost game $c \in CG$ the *excess* of a coalition $S \subset N$ with respect to a vector $x \in \mathbb{R}^N$ is defined as $e(S, x) = x(S) - c(S)$; it measures dissatisfaction of S under proposal x. Arrange the excesses of all coalitions $S \neq N, \varnothing$ in decreasing order and call the resulting vector $\vartheta(x) \in \mathbb{R}^{2^n-2}$. A vector of cost shares x will be preferred to a vector $\mathbf{y}$, notation $x \succ y$, whenever $\vartheta(x)$ is smaller than $\vartheta(y)$ in the lexicographic order, i. e. there exists i^* such that for $i \leq i^* - 1$ it holds $\vartheta_i(x) = \vartheta_i(y)$ and $\vartheta_{i^*}(x) < \vartheta_{i^*}(y)$. Schmeidler [115] showed that in the set of individual rational cost sharing vectors there is a unique element that is maximal with respect to $\succ$, which is called the *nucleolus*. This allocation, denoted by $\nu(c)$, is based on the idea of egalitarianism that the largest complaints of coalitions should consistently be minimized. The concept gained much popularity as a core-selector, i. e. it is a one-point solution contained in the core when it is nonempty. This contrasts with the constrained egalitarian solution which might not be well defined, and the Shapley-value which may lay outside the core.

Example 14 Consider in Example 7 the excesses of the different coalitions with respect to the constrained egalitarian solution $\mu^{\mathrm{E}}(c) = (10, 10, 13)$ and the nucleolus $\nu(c) = (6, 7, 20)$:

S	1	2	3	12	13	23
$e(S, \mu^{\mathrm{E}}(c))$	-2	10	-20	0	-10	-10
$e(S, \nu(c))$	-6	-13	-13	-7	-7	-6

Then the ordered excess vectors are

$$\vartheta(10, 10, 13) = (10, 0, -2, -10, -10, -20)\,,$$
$$\vartheta(6, 7, 20) = (-6, -6, -7, -7, -13, -13)\,.$$

Note that indeed $\vartheta(\nu(c)) \succ \vartheta(\mu^{\mathrm{E}}(c))$ since

$$\vartheta_1(6, 7, 20) = -6 < 10 = \vartheta_1(10, 10, 13)\,.$$

The nucleolus of standard fixed tree games may be calculated as a particular *home-down* allocation, as was pointed out by Maschler et al. [75].

For standard fixed tree games and minimum cost spanning tree games the special structure of the technology makes it possible to calculate the nucleolus in polynomial time, i. e. with a number of calculations bounded by a multiple of n^2 (see [39]). Sometimes one may even express the nucleolus through a nice formula; Legros [66] showed a class of cost sharing problems for which the nucleolus equals the SCRB solution. But in general calculations are hard and involve solving a linear program with a number of inequalities which is exponential in n. [124] suggests to use the nucleolus on the cost game corresponding to hub-games.

Instead of the direct comparison of excesses like above, the literature also discusses weighted excesses as to model the asymmetries of justifiable complaints within coalitions. For instance the *per capita nucleolus* minimizes maximal excesses which are divided by the number of players in the coalition (see [105]).

Cost Sharing Rules Induced by Solutions

Most of the above numerical examples deal with cost sharing problems which have a natural and intuitive representation as a cost game. Then basically on such domains of cost sharing problems there is no difference between cost sharing rules and solutions. It may seem that the cooperative solutions are restricted to this kind of situations. But recall that each cost sharing problem (x, c) is associated its stand-alone cost game $c_x \in CG$, as in Eq. (1). Now let μ be a solution on a subclass of $A \subset CG$ and B a class of cost sharing problems (x, c) for which $c_x \in A$. Then a cost sharing rule $\overline{\mu}$ is defined on B through

$$\begin{tikzcd}
(x,\, c) \arrow[r, "f"] \arrow[dr, dashed, "\overline{\mu}"'] & c_x \arrow[d, "\mu"] \\
& \mu(c_x)
\end{tikzcd}$$

Cost Sharing, Figure 6
Induced cost sharing rules

$$\overline{\mu}(x, c) = \mu(c_x)\,. \tag{10}$$

The general idea is illustrated in the diagram on the left. For example, since the Shapley value is defined on the class of all cost games, it defines a cost sharing rule $\overline{\Phi}$ on the class of all cost sharing problems. The cost sharing rule $\overline{\mu}^{\mathrm{E}}$ is defined on the general class of cost sharing problems with corresponding concave cost game. Cost sharing rules derived in this way, *game-theoretical rules* according to [130], will be most useful below.

Non-cooperative Cost Games

Formulating the cost sharing problem through a cooperative cost game assumes inelastic demands of the players. It might well be that for some player the private merits of service do not outweigh the cost share that is calculated by the planner. She will try to block the payment when no service at no cost is a preferred outcome. Another aspect is that the technology may operate at a sub-optimal level if benefits of delivered services are not taken into account. Below the focus is on a broader framework with elastic demands, which incorporates preferences of a player are defined over *combinations* of service levels and cost shares. The theory of non-cooperative games will provide a proper framework in which we can discuss individual aspirations and efficiency of outcomes on a larger scale.

Strategic Demand Games

At the heart of this non-cooperative theory is the notion of a *strategic game*, which models an interactive decision-making process among a group of *players* whose decisions may impact the consequences for others. Simultaneously, each player i independently chooses some available action

a_i and the so realized *action profile* $a = (a_1, a_2, \dots, a_n)$ is associated with some consequence $f(a)$. Below we will have in mind demands or offered contributions as actions, and consequences are combinations of service levels with cost shares.

Preferences Over Consequences Denote by A the set of possible action profiles and C as the set of all consequences of action. Throughout we will assume that players have preferences over the different consequences of action. Moreover, such preference relation can be expressed by a *utility function* $u_i : C \to \mathbb{R}$ such that for $z, z' \in C$ it holds $u_i(z) \le u_i(z')$ if agent i weakly prefers z' to z. Below the set of consequences for agent $i \in N$ will consist of pairs (x, y) where x is the level of service and y a cost share, so that a utilities are specified through multi-variable functions, $(x, y) \mapsto u_i(x, y)$.

Preferences Over Action Profiles In turn define for each agent i and all $a \in A$, $U_i(a) = u_i(f(a))$; then U_i assigns to each action profile the utility of its consequence. We will say that the action profile a' is weakly preferred to a by agent i if $U_i(a) \le U_i(a')$; U_i is called agent i's utility function *over action profiles*.

Strategic Game and Nash Equilibrium A strategic game is an ordered triple $G = \langle N, (A_i)_{i \in N}, (U_i)_{i \in N} \rangle$ where

- $N = \{1, 2, \dots, n\}$ is the set of *players*,
- A_i is the set of available *actions* for player i,
- U_i is player i's utility function over action profiles.

Rational players in a game will choose optimal actions in order to maximize utility. The most commonly used concept in game theory is that of *Nash-equilibrium*, a profile of strategies from where unilateral deviation by a single player does not pay. It can be seen as a *steady state* of action in which players hold correct beliefs about the actions taken by others and act rationally. An important assumption here is the level at which the players understand the game; usually it is taken as a starting point that players *know* the complete description of the game, including the action spaces and preferences of others.

Nash Equilibrium (Nash 1950) An action profile a^* in a strategic game $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is a Nash-equilibrium if, for every player i it holds $u_i(a^*) \ge u_i(a_i, a^*_{-i})$ for every $a_i \in A_i$.

The literature discusses several refinements of this equilibrium concept. One that will play a role in the games below is that of *strong Nash equilibrium* due to Aumann [9]; it is a Nash equilibrium a^* in a strategic game G such that for all $S \subset N$ and action profile a_S there exists a player $i \in S$ such that $u_i(a_S, a^*_{N \setminus S}) \le u_i(a^*)$. This means that a strong Nash-equilibrium guarantees stability against coordinated deviations, since within the deviating coalition there is at least one agent who does not strictly improve.

Example 15 Consider the following two-player strategic game with $N = \{1, 2\}$, $A_1 = \{T, B\}$ and $A_2 = \{L, M, R\}$. Let the utilities be as in the table below

	L	***M***	***R***
T	5,4	2,1	3,2
B	4,3	5,2	2,5

Here player 1 chooses a row, and player 2 a column. The numbers in the cells summarize the individual utilities corresponding to the action profiles; the first number is the utility of player 1, the second that of player 2. In this game there is a unique Nash-equilibrium, which is the action profile (T, L).

Dominance in Strategic Games In the game $G = \langle N, (A_i)_{i \in N}, (U_i)_{i \in N} \rangle$, the action $a_i \in A_i$ is *weakly dominated* by $a'_i \in A_i$ if $U_i(a_i, a_{-i}) \le U_i(a'_i, a_{-i})$ for all $a_{-i} \in A_{-i}$, with strict inequality for some profile of actions a_{-i}. If strict inequality holds for all a_{-i} then a_i is strictly dominated by $\tilde{a}_i$. Rational players will not use strictly dominated strategies, and, as far as prediction of play is concerned, these may be eliminated from the set of possible actions. If we do this elimination step for each player, then we may reconsider whether some actions are dominated within the reduced set of action profiles. This step-by-step reduction of action sets is called the procedure of *successive elimination of (strictly) dominated strategies*. The set of all action profiles surviving this procedure is denoted by D^∞.

Example 16 In Example 15 action M of player 2 is strictly dominated by L and R. Player 1 has no dominated actions. Now eliminate M from the actions for player 2. Then the reduced game is

	L	***R***
T	5,4	3,2
B	4,3	2,5

Notice that action L for player 1 was not dominated in the original game, for the reason that B was the better of the two actions against M. But if M is never played, T is strictly better than B. Now eliminate B, yielding the reduced game

	L	***R***
T	5,4	3,2

In this game, L dominates R; hence the only action profile surviving the successive elimination of strictly dominated strategies is (T, L).

A stronger notion than dominance is the following. Call an action $a_i \in A_i$ *overwhelmed* by $a'_i \in A_i$ if

$$\max\{U_i(a_i, a_{-i}) | a_{-i} \in A_{-i}\} < \min\left\{U_i(a'_i, a_{-i}) | a_{-i} \in A_{-i}\right\} .$$

Then $\mathcal{O}^\infty$ is the set of all action profiles surviving the successive elimination of overwhelmed actions. This notion is due to [34,37]. In Example 15 the action M is overwhelmed by L, not by R. Moreover, the remaining actions in $\mathcal{O}^\infty$ are B, T, L, and R.

Demand Games

Strategic games in cost sharing problems arise when we assume that the users of the production technology choose their demands strategically and a cost sharing rule sees to an allocation of the corresponding service costs. The action profiles A_i are simply specified by the demand spaces of the agents, and utilities are specified over combinations of (level of) received service and accompanying cost shares. Hence utilities are defined over consequences of action, $u_i(q_i, x_i)$ denotes i's utility at receiving service level q_i and cost share x_i; u_i is increasing in the level of received service x_i, and decreasing in the allocated cost y_i. Now assume a cost function c and a cost sharing rule μ. Then given a demand profile $a = (a_1, a_2, \ldots, a_n)$ the cost sharing rule determines a vector of cost shares $\mu(a, c)$, and in return also the corresponding utilities over demands $U_i(a) = u_i(a_i, \mu_i(a, c))$. Observe that agents influence each others utility via the cost component. The *demand game* for this situation is then the strategic game

$$G(\mu, c) = \langle N, (A_i)_{i \in N}, (U_i)_{i \in N}\rangle \ . \tag{11}$$

Example 17 Consider the airport problem in Example 7. Each player now may request service (1) or not (0). Then the cost function is fully described by the demand of the largest player. That is, $c(x) = 33$ if 3 requires service, $c(x) = 20$ for all profiles with $x_2 = 1, x_3 = 0$ and $c(x) = 12$ if $x = (1, 0, 0)$, $c(0, 0, 0) = 0$. Define the cost sharing rule $\overline{\Phi}(x, c) = \Phi(c_x)$, that is, $\overline{\Phi}$ calculates the Shapley-value for the underlying cost game c_x as in Eq. (1). Assume that the players' preferences over ordered pairs of service level and cost shares are fully described by

$$\begin{aligned} u_1(q_1, x_1) &= 8q_1 - x_1 \ , \\ u_2(q_2, x_2) &= 6q_2 - x_2 \ , \\ u_3(q_3, x_3) &= 30q_3 - x_3 \ . \end{aligned}$$

Here q_i takes values 0 (no service) or 1 (service) and x_i stands for the allocated cost. So player 1 prefers to be served at unit cost instead of not being served at zero cost, $u_1(0, 0) = 0 < 7 = u_1(1, 1)$. The infrastructure is seen as an excludable public good, so those with demand 0 do not get access to the technology. Each player now actively chooses to be served or not, so her action set is specified by $A_i = \{0, 1\}$. Recall the definition of c_x as in Eq. (1). Then given a profile of such actions $a = (a_1, a_2, a_3)$ and cost shares $\overline{\Phi}(a, c)$, utilities of the players in terms of action profiles become $U_i(a) = u_i(a_i, \Phi_i(a, c))$, so that

$$\begin{aligned} U_1(a) &= 8a_1 - \overline{\Phi}_1(a, c) \ , \\ U_2(a) &= 6a_1 - \overline{\Phi}_2(a, c) \ , \\ U_3(a) &= 30a_3 - \overline{\Phi}_3(a, c) \ . \end{aligned}$$

Now that we provided all details of the demand game $G(\overline{\Phi}, c)$, let us look for (strong) Nash-equilibria. Suppose that the action profile $a^* = (1, 0, 1)$ is played in the game. Then in turn the complete infrastructure is realized just for players 1 and 3 and the cost allocation is given by $(6, 0, 27)$. Then the vectors of individual utilities is given by $(2, 0, 3)$. Now if we consider unilateral deviations from a^*, what happens to the individual utilities?

$$\begin{aligned} U_1(0, 0, 1) &= 0 < 2 = U_1(1, 0, 1) \ , \\ U_2(1, 1, 1) &= 6 - \overline{\Phi}_2((1, 1, 1), c) \\ &= 6 - \Phi(c) = 6 - 8 \\ &= -2 < 0 = U_2(1, 0, 1) \ , \\ U_3(1, 0, 0) &= 0 < 3 = U_3(1, 0, 1) \ . \end{aligned}$$

This means that for each player unilateral deviation does not pay, a^* is a Nash-equilibrium. The first inequality shows as well why the action profile $(0, 0, 1)$ is not. It is easy to see that the other Nash equilibrium of this game is the action profile $(0, 0, 0)$, no player can afford the completion of the infrastructure just for herself. Notice however that this zero profile is not a strong Nash equilibrium as players 1 and 3 may well do better by choosing for service at the same time, ending up in $(1, 0, 1)$. The latter profile is the unique strong Nash-equilibrium of the game.

Similar considerations in the demand game $G(\overline{\mu}^{\mathrm{E}}, c)$ induced by the constrained egalitarian solution lead to the unique strong Nash-equilibrium $(0, 0, 0)$, nobody wants service.

With cost sharing rules as decentralization tools, the literature postulates Nash-equilibria of the related demand game as the resulting behavioral mode. This is a delicate step because – as the example above shows – it is easy to find games with many equilibria, which causes a selection problem. And what can we do if there are no equilibria? This will not be the topic of this text and the interested reader is referred to any standard textbook on game theory, for instance see [103,104,109]. If there is a unique equilibrium then it is taken as the prediction of actual play.

Demand Revelation Games

For a social planner one way to retrieve the level at which to operate the production facility is via a pre-specified demand game. Another way is to ask each of the agents for the maximal amount that she is willing to contribute in order to get service, and then, contingent on the reported amounts, install an adequate level of service together with a suitable allocation of costs. Opposed to demand games ensuring the level of service, in a demand revelation game each player is able to ensure a maximum charge for service.

The approach will be discussed under the assumption of a discrete production technology with binary demands, so that the cost function c for the technology is basically the characteristic function of a cooperative game. Moreover assume that the utilities of the agents in N are quasi-linear and given by

$$u_i(q_i, x_i) = \alpha_i q_i - x_i \tag{12}$$

where $q_i \in \{0, 1\}$ denotes the service level, x_i stands for the cost share, and α_i is a non-negative real number. [93] discusses this framework and assume that c is concave, and [64,148] moreover take c as the joint cost function for the realization of several discrete public goods.

Demand Revelation Mechanisms Formally, a *revelation mechanism M* assigns to each profile η of reported maximal contributions a set $S(\eta)$ of agents receiving service and $x(\eta)$ a vector of monetary compensations. Here we will require that these monetary compensations are cost shares; given some cost sharing rule μ the vector $x(\eta)$ is given by $\mu(\mathbf{1}_{S(\eta)}, c)$ where c is the relevant cost function.

Moreover, note that by restricting ourselves to cost share vectors we implicitly assume non-positive monetary transfers. The *budget balance* condition is crucial here, otherwise mechanisms of a different nature must be considered as well, see [23,40,41]. There are other ways that a planner may use to determine a suitable service level is by demanding pre-payment from the players, and determine a suitable service level on the basis of these labeled contributions, see [64,148].

Many mechanisms come to mind, but in order to avoid too much arbitrariness from the planner's side, the more sensible ones will grant the players some control over the outcomes. We postulate the following properties:

- **Voluntary Participation (VP)** Each agent i can guarantee herself the welfare level $u_i(0, 0)$ (no service, no payment) by reporting truthfully the maximal willingness to pay, which is α_i under Eq. (12).
- **Consumer Sovereignty (CS)** For each agent i a report y_i exists so that she receives service, irrespective of the reports by others.

Now suppose that the planner receives the message $\eta = \alpha$, known to her as the profile of true player characteristics. Then for economic reasons she could choose to serve a coalition S of players that maximizes the *net benefit* at α, $\pi(S, \alpha) = \alpha(S) - c(S)$. However, problems will arise when some player i is supposed to pay more than α_i, so the planner should be more careful than that. She may want to choose a coalition S with maximal $\pi(S, \alpha)$ such that $\mu(\mathbf{1}_S, c) \le \alpha$ holds; such set S is called *efficient*. But in general the planner cannot tell whether the players reported truthfully or not, then what should she do then? One option is that she applies the above procedure thereby naively holding each reported profile η for the *true* player characteristics. In other words, she will pick a coalition that solves the following optimization problem

$$\begin{aligned} \max_{S \subset N} \quad & \pi(S, \eta) = \eta(S) - c(S) \\ \text{s.t.} \quad & \mu(\mathbf{1}_S, c) \le \eta\,. \end{aligned} \tag{13}$$

Denote such a set by $S(\mu, \eta)$. If this set is unique then the demand revelation mechanism $M(\mu)$ selects $S(\mu, \eta)$ to be served at cost shares determined by $x(\eta) = \mu(\mathbf{1}_S, c)$. This procedure will be explained through some numerical examples.

Example 18 Consider the airport problem and utilities of players over service levels and cost shares as in Example 7. Moreover assume the planner uses the Shapley cost sharing rule $\overline{\Phi}$ as in Example 17 and that she receives the true profile of preferences from the players, $\alpha = (8, 6, 30)$. Calculate for each coalition S the net benefits at α:

S	∅	1	2	3	12	13	23	N
$\pi(S,\alpha)$	0	-4	-14	-3	-6	8	3	11

Not surprisingly, the net benefits are the highest for the grand coalition. But if N were selected by the mechanism the corresponding cost shares are given by $\overline{\Phi}(\mathbf{1}_N, c) = (4, 8, 21)$, and player 2 is supposed to contribute more than she is willing to. Then the second highest net benefits are generated by serving $S = \{1, 3\}$, with cost shares $\overline{\Phi}(\mathbf{1}_S, c) = (6, 0, 27)$. Then $\{1, 3\}$ is the solution to Eq. (13).

What happens if some of the players misrepresent their preferences, for instance like in $\eta = (13, 6, 20)$? The planner determines the *conceived* net benefits

S	$\varnothing$	1	2	3	12	13	23	N
$\pi(S, \eta)$	0	1	-14	-13	-1	0	-7	6

Again, if the planner served the coalition with the highest net benefit, N, then player 2 would refuse to pay. Second highest net benefit corresponds to the singleton $S = \{1\}$, and this player will get service under $M(\overline{\Phi})$ since $\eta_3 = 13 > 12 = c(1, 0, 0)$.

Example 19 Consider the same situation but now with $\overline{\mu}^{\mathrm{E}}$ instead of $\overline{\Phi}$ as cost sharing rule. Now consider the following table, restricted to coalitions with non-negative net benefits (all other will not be selected):

S	$\varnothing$	13	23	N
$\pi(S, \alpha)$	0	8	3	11
$\overline{\mu}^{\mathrm{E}}(\mathbf{1}_S, c)$	(0,0,0)	(12,0,21)	$(0, \frac{33}{2}, \frac{33}{2})$	(10, 10, 13)

Here only the empty coalition $S = \varnothing$ satisfies the requirement $\overline{\mu}^{\mathrm{E}}(\mathbf{1}_S, c) = (0, 0, 0) \leq (8, 6, 30)$; hence according to the mechanism $M(\overline{\mu}^{\mathrm{E}})$ nobody will get service.

In general, the optimization problem Eq. (13) does not give unique solutions in case of which a planner should still further specify what she does in those cases. For concave cost functions, consider the following sequence of coalitions

$$S_1 = N, \; S_t = \{\, i \in S_{t-1} \,|\, \eta_i \geq \mu_i(\mathbf{1}_{S_t}, c) \} \; .$$

So, starting with the grand coalition N, at each consecutive step those players are removed whose maximal contributions are not consistent with the proposed cost share – until the process settles down. The set of remaining players defines a solution to Eq. (13) and taken to define $S(\mu, \eta)$.

Strategyproofness The essence of a demand revelation mechanism M is that its rules are set up in such a way that it provides enough incentives for the players not to lie about their true preferences. We will now discuss the most common non-manipulability properties of revelation mechanisms in the literature.

Fix two profiles $\alpha', \alpha \in \mathbb{R}_+^N$, where α corresponds to the true maximal willingness' to pay. Let (q', x') and (q, x) be the allocations implemented by the mechanism M on receiving the messages α' and α, respectively. The mechanism M is called *strategy-proof* if it holds for all $i \in N$ that $\alpha'_{N\backslash\{i\}} = \alpha_{N\backslash\{i\}}$ implies $u_i(q'_i, x'_i) \leq u_i(q_i, x_i)$.

So, given the situation that the other agents report truthfully, unilateral deviation by agent i from the true preference never produces better outcomes for her. Similarly, M is *group strategy-proof* if deviations of groups of agents does not pay for all deviators, i. e., for all $T \subset N$ the fact that $\alpha'_{N\backslash T} = \alpha_{N\backslash T}$ implies $u_i(q'_i, x'_i) \leq u_i(q_i, x_i)$ for all $i \in T$. So, under a (group) strategy-proof mechanism, there is no incentive to act untruthfully by misrepresenting the true preferences and this gives a benevolent planner control over the outcome.

Cross-Monotonicity Cost sharing rule μ is called *cross-monotonic* if the cost share of an agent i is not increasing if other agents demand more service, in case of a concave cost function c. Formally, if $\overline{x} \geq x$ and $\overline{x}_i = x_i$ then $\mu_i(\overline{x}, c) \leq \mu_i(x, c)$; each agent is granted a (fair) share of the positive externality due to an increase in demand by others.

Proposition 1 (Moulin & Shenker [93]) *The only group strategy-proof mechanisms $M(\mu)$ satisfying VP and CS are those related to cross-monotonic cost sharing rules μ.*

There are many different cross-monotonic cost sharing rules, and thus just as many mechanisms that are group strategy-proof. Examples include the mechanisms $M(\overline{\Phi})$ and $M(\overline{\mu}^{\mathrm{E}})$, because $\overline{\Phi}$ and $\overline{\mu}^{\mathrm{E}}$ are cross-monotonic. However, the nucleolus is not cross-monotonic and does therefore not induce a strategy-proof mechanism.

Above we discussed two instruments a social planner may invoke to implement a desirable outcome without knowing the true preferences of the agents. Basically, the demand revelation games define a so-called *direct mechanism*. The announcement of the maximal price for service pins down the complete preferences of an agent. So in fact the planner decides upon the service level based on a complete profile of preferences. In case of a cross-monotonic cost sharing rule, under the induced mechanism truth-telling is a weakly dominant strategy; announcing the true maximal willingness to pay is optimal for the agent regardless of the actions of others. This means good news for the social planner as the mechanism is self-organizing: the agents need not form any conjecture about the behavior of others in order to know what to do. In the literature [24] such a mechanism is called *straightforward*. The demand games define an *indirect mechanism*, as by reporting a demand the agents do no more than signaling their preferences to the planner.

Although in general there is a clear distinction between direct and indirect mechanisms, in the model presented in this section these are nevertheless strongly connected. Focus on a demand game $G(\mu, c)$; recall that in this game the agents simultaneously and independently decide upon requesting service or not and costs are shared using the rule μ amongst those agents receiving service. Suppose that for each profile of utility functions as in Eq. (12) the resulting game $G(\mu, c)$ has a unique (strong) Nash-equilibrium. Then this equilibrium can be taken to define a mechanism. That is, the mechanism elicits u and chooses the unique equilibrium *outcome* of the reported demand game. Then this mechanism is equivalent with the demand revelation mechanism. Observe that indeed the strong equilibrium $(1, 0, 1)$ in the game $G(\overline{\Phi}, c)$ in Example 17 corresponds to players chosen by $M(\overline{\Phi})$ under truthful reporting. And where no player is served in the strong equilibrium of $G(\overline{\mu}^{\mathrm{E}}, c)$, none of the players is selected by $M(\overline{\mu}^{\mathrm{E}})$. It is a general result in implementation theory due to [24] that a direct mechanism constructed in this way is (group) strategy-proof *provided* the underlying space of preferences is *rich*. It is easily seen that the above sets of preferences meet the requirements. To stress importance of such a structural property as richness, it is instructive to point at what is yet to come in Sect. "Uniqueness of Nash-Equilibria in $\mathcal{P}^1$-Demand Games". Here, the strategic analysis of the demand game induced by the proportional rule shows uniqueness of Nash equilibrium on the domain of preferences $\mathcal{L}^*$ if only costs are convex. However, this domain is *not* rich, and the direct mechanism defined in the same fashion as above by the Nash equilibrium selection is *not* strategyproof.

Efficiency and Strategy-Proof Cost Sharing Mechanisms

Suppose *cardinal utility* for each agent, so that inter-comparison of utility is allowed. Proceeding on the net benefit of a coalition, we may define its *value* at α by

$$v(N, \alpha) = \max_{S \subset N} \pi(S, \alpha) \,, \tag{14}$$

where $\pi(S, \alpha)$ is the net benefit of S at α. A coalition S such that $v(N, \alpha) = \pi(S, \alpha)$ is called *efficient*. It will be clear that a mechanism $M(\mu)$ that is defined through the optimization problem Eq. (13) will not implement an efficient coalition of served players, due to the extra constraint on the cost shares.

For instance, in Example 7 the value of the grand coalition at $\alpha = (8, 6, 30)$ is given by $v(N, \alpha) = \alpha(N) - c(N) = 44 - 33 = 11$. At the same profile the implemented outcome by mechanism $M(\overline{\Phi})$ gives rise to a total surplus of $38 - 30 = 8$ for the grand coalition – which is not optimal. The mechanism $M(\overline{\mu}^{\mathrm{E}})$ performs even worse as it leads to the stand alone surplus 0, none is served.

This observation holds for far more general settings, and, moreover, it is a well known result from implementation theory that – under non-constant marginal cost – any strategy-proof mechanism based on full coverage of total costs will not always implement efficient outcomes. For the constant marginal cost case see [67,71]. Then, if there is an unavoidable loss in using demand revelation mechanisms, can we still tell which mechanisms are more efficient? Is it a coincidence that in the above examples the Shapley value performs better than the egalitarian solution?

The *welfare loss* due to $M(\mu)$ at a profile of true preferences α is given by

$$L(\mu, \alpha) = v(N, \alpha) - \{\alpha(S(\mu, \alpha)) - c(S)\} \,. \tag{15}$$

For instance, with $\alpha = (8, 6, 30)$ in the above examples we calculate $L(\overline{\Phi}, \alpha) = 11 - 8 = 3$ and $L(\overline{\mu}^{\mathrm{E}}, \alpha) = 11 - 0 = 11$. An overall measure of quality of a cost sharing rule μ in terms of efficiency loss is defined by

$$\gamma(\mu) = \sup_\alpha L(\mu, \alpha) \,. \tag{16}$$

Theorem 6 (Moulin & Shenker [93]) *Among all mechanisms $M(\mu)$ derived from cross-monotonic cost sharing rules μ, the Shapley rule $\overline{\Phi}$ has the unique smallest maximal efficiency loss, or $\gamma(\overline{\Phi}) < \gamma(\mu)$ if $\mu \neq \overline{\Phi}$.*

Notice that this makes a strong case for the Shapley-value against the egalitarian solution. The story does, however, not end here.

[98] considers a model where the valuations of the agents for the good are independent random variables, drawn from a distribution function F satisfying the *monotone hazard condition*. This means that the function defined by $h(x) = f(x)/(1 - F(x))$ is non-decreasing, where F has f as density function. It is shown that the constrained egalitarian solution maximizes the probability that all members of any given coalition accept the cost shares imputed to them. Moreover, [98] characterized the solution in terms of efficiency. Suppose for the moment that the planner calculated cost share vector $\mathbf{x}$ for the coalition S, and that its members are served conditional on acceptance of the proposed cost shares. The probability that all members of S accept the shares is given by $P(x) = \prod_{i \in S}(1 - F(x_i))$, and if we assume that the support of F is $(0, m)$ then the expected surplus from such an offer can be calculated as follows

$$W(x) = P(x) \cdot \left[\sum_{i \in S} \int_{x_i}^{m} \frac{u_i}{1 - F(x_i)} \mathrm{d}F(u_i) - c(S) \right] . \tag{17}$$

The finding of [98] is that for *log-concave* f, i.e. $x \mapsto \ln(f(x))$ is concave [5], the mechanism based on the constrained egalitarian solution not only maximizes the probability that a coalition accepts the proposal, but it maximizes its expected surplus as well. Formally, the result is the following.

Theorem 7 (Mutuswami (2004)) *If the profile of valuations $(u_i)_{i \in N}$ are independently drawn form a common distribution function F with log-concave and differentiable density function f, then $W(\overline{\mu}^{\mathrm{E}}(\mathbf{1}_S, c)) \geq W(\mu(\mathbf{1}_S, c))$ for all cross monotonic solutions μ and all $S \subset N$.*

Extension of the Model: Discrete Goods

Suppose the agents consume idiosyncratic goods produced in indivisible units. Then given a profile of demands the cost associated with the joint production must be shared by the users. Then this model generalizes the binary good model discusses so far and it is a launch-pad to the continuous framework in the next section. In this discrete good setting [86] characterizes the cost sharing rules which induce strategyproof social choice functions defined by the equilibria of the corresponding demand game. As it turns out, these rules are *basically* the *sequential stand alone rules*, according to which costs are shared in an incremental fashion with respect to a fixed ordering of the agents. This means that such a rule charges the first agent for her stand alone costs, the second for the stand alone cost for the first two users minus the stand alone cost of the first, et cetera. Here the word 'basically' refers to all discrete cost sharing problems other than those with binary goods. Then here the sufficient condition for strategyproofness is that the underlying cost sharing rule be cross monotonic, which admits other rules than the sequential ones – like $\overline{\mu}^{\mathrm{E}}$ and $\overline{\Phi}$.

Continuous Cost Sharing Models

Continuous Homogeneous Output Model, $\mathcal{P}^1$

This model deals with production technologies for one single perfectly divisible output commodity. Moreover, we will restrict ourselves to private goods. Many ideas below have been studied for *public goods* as well, for further references see, e. g., [33,72,82].

The demand space of an individual is given by $X = \mathbb{R}_+$. The technology is described by a non-decreasing cost function $c\colon \mathbb{R}_+ \to \mathbb{R}$ such that $c(0) = 0$, i. e. there are no fixed costs. Given a profile of demands $x \in \mathbb{R}_+^N$ costs $c(x(N))$ have to be shared. Moreover, the space of cost functions will be restricted to those c being *absolutely continuous*. Examples include the differentiable and Lipschitz-continuous functions. Absolute continuity implies that aggregate costs for production can be calculated by the total of marginal costs

$$c(y) = \int_0^y c'(t)\,\mathrm{d}t\,.$$

Denote the set of all such cost functions by C^1 and the related cost sharing problems by $\mathcal{P}^1$. Several cost sharing rules on $\mathcal{P}^1$ have been proposed in the literature.

Average Cost Sharing Rule This is the most popular and oldest concept in the literature and advocates Aristotle's principle of proportionality.

$$\mu^{\mathrm{AV}}(x, c) = \begin{cases} \frac{x}{x(N)} \cdot c(x(N)) & \text{if } x \neq 0_N \\ 0 & \text{if } x = 0_N \end{cases} \tag{18}$$

Shapley–Shubik Rule Each cost sharing problem $(x, c) \in \mathcal{P}^1$ is related to the stand-alone cost game c_x such that $c_x(S) = c(x(S))$. Then the Shapley–Shubik rule is determined by application of the Shapley-value to this game:

$$\mu^{\mathrm{SS}}(x, c) = \Phi(c_x)\,.$$

Serial Rule This rule, due to Moulin and Shenker [91], determines cost shares by considering particular intermediate cost levels. More precisely, given $(x, c) \in \mathcal{P}^1$ it first relabels the agents by increasing demands such that $x_1 \leq x_2 \leq \cdots \leq x_n$. The intermediate production levels are

$$\begin{aligned} y^1 &= nx_1,\ y^2 = x_1 + (n-1)x_2,\ \ldots\,, \\ y^k &= \sum_{j=1}^{k-1} x_j + (n-k+1)x_k,\ \ldots\,, \\ y^n &= x(N)\,. \end{aligned}$$

These levels are chosen such that at each new level one agent more is fully served his demand: at y^1 each agent is handed out x_1, at y^2 agent 1 is given x_1 and the rest x_2, etc. The serial cost shares are now given by

$$\mu_i^{\mathrm{SR}}(x, c) = \sum_{\ell=1}^{k} \frac{c(y^\ell) - c(y^{\ell-1})}{n - \ell + 1}\,.$$

So according to μ^{SR} each agent pays a fair share of the incremental costs in each stage that she gets new units.

Example 20 Consider the cost sharing problem (x, c) with $x = (10, 20, 30)$ and $c(y) = \frac{1}{2}y^2$. Then first calculate the intermediate production levels $y^0 = 0$, $y^1 = 30$, $y^2 = 50$, and $y^3 = 60$. Then the cost shares are calculated as follows

$$\mu_1^{\mathrm{SR}}(x, c) = \frac{c(y^1) - c(y^0)}{3} = 150\,,$$

$$\begin{aligned}\mu_2^{\mathrm{SR}}(x,c) &= \mu_1^{\mathrm{SR}}(x,c) + \frac{c(y^2)-c(y^1)}{2}\\ &= 150 + \frac{1250-450}{2} = 550\,,\\ \mu_3^{\mathrm{SR}}(x,c) &= \mu_2^{\mathrm{SR}}(x,c) + c(y^3) - c(y^2) = 550 + 550\\ &= 1100\,.\end{aligned}$$

The serial rule has attracted much attention lately in the network literature, and found its way in fair queuing packet scheduling algorithms in routers [27].

Decreasing Serial Rule De Frutos [26] proposes serial cost shares where demands of agents are put in decreasing order. Resulting is the *decreasing serial rule*. Consider a demand vector $x \in \mathbb{R}_+^N$ such that $x_1 \leq x_2 \leq \cdots \leq x_n$. Define recursively the numbers y^ℓ for $\ell = 1, 2, \ldots, n$ by $y^\ell = \ell x_\ell + x_{\ell+1} + \cdots + x_n$, and put $y^{n+1} = 0$. Then the decreasing serial rule is defined by

$$\mu_i^{\mathrm{DSR}}(x,c) = \sum_{\ell=i}^{n} \frac{c(y^\ell) - c(y^{\ell+1})}{\ell}\,. \tag{19}$$

Example 21 For the cost sharing problem in Example 20 calculate $y^1 = 90$, $y^2 = 70$, $y^1 = 60$, then

$$\begin{aligned}\mu_3^{\mathrm{DSR}}(x,c) &= \frac{c(y^3)-c(y^4)}{3} = \frac{4050-0}{3} = 1350\,,\\ \mu_2^{\mathrm{DSR}}(x,c) &= \mu_3^{\mathrm{DSR}}(x,c) + \frac{c(y^2)-c(y^3)}{2}\\ &= 1350 + \frac{2450-4050}{2} = 550\,,\\ \mu_1^{\mathrm{DSR}}(x,c) &= \mu_2^{\mathrm{DSR}}(x,c) + (c(y^1)-c(y^2))\\ &= 550 + (1800-2450) = -100\,.\end{aligned}$$

Notice that here the cost share of agent 1 is negative, due to the convexity of c! This may be considered as an undesirable feature of the cost sharing rule. Not only are costs increasing in the level of demand, in case of a convex cost function each agent contributes to the negative externality. It seems fairly reasonable to demand a non-negative contribution in those cases, so that none profits for just being there. The mainstream cost sharing literature includes positivity of cost shares into the very definition of a cost sharing rule. Here we will add it as a specific property:

Positivity μ is *positive* if $\mu(x,c) \geq 0_N$ for all (x,c) in its domain. All earlier discussed cost sharing rules have this property, except for the decreasing serial rule.

The decreasing serial rule is far more intuitive in case of economies of scale, in presence of a concave cost function. The larger agents now are credited with a lower price per unit of the output good. [47,57] propose variations on the serial rule that coincide with the increasing (decreasing) serial rule in case of a convex (concave) cost function, meeting the positivity requirement.

Marginal Pricing Rule A popular way of pricing an output of a production facility is *marginal cost pricing*. The price of the output good is set to cover the cost producing one extra unit. It is frequently used in the domain of public services and utilities. However, a problem is that for concave cost functions the method leads to budget deficits. An adapted form of marginal cost pricing splits these deficits equally over the agents. The *marginal pricing rule* is defined by

$$\mu_i^{\mathrm{MP}}(x,c) = x_i c'(x(N)) + \tfrac{1}{n}\left[c(x(N)) - x(N)c'(x(N))\right]. \tag{20}$$

Note that in case of convex cost functions agents can receive negative cost shares, just like it is the case with decreasing serial cost sharing.

Additive Cost Sharing and Rationing

The above cost sharing rules for homogeneous production models share the following properties:

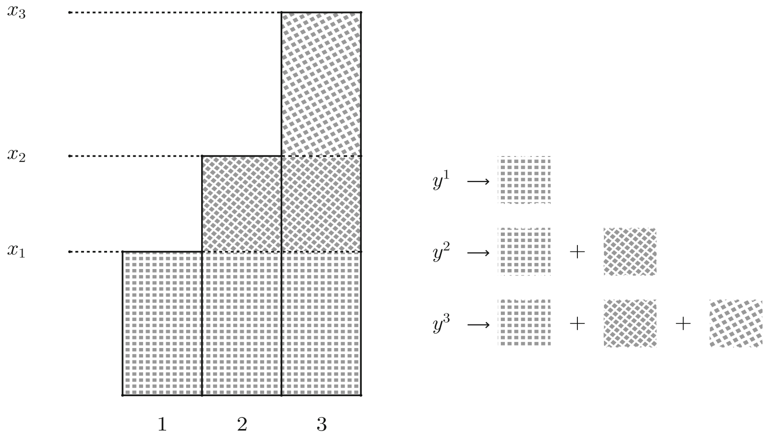
Additivity $\mu(x, c_1 + c_2) = \mu(x, c_1) + \mu(x, c_2)$ for all relevant cost sharing problems. This property carries the same flavor as the homonymous property for cost games.

Constant Returns $\mu(x,c) = \vartheta x$ for linear cost functions c such that $c(y) = \vartheta y$ for all y. So if the agents do not cause any externality, the fixed marginal cost is taken as a price for the good.

It turns out that the class of all positive cost sharing rules with these properties can be characterized by solutions to *rationing problems* – which are the most basic of all models of distributive justice. A rationing problem amongst the agents in N consists of a pair $(x,t) \in \mathbb{R}_+^N \times \mathbb{R}_+$ such that $x(N) \geq t$; t is the available amount of some (in)divisible good and x is the set of demands. The inequality sees to the interpretation of rationing as not every agent may get all she wants.

A *rationing method* r is a solution to rationing problems, such that each problem (x,t) is assigned a vector of shares $r(x,t) \in \mathbb{R}_+^N$ such that $\mathbf{0}_N \leq r(x,t) \leq x$. The latter restriction is a weak solidarity statement assuring that everybody's demand be rationed in case of shortages.

For $t \in \mathbb{R}_+$ define the special cost function Γ_t by $\Gamma_t(y) = \min\{y, t\}$. The cone generated by these base



Cost Sharing, Figure 7
Intermediate production levels

functions lays dense in the space of all *absolutely continuous* cost functions c; if we know what the values $\mu(x, \Gamma_t)$ are, then basically we know $\mu(x, c)$. Denote by $\mathcal{M}$ the class of all cost sharing rules with the properties positivity, additivity, and constant returns.

Theorem 8 (Moulin & Shenker [88,92]) *Consider the following mappings associating rationing methods with cost sharing rules and vice versa,*

$$\begin{aligned} r \mapsto \mu &: \mu(x, c) = \int_0^{x(N)} c'(t) dr(x, t) \,, \\ \mu \mapsto r &: r(x, t) = \mu(x, \Gamma_t) \,. \end{aligned}$$

These define an isomorphism between $\mathcal{M}$ and the space of all monotonic rationing methods.

So each monotonic rationing method relates to a cost sharing rule and vice versa. In this way μ^{P} is directly linked with the proportional rationing method, μ^{SR} to the *uniform gains* method, and μ^{SS} to the *random priority* method. Properties of rationing methods lead to properties of cost sharing rules and vice versa [61].

Incentives in Cooperative Production

Stable Allocations, Stand-Alone Core Suppose again, like in the framework of cooperative cost games, that (coalitions of) agents can decide to leave the cost sharing and organize their own production facility. Under the ability to replicate the technology the question arises whether cost sharing rules induce stable cost share vectors.

Theorem 9 (Moulin [85]) *For concave cost functions c, if μ is an anonymous and cross-monotonic cost sharing rule then $\mu(x, c) \in$* core(c_x).

Under increasing returns to scale, this implies that $\mu^{\mathrm{P}}, \mu^{\mathrm{SR}}$ are core-selectors but μ^{MC} is not. [137] associates to each cost sharing problem (x, c) a *pessimistic* one, (x, c^*); here $c^*(y)$ reflects the maximum of marginal cost on $[0, x(N)]$ to produce y units,

$$c^*(y) = \begin{cases} \sup\left\{\int_T c'(t)\mathrm{d}t \mid T \subset [0, x(N)], \lambda(T) = y\right\} & \\ & \text{if } y \leq x(N)\,, \\ c(y) & \text{else}\,. \end{cases} \tag{21}$$

Here λ denotes the Lebesgue measure.

Theorem 10 (Koster [60]) *For any cost sharing problem (x, c), it holds* core$(c_x^*) = \{\mu(x, c) \mid \mu \in \mathcal{M}\}$.

In particular this means that for $\mu \in \mathcal{M}$ it holds that $\mu(x, c) \in$ core(c_x) whenever c is concave, since this implies $c^* = c$. This result appeared earlier as a corollary to Theorem 8, see [88]. [47,57,61] show *non-linear* cost sharing rules yielding core elements for concave cost functions as well, so additivity is only a sufficient condition in the above statement. For average cost sharing, one can show more, $\mu^{\mathrm{P}}(x, c) \in$ core(c_x) for all x precisely when the average cost $c(y)/y$ is decreasing in y.

Strategic Manipulation Through Reallocation of Demands In the cooperative production model, there are other ways that agents may use to manipulate the final allocation. In particular, note that the serial procedure gives the larger demanders an advantage in case of positive externalities; as marginal costs decrease, the price paid by the larger agents per unit of output is lower than that of the smaller agents. In the other direction larger demanders are punished if costs are convex. Then as the examples below show, this is just why the serial ideas are vulnerable to misrepresentation of demands, since combining demands and redistribute the output afterwards can be beneficial.

Example 22 Consider the cost function c given by $c(y) = \min\{5y, 60 + 2y\}$. Such cost functions are part of daily life, whenever one has to decide upon telephone or energy supply contracts: usually customers get to choose between a contract with high fixed cost and a low variable cost, and another with low or no fixed cost and a high variable price. Now consider the two cost sharing problems (x, c) and (x', c) where $x = (10, 20, 30)$, $x' = (0, 30, 30)$. The cost sharing problem (x', c) arises from (x, c) if agent 2 places a demand on behalf of agent 1 – without letting agent 3 know. The corresponding average and serial cost shares are given by

$$\begin{aligned} \mu^{\mathrm{P}}(x,c) &= (30,60,90) \quad & \mu^{\mathrm{P}}(x',c) &= (0,90,90) \\ \mu^{\mathrm{SR}}(x,c) &= (40,60,80) \quad & \mu^{\mathrm{SR}}(x',c) &= (0,90,90)\,. \end{aligned}$$

Notice that the total of average cost shares for 1 and 2 is the same in both cost sharing problems. But if the serial rule were used, these agents can profit by *merging* their demands; if agent 2 returns agent 1's demand and requires a payment from agent 1 between €30 and €40, then both agents will have profited by such *merging* of demand.

Example 23 Consider the five-agent cost sharing problems (x, c) and $(\overline{x}, c)$ with $x = (1, 2, 3, 0, 0)$, $\overline{x} = (1, 2, 1, 1, 1)$ and convex cost function $c(y) = \frac{1}{2}y^2$. $(\overline{x}, c)$ arises out of (x, c) if agent 3 splits her demand over agents 4 and 5 as well. Then

$$\begin{aligned} \mu^{\mathrm{P}}(x,c) &= (6,12,18,0,0) \quad & \mu^{\mathrm{P}}(\overline{x},c) &= (6,12,6,6,6)\,, \\ \mu^{\mathrm{SR}}(x,c) &= (3,11,22,0,0) \quad & \mu^{\mathrm{SR}}(\overline{x},c) &= (5,16,5,5,5)\,. \end{aligned}$$

The aggregate of average cost shares for agents 3, 4, and 5 does not change. But notice that according to the serial cost shares, there is a clear advantage for the agents. Instead of paying 22 in the original case, now the total of payments equals 15. Agent 3 may consider a transfer between 0 and 7 to 3 and 4 for their collaboration and still be better of. In general, in case of a convex cost function the serial rule is vulnerable with respect to manipulation of demands through splitting.

Note that in the above cases the proportional cost sharing rule does prescribe the same cost shares. It is a non-manipulable rule: reshuffling of demands will not lead to different aggregate cost shares. The rule does not discriminate between units, when a unit is produced is irrelevant. It is actually a very special feature of the cost sharing rule that is basically not satisfied by any other cost sharing rule.

Theorem 11 *Assume that N contains at least three agents. The proportional cost sharing rule is the unique rule that charges nothing for a null demand and meets any one of the following properties:*

- *Independence of merging and splitting,*
- *No advantageous reallocation,*
- *Irrelevance of reallocation.*

The second property shows even a stronger property than merging and splitting: agents may redistribute the demands in any preferred way without changing the aggregate cost shares of the agents involved. The third property states that in such cases the cost shares of the other agents do not change. Then this makes proportional cost sharing compelling in situations where one is not capable of detecting the true demand characteristics of individuals.

Demand Games for $\mathcal{P}^1$

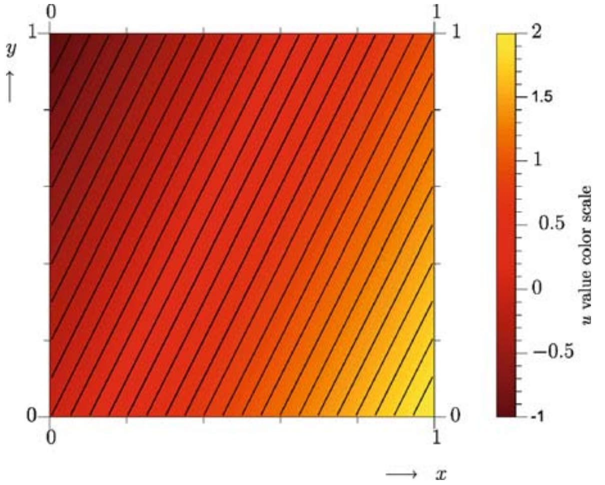
Consider demand games $G(\mu, c)$ as in Eq. (11), Sect. "Demand Games", where now μ is a cost sharing rule on $\mathcal{P}^1$. These games with uncountable strategy spaces are more complex than the demand games that we studied before.

The set of consequences for players is now given by $C = \mathbb{R}^2_+$, combinations of levels of production and costs (see Sect. "Strategic Demand Games"). Then an individual i's preference relation is *convex* if for the corresponding utility function u_i and all pairs $z, z' \in C$ it holds

$$\begin{aligned} &u_i(z) = u_i(z') \\ &\quad \Longrightarrow\; u(tz + (1-t)z) \geq u_i(z) \quad \text{for all } t \in [0,1]\,. \end{aligned} \tag{22}$$

This means that a weighted average of the consequences is weakly preferred to both consequences, if these are equivalent. Such utility functions u_i are called *quasi-concave*. An example of convex preferences are those related to *linear utility* functions of type $u_i(x, y) = \alpha x - y$. Moreover, *strictly convex* preferences are those with strict inequality in Eq. (22) for $0 < t < 1$; the corresponding utility functions are *strictly quasi-concave*. Special classes of preferences are the following.

- $\mathcal{L}$: the class of all convex and continuous preferences utility functions that are non-decreasing in the service component x, non-increasing in the cost component y, non-locally satiated and decreasing on $(x, c(x))$ for x large enough. The latter restriction is no more than assuring that agents will not place requests for unlimited amounts of the good.



Cost Sharing, Figure 8
Linear, convex preferences, $u(x, y) = 2x - y$. The contours indicate indifference curves, i.e. sets of type $\{(x, y)|u(x, y) = k\}$, the k-level curve of u

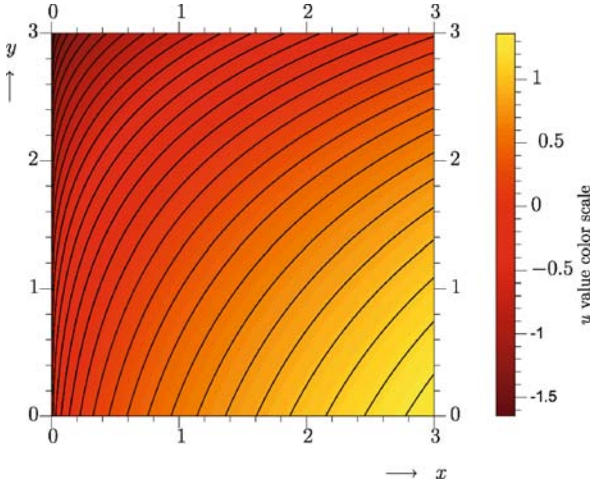
- $\mathcal{L}^*$: the class of *bi-normal* preferences in $\mathcal{L}$. Basically, if such a preference is represented by a differentiable utility function u then the slope $\mathrm{d}y/\mathrm{d}x$ of the indifference contours is non-increasing in x, non-decreasing in y. For a concise definition see [141]. Examples include Cobb–Douglas utility functions and also those of type $u_i(x, y) = \alpha(x) - \beta(y)$ where α and β are concave and convex functions, respectively. A typical plot of level curves of such utility functions is in Fig. 9. Note that the approach differs from the standard literature where agents have preferences over endowments. Here costs are 'negative' endowments. In the latter interpretation, the condition can be read as that the *marginal rate of substitution* is non-positive. At equal utility, an increase of the level of output has to be compensated by a decrease in the level of input good.

Nash-Equilibria of Demand Games in a Simple Case

Consider a production facility shared by three agents $N = \{1, 2, 3\}$ with cost function $c(y) = \frac{1}{2}y^2$. Assume that the agents have quasi-linear utilities in $\mathcal{L}$, i.e. $u_i(x_i, y_i) = \alpha_i x_i - y_i$ for all pairs $(x_i, y_i) \in \mathbb{R}^2_+$. Below the Nash-equilibrium in the serial and proportional demand game is calculated in two special cases. This numerical example is based on [91].

Proportional Demand Game Consider the corresponding proportional demand game, $G(\mu^{\mathrm{P}}, c)$, with utility over actions given by

$$U_i^{\mathrm{P}}(x) = \alpha_i x_i - \mu^{\mathrm{P}}(x, c) = \alpha_i x_i - \tfrac{1}{2} x_i x(N)\,. \qquad (23)$$



Cost Sharing, Figure 9
Strictly convex preferences, $u(x, y) = \sqrt{x} - \mathrm{e}^{0.5y}$. The straight line connecting any two points on the same contour lays in the lighter area – with higher utility. Fix the y-value, then an increase of x yields higher utility, whereas for fixed x an increase of y causes the utility to decrease

In a Nash-equilibrium x^* of $G(\mu^{\mathrm{P}}, c)$ each player i gives a best response on x^*_{-i}, the action profile of the other agents. That is, player i chooses $x^*_i \in \arg\max_t U_i^{\mathrm{P}}(t, x^*_{-i})$. Then first order conditions implies for an interior solution

$$\alpha_i - \tfrac{1}{2} x^*(N) - \tfrac{1}{2} x^*_i = 0 \qquad (24)$$

for all $i \in N$. Then $x^*(N) = \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3)$ and $x^*_i = 2\alpha_i - \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3)$.

Serial Demand Game Consider the same production facility and the demand game $G(\mu^{\mathrm{SR}}, c)$, corresponding to the serial rule. Then the utilities over actions are given by

$$U_i^{\mathrm{SR}}(x) = \alpha_i x_i - \mu^{\mathrm{SR}}(x, c)\,. \qquad (25)$$

Now suppose $\overline{x}$ is a Nash equilibrium of this game, and assume without loss of generality that $\overline{x}_1 \le \overline{x}_2 \le \overline{x}_3$. Then player 1 with the smallest equilibrium demand maximizes the expression

$$U_1^{\mathrm{SR}}((t, \overline{x}_2, \overline{x}_3) = \alpha_1 t - c(3t)/3 = \alpha_1 t - \tfrac{3}{2}t^2$$

at $\overline{x}_1$, from which we may conclude that $\alpha_1 = 3\overline{x}_1$. In addition, in equilibrium, player 2, maximizes

$$U_2^{\mathrm{SR}}(\overline{x}_1, t, \overline{x}_3) = \alpha_2 t - (\tfrac{1}{3}c(3\overline{x}_1) + \tfrac{1}{2}(c(\overline{x}_1 + 2t) - c(3\overline{x}_1))\,,$$

for $t \ge \overline{x}_1$, yielding $\alpha_2 = \overline{x}_1 + 2\overline{x}_2$. Finally, the equilibrium condition for player 3 implies $\alpha_3 = \overline{x}(N)$. Then it is not hard to see that actually this constitutes *the* serial equilibrium.

Comparison of Proportional and Serial Equilibria (I)

Now let's compare the serial and the proportional equilibrium in the following two cases:

$$\begin{aligned}&\text{(i)} && \alpha_1 = \alpha_2 = \alpha_3 = \alpha\\ &\text{(ii)} && \alpha_1 = \alpha_2 = 2, \alpha_3 = 4\,.\end{aligned}$$

Case (i): Then we get $x^* = (\frac{1}{2}\alpha, \frac{1}{2}\alpha, \frac{1}{2}\alpha)$ and $\overline{x}_i = (\frac{1}{3}\alpha, \frac{1}{3}\alpha, \frac{1}{3}\alpha)$ for all i. The resulting equilibrium payoff vectors are given by

$$U^{\mathrm{P}}(x^*) = (\tfrac{1}{8}\alpha^2, \tfrac{1}{8}\alpha^2, \tfrac{1}{8}\alpha^2) \text{ and}$$
$$U^{\mathrm{SR}}(\overline{x}) = (\tfrac{1}{6}\alpha^2, \tfrac{1}{6}\alpha^2, \tfrac{1}{6}\alpha^2)\,.$$

Not only the average outcome is less efficient than its serial counterpart, it is also Pareto-inferior to the latter.

Case (ii): The proportional equilibrium is a boundary solution, $x^* = (0, 0, 2)$ with utility profile $U^{\mathrm{P}}(x^*) = (0, 0, 8)$. The serial equilibrium strategies and utilities are given by

$$\overline{x} = (\tfrac{2}{3}, \tfrac{2}{3}, \tfrac{8}{3}), U^{\mathrm{SR}}(\overline{x}) = (\tfrac{2}{3}, \tfrac{2}{3}, 4)\,.$$

Notice that the serial equilibrium is now less efficient, but is not Pareto-dominated by the proportional utility distribution.

Uniqueness of Nash-Equilibria in $\mathcal{P}^1$-Demand Games

In the above demand games there is a unique Nash-equilibrium which serves as a prediction of actual play. This need not hold for any game. In the literature strategic characterizations of cost sharing rules are discussed in terms of uniqueness of equilibrium in the induced cost games, relative to specific domains of preferences and cost functions. Below we will discuss the major findings of [141]. These results concern a broader cost sharing model with the notion of a cost function as a differentiable function $\mathbb{R}_+ \to \mathbb{R}_+$. So in this paragraph such cost function can decrease, fixed cost need not be 0. This change in setup is not crucial to the overall exposition since the characterizations below are easily interpreted within the context of $\mathcal{P}^1$.

Demand Monotonicity The mapping $t \mapsto \mu_i((t, x_{-i}), c)$ is non-decreasing; μ is strictly demand monotonic if this mapping is increasing whenever c is increasing.

Smoothness The mapping $x \mapsto \mu(x, c)$ is continuously differentiable for all continuously differentiable $c \in C^1$.

Recall that $\mathcal{L}^*$ is the domain of all bi-normal preferences.

Theorem 12 (Watts [141]) *Fix a differentiable cost function c and a demand monotonic and smooth cost sharing rule μ. A cost sharing game $G(\mu, c)$ has a unique equilibrium whenever agents' preferences belong to $\mathcal{L}^*$, only if, for all $x = (x_1, \ldots, x_n)$*

- *Every principal minor of the matrix W with rows w^i is non-negative for all*

$$w^i \in \left\{ \left(\frac{\partial \mu_i}{\partial x_1}, \ldots, \frac{\partial \mu_i}{\partial x_n} \right), \left(\frac{\partial^2 \mu_i}{\partial x_1 \partial x_1}, \ldots, \frac{\partial^2 \mu_i}{\partial x_i \partial x_n} \right) \right\}. \tag{26}$$

- *The determinant of the Hessian matrix corresponding to the mapping $x \mapsto \mu(x, c)$ is strictly positive.*

A sufficient condition to have uniqueness of equilibrium is that the principle minor of the matrix W is strictly positive.

The impact of this theorem is that one can characterize the class of cost functions yielding unique equilibria if the domain of preferences is $\mathcal{L}^*$.

- $G(\mu^{\mathrm{SR}}, c), G(\mu^{\mathrm{DSR}}, c)$: Necessary condition for uniqueness of equilibrium is that c is strictly convex, i. e. $c'' > 0$. A sufficient condition is that c is increasing and strictly convex. Actually, [141] also shows that the conclusions for the serial rule do not change when $\mathcal{L}$ is used instead of $\mathcal{L}^*$. As will get more clear below, the serial games have unique strategic properties.
- $G(\mu^{\mathrm{P}}, c)$: The necessary and sufficient conditions are those for the serial demand game, including $c'(y) > c(y)/y$ for all $y \neq 0$. Notice, that the latter property does not pose additional restrictions on cost functions within the framework of $\mathcal{P}^1$.
- $G(\mu^{\mathrm{SS}}, c)$: Necessary condition is $c'' > 0$. In general it is hard to establish uniqueness if more than 2 players are involved.
- $G(\mu^{\mathrm{MP}}, c)$: even in 2 player games uniqueness is not guaranteed. For instance, uniqueness is guaranteed for cost functions $c(y) = y^{\alpha}$ only if $1 < \alpha \leq 3$. For $c(y) = y^4$ there are preference profiles in $\mathcal{L}^*$ such that multiple equilibria reside.

Decreasing Returns to Scale The above theorem basically shows that uniqueness of equilibrium in demand games related to $\mathcal{P}^1$ can be achieved for preferences in $\mathcal{L}^*$ if only costs are convex, i. e. the technology exhibits decreasing returns to scale. Starting point in the literature to characterize cost sharing rules in terms of their strategic properties is the seminal paper by Moulin and Shenker [91]. Their finding is that on $\mathcal{L}$ basically μ^{SR} is the only cost sharing rule of which the corresponding demand game passes the unique equilibrium test like in Theorem 12.

Call a smooth and strictly demand monotonic cost sharing rule μ *regular* if it is *anonymous*, so that the name of an agent does not have an impact on her cost share.

Theorem 13 (Moulin & Shenker [91]) *Let c be a strictly convex continuously differentiable cost function, and let μ be a regular cost sharing rule. The following statements are equivalent:*

- $\mu = \mu^{\mathrm{SR}}$,
- *for all profiles $(u_1, u_2, \dots, u_n)$ of utilities in $\mathcal{L}$, $G(\mu, c)$ has at most one Nash-equilibrium,*
- *for all profiles $(u_1, u_2, \dots, u_n)$ of utilities in $\mathcal{L}$ every Nash-equilibrium of $G(\mu, c)$ is also a strong equilibrium, i. e., no coalition can coordinate in order to improve the payoff for all its members.*

This theorem makes a strong case for the serial cost sharing rule, especially when one realizes that the serial equilibrium is the unique element surviving successive elimination of strictly dominated strategies. Then this equilibrium may naturally arise through evolutive or eductive behavior, it is a robust prediction of non-cooperative behavior. Recent experimental studies are in line with this theoretical support, see [22,108]. Proposition 1 in [141] shows how easy it is to construct preferences in $\mathcal{L}$ such that regular cost sharing rules other than μ^{SR} give rise to multiple equilibria in the corresponding demand game, even in two-agent cost sharing games.

Besides other fairness concepts in the distributive literature the most compelling is *envy-freeness*. An allocation passes the no envy test if no player prefers her own allocation less than that of other players. Formally, the definition is as follows.

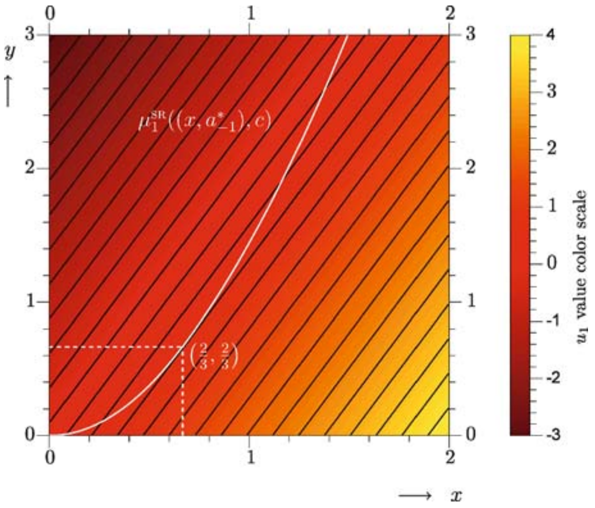
No Envy Test Let x be a demand profile and y a vector of cost shares. Then the allocation $(x_i, y_i)_{i \in N}$ is *envy-free* if for all $i, j \in N$ it holds $u_i(x_i, y_i) \geq u_i(x_j, y_j)$.

It is easily seen that the allocations associated with the serial equilibria are all envy-free.

Increasing Returns to Scale As Theorem 12 already shows, uniqueness of equilibrium in demand games for all utility profiles in $\mathcal{L}^*$ is in general inconsistent with concave cost functions.

Theorem 14 (de Frutos [26]) *Let c be a strictly concave continuously differentiable cost function, and let μ be a regular cost sharing rule. The following statements are equivalent:*

- $\mu = \mu^{\mathrm{DSR}}$ *or* $\mu = \mu^{\mathrm{SR}}$.
- *For all utility profiles $u = (u_i)_{i \in N}$ in $\mathcal{L}$ the induced demand game $G(\mu, c)$ has at most one Nash equilibrium, or*
- *For all utility profiles $u = (u_i)_{i \in N}$ in $\mathcal{L}$, every Nash equilibrium of the game $G(\mu, c)$ is a strong Nash equilibrium as well.*



Cost Sharing, Figure 10
Scenario (ii). The indifference curves of agent 1 together with the curve $\kappa : t \mapsto \mu_1^{\mathrm{SR}}((t, \bar{x}_{-1}), c)$. Best response of player 1 against $\bar{x}_{-1} = (\frac{2}{3}, \frac{8}{3})$ is the value of x where the graph of κ is tangent to an indifference curve of u_1

Moreover, if the curvature of the indifference curves is bigger than that of the curve generated by the cost sharing rule as in Fig. 10 then the second and third statement are equivalent with $\mu = \mu^{\mathrm{DSR}}$.

Theorem 15 (Moulin [85]) *Assume agents have preferences in $\mathcal{L}^*$. The serial cost sharing rule is the unique continuous, cross-monotonic and anonymous cost sharing rule for which the Nash-equilibria of the corresponding demand games all pass the no-envy test.*

Comparison of Serial and Proportional Equilibria (II)

Just as in the earlier analysis in Sect. "Efficiency and Strategy-Proof Cost Sharing Mechanisms", performance of cost sharing rules can be measured by the related surpluses in the Nash-equilibria of the corresponding demand games. Assume in this section that the preferences of the agents are quasi-linear in cost shares, and represented by functions $U_i(x_i, y_i) = u_i(x_i) - y_i$. Moreover, assume that u_i is non-decreasing and concave, $u_i(0) = 0$. Then the surplus at the demand profile x and utility profile is the number $\sum_{i \in N} u_i(x_i) - c(x(N))$. Define the *efficient surplus* or *value* of N relative to c and U by

$$v(c, U) = \sup_{x \in \mathbb{R}^N_+} \sum_{i \in N} u_i(x_i) - c(x(N)) \,. \tag{27}$$

Denote the set of Nash-equilibria in the demand game $G(\mu, c)$ with profile of preferences U by $\mathrm{NE}(\mu, c, U)$. Given c, μ the *guaranteed (relative) surplus* of the cost sharing rule μ for N is defined by

$$\gamma(c, \mu) = \inf_{U, x \in \mathrm{NE}(\mu, c, U)} \frac{\sum_{i \in N} u_i(x_i) - c(x(N))}{v(c, U)} . \quad (28)$$

Here the infimum is taken over all utility profiles discussed above. This measure is also called the *price of anarchy* of the game, see [65]. Let C^* be the set of all convex increasing cost functions with $\lim_{y \to \infty} c(y)/y = \infty$. Then [89] shows that for the serial and the proportional rule the guaranteed surplus is at least $1/n$. But sometimes the distinction is eminent. Define the number

$$\delta(y) = \frac{y c''(y)}{c'(y) - c'(0)} ,$$

which is a kind of elasticity. The below theorem shows that on certain domains of cost functions with bounded δ the serial rule prevails over the proportional rule. For large n the guaranteed surplus at μ^{SR} is of order $1/\ln(n)$, that of μ^{AV} of order $1/n$. More precise, write $K_n = 1 + \frac{1}{3} + \cdots + \frac{1}{2n-1} \approx 1 + \frac{\ln n}{2}$, then:

Theorem 16 (Moulin [89]) *For any convex increasing cost function c with* $\lim_{y \to \infty} c(y)/y = \infty$ *it holds that*

- *If c' is concave and* $\inf\{\delta(y) | y \geq 0\} = p > 0$, *then*

$$\gamma(c, \mu^{\mathrm{SR}}) \geq \frac{p}{K_n} \qquad \gamma(c, \mu^{\mathrm{AV}}) \leq \frac{4}{n+3}$$

- *If c' is convex and* $\sup\{\delta(y) | y \geq 0\} = p < \infty$ *then*

$$\gamma(c, \mu^{\mathrm{SR}}) \geq \frac{1}{2p-1} \frac{1}{K_n}$$
$$\frac{4}{n+3} \leq \gamma(c, \mu^{\mathrm{AV}}) \leq \frac{4(2p-1)}{n}$$

A Word on Strategy-Proofness in $\mathcal{P}^1$

Recall the discussion on strategyproofness in Sect. "Strategyproofness". The serial demand game has a unique strong Nash equilibrium in case costs are convex and preferences are drawn from $\mathcal{L}$. Suppose the social planner aims at designing a mechanism to implement the outcomes associated with these equilibria. [91] show an efficient way to implement this *serial choice function* by an *indirect* mechanism. It is defined through a multistage game which mimics the way the serial Nash equilibria are calculated. It is easily seen that here each agent has a unique dominant strategy, in which demands result from optimization of the true preferences. Then this gives rise to a strategyproof mechanism.

Note that the same approach can not be used for the proportional rule. The strategic properties of the proportional demand game are weaker than that of the serial demand game in several aspects. First of all, it is not hard to find preference profiles in $\mathcal{L}$ leading to multiple equilibria. Whereas uniqueness of equilibrium can be repaired by restricting $\mathcal{L}$ to $\mathcal{L}^*$, the resulting equilibria are, in general, not strong (like the serial counterparts). In the proportional equilibria there is overproduction, see e. g. the example in Sect. "Proportional Demand Game" where a small uniform reduction of demands yields higher utility for all the players. Besides, a single-valued Nash equilibrium selection corresponds to a strategyproof mechanism *provided* the underlying domain of preferences is rich, and $\mathcal{L}^*$ is not. Though richness is not a necessary condition, the proportional rule is not consistent with a strategyproof demand game.

Bayesian $\mathcal{P}^1$-Demand Games

Recall that at the basis of a strategic game there is the assumption that each player *knows* all the ingredients of the game. However, as [56] argues, production cost and output quality may vary unpredictably as a consequence of the technology and worker quality. Besides that, changes in the available resources and demands will have not foreseen influences on individual preferences. On top of that, the players may have asymmetrical information regarding the nature of uncertainty. [56] study the continuous homogeneous cost sharing problem within the context of a *Bayesian demand game* [43], where these uncertainties are taking into account. The qualifications of the serial rule in the stochastic model are roughly the same as in the deterministic framework.

Continuous Heterogeneous Output Model, $\mathcal{P}^n$

The analysis of continuous cost sharing problems for multi-service facilities are far more complex than the single-output model. The literature discusses two different models, one where each agent i demands a different good, and one where agents may require mixed bundles of goods. As the reader will notice, the modeling and analysis of solutions differs in abstraction and complexity. In order to concentrate on the main ideas, here we will stick to the first model, where goods are identified with agents. This means that a demand profile is a vector $x \in \mathbb{R}_+^N$, where x_i denotes the demand of agent i for good i. From now we deal with technologies described by continuously differentiable cost functions $c\colon \mathbb{R}^N_+ \to \mathbb{R}_+$, non-decreasing and $c(0_N) = 0$; the class of all such functions is denoted by C^n.

Extensions of Cost Sharing Rules The single-output model is connected to the multi-output model via the homogeneous cost sharing problems. Suppose that for $c \in C^n$ there is a function $c_0 \in C$ such that $c(x) = c_0(x(N))$ for all x. For instance, such functions are found if we distinguish between production of blue and red cars; the color of the car does not affect the total production costs. Essentially, a homogeneous cost sharing problem (x, c) may be solved as if it were in $\mathcal{P}^1$. If μ is the compelling solution on $\mathcal{P}^1$ then any cost sharing rule on $\mathcal{P}^n$ should determine the same solution on the class of *homogeneous* problems therein. Formally the cost sharing rule on $\mathcal{P}^n$ extends $\overline{\mu}$ on $\mathcal{P}^1$ if for all homogeneous cost sharing problems (x, c) it holds $\overline{\mu}(x, c) = \mu(x, c_0)$. In general a cost sharing rule μ on $\mathcal{P}^1$ allows for a whole class of extensions. Below we will focus on extensions of $\mu^{\text{SR}}, \mu^{\text{P}}, \mu^{\text{SS}}$.

Measurement of Scale Around the world quantities of goods are measured by several standards. Length is expressed in inches or centimeters, volume in gallons or liters, weight in ounces to kilos. Here measurement conversion involves no more than multiplication with a fixed scalar. When such linear scale conversions do not have any effect on final cost shares, a cost sharing rule is called *scale invariant*. It is an *ordinal* rule if this invariance extends to essentially all transformations of scale. Scale invariance captures the important idea that the relative cost shares should not change, whether we are dividing 1 Euro or 1,000 Euros. Ordinality may be desirable, but for many purposes too strong as a basic requirement. Formally, a *transformation of scale* is a mapping $f\colon \mathbb{R}^N_+ \to \mathbb{R}^N_+$ such that $f(x) = (f_1(x_1), f_2(x_2), \ldots, f_n(x_n))$ for all x and each of the coordinate mappings f_j is differentiable and strictly increasing.

Ordinality A cost sharing rule μ on $\mathcal{P}^n$ is *ordinal* if for all transformations of scale f and all cost sharing problems $(x, c) \in \mathcal{P}^n$, it holds that

$$\mu(x, c) = \mu(f(x), c \circ f^{-1})\,. \tag{29}$$

Scale Invariance A cost sharing rule μ on $\mathcal{P}^n$ is scale invariant if Eq. (29) holds for all linear transforms f. Under a scale invariant cost sharing rule final cost shares do not change by changing the units in which the goods are measured.

Path-Generated Cost Sharing Rules Many cost sharing rules on $\mathcal{P}^n$ calculate the cost shares for $(x, c) \in \mathcal{P}^n$ by the total of marginal costs along some *production path* from 0 toward x. Here a path for x is a non-decreasing mapping $\gamma^x\colon \mathbb{R}_+ \to \mathbb{R}^N_+$ such that $\gamma(0) = 0_N$ and there is a $T \in \mathbb{R}_+$ with $\gamma^x(T) = x$. The cost sharing rule *generated by the path collection* $\gamma = \{\gamma^x | x \in \mathbb{R}^N_+\}$ is defined by

$$\mu^\gamma_i(x, c) = \int_0^\infty \partial_i c(\gamma^x(t))(\gamma^x_i)'(t)\mathrm{d}t\,. \tag{30}$$

Special path-generated cost sharing rules are the *fixed-path cost sharing rules*; a single path $\gamma^*\colon \mathbb{R}_+ \to \mathbb{R}^N_+$ with the property that $\lim_{t\to\infty} \gamma^*_i(t) = \infty$ defines the whole underlying family of paths. More precisely, the *fixed path cost sharing rule* μ generated by γ^* is the path-generated rule for the family of paths $\{\gamma^x | x \in \mathbb{R}^N_+\}$ defined by $\gamma^x(t) = \gamma^*(t) \wedge x$, the vector with coordinates $\min\{\gamma^*_i(t), x_i\}$. So the paths are no more than the projections of $\gamma^*(t)$ on the cube $[0, x]$. Below we will see many examples of (combinations of) such fixed-path methods.

Aumann–Shapley Rule The early characterizations by [15,79] on this rule set off a vast growing literature on cost sharing models with variable demands. [16] suggested to use the Aumann–Shapley rule to determine telephone billing rates in the context of sharing the cost of a telephone system. This extension of proportional cost sharing calculates marginal costs along the path $\gamma^{\text{AS}}(t) = tx$ for $t \in [0, 1]$. Then

$$\mu^{\text{AS}}_i(x, c) = x_i \int_0^1 \partial_i c(tx)\mathrm{d}t\,. \tag{31}$$

The Aumann–Shapley rule can be interpreted as the Shapley-value of the *non-atomic* game where each unit of the good is a player, see [11]. It is the uniform average over marginal costs along all increasing paths from 0_N to x. The following is a classic result in the cost sharing literature:

Theorem 17 (Mirman & Tauman [79], Billera & Heath [15]) *There is only one additive, positive, and scale invariant cost sharing rule on $\mathcal{P}^n$ that extends the proportional rule, and this is μ^{AS}.*

Example 24 If c is positively homogeneous, i. e. $c(\alpha y) = \alpha c(y)$ for $\alpha \geq 0$ and all $y \in \mathbb{R}^N_+$, then

$$\mu^{\text{AS}}_i(x, c) = \frac{\partial_i c}{\partial x_i}(x)$$

i. e. μ^{AS} calculates the marginal costs of the ith good at the final production level x. The risk measures (cost functions) as in [28] are of this kind.

Friedman–Moulin Rule This serial extension [36] calculates marginal costs along the diagonal path, i. e. $\gamma^{\text{FM}}(t) = t\mathbf{1}_N \wedge x$

$$\mu_i^{\text{FM}}(x, c) = \int_0^{x_i} \partial_i c(\gamma^x(t))\mathrm{d}t \,. \tag{32}$$

This fixed path cost sharing rule is demand monotonic. As far as invariance with the choice of unit is concerned, the performance is bad as it is not a scale invariant cost sharing rule.

Moulin–Shenker Rule This fixed-path cost sharing rule is proposed as an ordinal serial extension by [125]. Suppose that the partial derivatives of $c \in C^n$ are bounded away from 0, i. e. there is a such that $\partial_i c(x) > a$ for all $x \in \mathbb{R}_+^N$. The Moulin–Shenker rule μ^{MS} is generated by the path γ^{MS} as solution to the system of ordinary differential equations

$$\gamma_i'(t) = \begin{cases} \frac{\sum_{j\in N}\partial_j c(\gamma(t))}{\partial_i c(\gamma(t))} & \text{if } \gamma_i(t) < x_i \,, \\ 0 & \text{else} \,. \end{cases} \tag{33}$$

The interpretation of this path is that at each moment the total expenditure for production of extra units for the different agents is equal; if good 2 is twice as expensive as good 1, then the production device γ^{MS} will produce twice as much of the good 1. The serial rule embraces the same idea – as long as an agent desires extra production, the corresponding incremental costs are equally split. This makes μ^{MS} a natural extension of the serial rule. Call t_i^* the completion time of production for agent i, i. e. $\gamma_i^{\text{MS}}(t) < x_i$ if $t < t_i^*$ and $\gamma_i^{\text{MS}}(t_i^*) = x_i$. Assume without loss of generality that these completion times are ordered such that $0 = t_0^* \le t_1^* \le t_2^* \le \dots \le t_n^*$, then the Moulin–Shenker rule is given by

$$\mu_i^{\text{MS}}(x, c) = \sum_{\ell=1}^{i} \frac{c(\gamma^{\text{MS}}(t_\ell^*)) - c(\gamma^{\text{MS}}(t_{\ell-1}^*))}{n - \ell + 1} \,. \tag{34}$$

Note that the path varies with the cost function, and that this is the reason why μ^{MS} is a non-additive solution. Such solutions – though intuitive – are in general notoriously hard to analyze. There are two axiomatic characterizations of the Moulin–Shenker rule. The first is by [125], in terms of the *serial principle* and the technical condition that a cost sharing rule be a partial differentiable functions of the demands. The other characterization by [60] is more in line with the ordinal character of this *serial extension.*

Continuity A cost sharing rule μ on $\mathcal{P}^n$ is *continuous* if $q \mapsto \mu(q, c)$ is continuous on $\mathbb{R}_+^N$ for all c.

Continuity is weaker than partial differentiability, as it requires stability of the solution with respect to small changes in the demands.

Upperbound A cost sharing rule μ satisfies *upperbound* if for all $(q, c) \in \mathcal{P}^n$, $i \in N$

$$\mu_i(q, c) \le \max_{y\in[\mathbf{0},q]} \partial_i c(y) \,.$$

An upperbound provides each agent with a conservative and rather pessimistic estimate of her cost share, based on the maximal value of the corresponding marginal cost toward the aggregate demand.

Suppose that d is a demand profile smaller than q. A reduced cost sharing problem is defined by $(q - d, c^d)$ where c^d is defined by $c^d(y) = c(y + d) - c(d)$. So c^d measures the incremental cost of production beyond the level d.

Self-consistency A cost sharing rule μ is *self-consistent* if for all cost sharing problems $(q, c) \in \mathcal{P}^n$ with $q_{N\setminus S} = \mathbf{0}_{N\setminus S}$ for some $S \subseteq N$, and $d \le q$ such that $\mu_i(d, c) = \mu_j(d, c)$ for all $\{i, j\} \subset S$, then $\mu(q, c)_S = \mu(d, c)_S + \mu(q - d, c^d)_S$.

So, self-consistency is expresses the idea that if cost shares of agents with non-zero demand differ, then this is not due to the part of the problem that they are equally charged for, but due to the asymmetries in the related reduced problem. The property is reminiscent of the *step-by-step negotiation property* in the bargaining literature, see [54].

Theorem 18 (Koster [60]) *There is only one continuous, self-consistent and scale invariant cost sharing rule satisfying upper bounds, which is the Moulin–Shenker rule.*

Shapley–Shubik Rule For each demand profile x the stand-alone cost game c_x is defined as before. Then the Shapley–Shubik rule is no more than the Shapley value of this game, i. e. $\mu^{\text{SS}}(x, c) = \Phi(c_x)$. The Shapley–Shubik rule is ordinal.

A Numerical Example Consider the cost sharing problem (x, c) with $N = \{1, 2\}, x = (5, 10)$, and $c \in C^2$ is given by $c(t_1, t_2) = \mathrm{e}^{2t_1+t_2} - 1$ on $[0, 10] \times [0, 10]$. We calculate the partial derivatives $\partial_1 c(t_1, t_2) = 2\mathrm{e}^{2t_1+t_2} = 2\partial_2 c(t_1, t_2)$ for all $(t_1, t_2) \in \mathbb{R}_+^2$. The Aumann–Shapley path is given by $\gamma(t) = (5t, 10t)$ for $t \in [0, 1]$ and

$$\mu_1^{\text{AS}}(x, c) = \int_0^1 \partial_1 c(5t, 10t) \cdot 5\mathrm{d}t$$

$$= \int_0^1 2\mathrm{e}^{20t}\cdot 5\mathrm{d}t = \tfrac{1}{2}\left(\mathrm{e}^{20}-1\right)$$

$$\mu_2^{\mathrm{AS}}(x,c) = \int_0^1 \partial_2 c(5t,10t)\cdot 10\mathrm{d}t$$

$$= \int_0^1 \mathrm{e}^{20t}\cdot 10\mathrm{d}t = \tfrac{1}{2}\left(\mathrm{e}^{20}-1\right) .$$

The Friedman–Moulin rule uses the path

$$\gamma^{\mathrm{FM}}(t) = t\mathbf{1}_N \wedge q = \begin{cases} (t,t) & \text{if } 0 \le t < 5\,, \\ (5, 5+t) & \text{if } 5 \le t\,, \end{cases}$$

and the corresponding cost shares are calculated as follows:

$$\mu_1^{\mathrm{FM}}(x,c) = \int_0^5 \partial_1 c(t,t)\mathrm{d}t$$

$$= \int_0^5 2\mathrm{e}^{3t}\mathrm{d}t = \tfrac{2}{3}\left(\mathrm{e}^{15}-1\right) ,$$

$$\mu_2^{\mathrm{FM}}(x,c) = \int_0^5 \partial_2 c(t,t)\mathrm{d}t + \int_5^{10} \partial_2 c(5,5+t)\mathrm{d}t$$

$$= \tfrac{1}{3}\left(\mathrm{e}^{15}-1\right) + \mathrm{e}^{20} - \mathrm{e}^{15} .$$

Note that both discussed cost sharing rules use one and the same path for all cost sharing problems with demand profile x. This is characteristic for *additive* cost sharing rules (see e. g. [35,42]).

Now turn to the Moulin–Shenker rule. Since $\partial_1 c = 2\partial_2 c$ everywhere on $[0,10]\times[0,10]$, according the solution γ^{MS} of Eq. (33), until one of the demands is reached, for each produced unit of good 1 two units of good 2 are produced. In particular there is a parametrization $\overline{\gamma}$ of γ^{MS} such that $\overline{\gamma}(t) = (t, 2t)$ for $0 \le t \le 5$. The corresponding cost shares are equal since $\overline{\gamma}$ reaches both coordinates of x at the same time, so $\mu_1^{\mathrm{MS}}(x,c) = \mu_2^{\mathrm{MS}}(x,c) = \frac{1}{2}c(\overline{\gamma}(5)) = \frac{1}{2}c(5,10) = \frac{1}{2}(\mathrm{e}^{20}-1)$. Now suppose that the demands are summarized by $x^* = (10,10)$. In order to calculate $\mu^{\mathrm{MS}}(x^*,c)$, notice that there is a parametrization of γ^* of the corresponding path γ^{MS} such that

$$\gamma^*(t) = \begin{cases} (t,2t) & \text{if } t \le 5\,, \\ (t,10) & \text{for } 5 < t \le 10\,, \end{cases}$$

Notice that this path extends $\overline{\gamma}$ just to complete service for agent 1, so that – like before – agent 2 only contributes while $t < 5$. Then the cost shares are given by

$$\mu_2^{\mathrm{MS}}(x^*,c) = \tfrac{1}{2}c(\gamma^*(5)) = \tfrac{1}{2}c(5,10) = \tfrac{1}{2}(\mathrm{e}^{20}-1)\,,$$

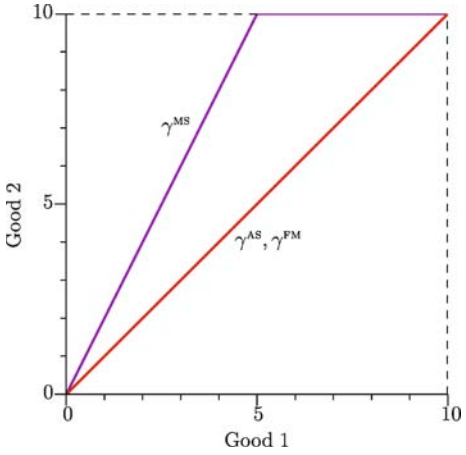
$$\mu_1^{\mathrm{MS}}(x^*,c) = \mu_2^{\mathrm{MS}}(x^*,c) + c(\gamma^*(10)) - c(\gamma^*(5))$$

$$= \mathrm{e}^{30} - \tfrac{1}{2}(\mathrm{e}^{20}+1)\,.$$

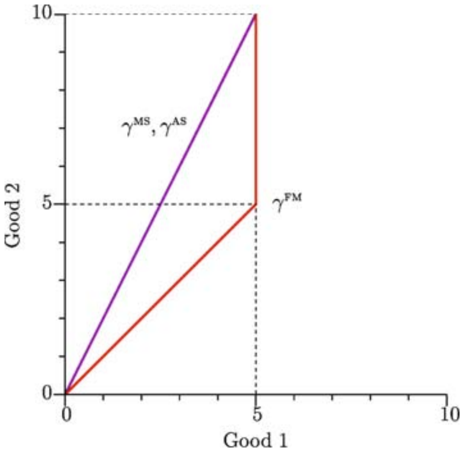
For x^* the cost sharing rules μ^{AS} and μ^{FM} use essentially the same symmetric path $\gamma(t) = (t,t)$, so that it is easily calculated that $\mu^{\mathrm{AS}}(x^*,c) = \mu^{\mathrm{FM}}(x^*,c) = (\frac{2}{3}\left(\mathrm{e}^{30}-1\right), \frac{1}{3}\left(\mathrm{e}^{30}-1\right)$.

Axiomatic Characterization of Fixed-Path Rules Recall demand monotonicity as a weak incentive constraint for cost sharing rules. Despite the attention that the Aumann–Shapley rule received, it fails to meet this standard. To see this consider the following

$$c(y) = \frac{y_1 y_2}{y_1 + y_2}\,, \quad \text{and} \quad \mu_1^{\mathrm{AS}}(x,c) = \frac{x_1 x_2^2}{(x_1+x_2)^2}\,.$$



Cost Sharing, Figure 11
Paths for $\mu^{\mathrm{MS}}, \mu^{\mathrm{AS}}, \mu^{\mathrm{FM}}$



Cost Sharing, Figure 12
Paths for $\mu^{\mathrm{MS}}, \mu^{\mathrm{AS}}, \mu^{\mathrm{FM}}$

Then the latter expression is not monotonic in x_1. One may show that the combination of properties in Theorem 17 are incompatible with demand monotonicity. Now what kind of rules are demand monotonic? The classification of all such rules is too complex. We will restrict our attention to the additive rules with the *dummy* property, which comprises the idea that a player pays nothing if her good is for free:

Dummy If $\partial_i c(y) = 0$ for all y then $\mu_i(x, c) = 0$ for all cost sharing problems $(x, c) \in \mathcal{P}^n$.

Theorem 19 (Friedman [35])

- *A cost sharing rule μ satisfies dummy, additivity and demand monotonicity if and only if it is an (infinite) convex combination of rules generated by fixed paths which do not depend on the cost structure.*
- *A cost sharing rule μ satisfies dummy, additivity and scale invariance if and only if it is an (infinite) convex combination of rules generated by scale invariant fixed paths which do not depend on the cost structure.*

This theorem has some important implications. The Friedman–Moulin rule is the unique serial extension with the properties additivity, dummy and demand monotonicity. As we mentioned before, μ^{FM} is not scale invariant. The only cost sharing rules satisfying all four of the above properties are *random order values*, i. e. a convex combination of marginal vectors of the stand-alone cost game [142]. μ^{SS} is the special element in this class of rules, giving equal weight to each marginal vector. Consider the following weak fairness property:

Equal Treatment Consider $(x, c) \in \mathcal{P}^n$. If $c_x(S \cup \{i\}) = c_x(S \cup \{j\})$ for all i, j and $S \subset N \setminus \{i, j\}$ then $\mu_i(x, c) = \mu_j(x, c)$.

Within the class of random order values, μ^{SS} is the unique cost sharing rule satisfying equal treatment.

Strategic Properties of Fixed-Path Rules [34] shows that the fixed-path cost sharing rules essentially have the same strategic properties as the serial rule. The crucial result in this respect is the following.

Theorem 20 (Friedman [34]) *Consider the demand game $G(\mu, c)$ where μ is a fixed path cost sharing rule and $c \in C^n$ has strictly increasing partial derivatives. Then the corresponding set $\mathcal{O}^\infty$ of action profiles surviving the successive elimination of overwhelmed actions consists of a unique element.*

As a corollary one may prove that the action profile in $\mathcal{O}^\infty$ is actually the unique Nash-equilibrium of the game $G(\mu, c)$, and that it is strong as well. Moreover, [34] shows that this Nash-equilibrium can be reached through some learning dynamics. Then this means that the demand games induced by μ^{FM} and μ^{MS} have strong strategic properties. Notice that the above theorem is only one-way. There are other cost sharing rules, like the *axial serial rule*, having the same strategic characteristics.

Future Directions

So far, a couple of standard stylized models have been launched providing a theoretical basis for defining and studying cost sharing principles at a basic level. The list of references below indicates that this field of research is full in swing, both in theoretical and applied directions. Although it is hard to make a guess where developments lead to, a couple of future directions will be highlighted.

Informational Issues

So far most of the literature is devoted to deterministic cost sharing problems. The cost sharing problems we face in practice are shaped by unsure events. Despite its relevance, this stochastic modeling in the literature is quite exceptional, see [56,138].

The presented models assume the information of costs for every contingent demand profile. Certainly within the continuous framework this seems too much to ask for. Retrieving the necessary information is hindered not only by technical constraints, but leads to new costs as well. [48] discusses data envelopment in cost sharing problems. A stochastic framework will be useful to study such *estimated* cost sharing problems. Other work focusing on informational coherence in cost sharing problems is [126]. Related work is [4], discussing mixtures of discrete and continuous cost sharing problems.

Budget Balance

In this overview, the proposed mechanisms are based on cost *sharing* rules. Another stream in implementation theory – at the other extreme of the spectrum – deals with cost *allocation* rules with no restrictions on the budget. [89] compares the size of budget deficits relative to the overall efficiency of a mechanism.

Performance

Recall the performance indices measuring the welfare impact of different cost sharing rules. [90] focuses on the continuous homogeneous production situations, with cost functions of specific types. There is still a need for a more general theory. In particular this could prove to be indispensable for analyzing the quality of cost sharing rules in a broader set-up, the heterogeneous and Bayesian cost sharing problems.

Non-linear Cost Sharing Rules

Most of the axiomatic literature is devoted to the analysis of cost sharing rules as linear operators. The additivity property is usually motivated as an accounting convention, but it serves merely as a tool by which some mathematical representation theorems apply. Besides the practical motivation, it is void of any ethical content. As [88] underlines, there are hardly results on non-additive cost sharing rules – one of the reasons is that the mathematical analysis becomes notoriously hard. But – as a growing number of authors acknowledges – the usefulness of these mathematical techniques alone cannot justify the contribution of the property.

I thank Hervé Moulin as a referee of this article, and his useful suggestions.

Bibliography

1. Aadland D, Kolpin V (1998) Shared irrigation cost: an empirical and axiomatical analysis. Math Soc Sci 35:203–218
2. Aadland D, Kolpin V (2004) Environmental determinants of cost sharing. J Econ Behav Organ 53:495–511
3. Albizuri MJ, Zarzuelo JM (2007) The dual serial cost-sharing rule. Math Soc Sci 53:150–163
4. Albizuri MJ, Santos JC, Zarzuelo JM (2003) On the serial cost sharing rule. Int J Game Theory 31:437–446
5. An M (1998) Logconcavity versus logconvexity, a complete characterization. J Econ Theory 80:350–369
6. Archer A, Feigenbaum J, Krishnamurthy A, Sami R, Shenker S (2004) Approximation and collusion in multicast costsharing. Games Econ Behav 47:36–71
7. Arin J, Iñarra E (2001) Egalitarian solutions in the core. Int J Game Theory 30:187–193
8. Atkinson AB (1970) On the measurement of inequality. J Econ Theory 2:244–263
9. Aumann RJ (1959) Acceptable points in general cooperative *n*-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games, vol IV. Princeton University Press, Princeton
10. Aumann RJ, Maschler M (1985) Game theoretic analysis of a bankruptcy problem from the Talmud. J Econ Theory 36:195–213
11. Aumann RJ, Shapley LS (1974) Values of Non-atomic Games. Princeton University Press, Princeton
12. Baumol W, Bradford D (1970) Optimal departure from marginal cost pricing. Am Econ Rev 60:265–283
13. Baumol W, Panzar J, Willig R (1988) Contestable Markets & the Theory of Industry Structure. 2nd edn Hartcourt Brace Jovanovich. San Diego
14. Bergantino A, Coppejans L (1997) A game theoretic approach to the allocation of joint costs in a maritime environment: A case study. Occasional Papers, vol 44. Department of Maritime Studies and International Transport, University of Wales, Cardi
15. Billera LJ, Heath DC (1982) Allocation of shared costs: A set of axioms yielding a unique procedure. Math Oper Res 7:32–39
16. Billera LJ, Heath DC, Raanan J (1978) Internal telephone billing rates: A novel application of non-atomic game theory. Oper Res 26:956–965
17. Bird CG (1976) On cost allocation for a spanning tree: A game theoretic approach. Networks 6:335–350
18. Binmore K (2007) Playing for real: A text on game theory. Oxford University Press, Oxford
19. Bjorndal E, Hamers H, Koster M (2004) Cost allocation in a bank ATM network. Math Methods Oper Res 59:405–418
20. Bondareva ON (1963) Some applications of linear programming to the theory of cooperative games. Problemy Kybernetiki 10:119–139 (in Russian)
21. Brânzei R, Ferrari G, Fragnelli V, Tijs S (2002) Two approaches to the problem of sharing delay costs in joint projects. Ann Oper Res 109:359–374
22. Chen Y (2003) An experimental study of serial and average cost pricing mechanisms. J Publ Econ 87:2305–2335
23. Clarke EH (1971) Multipart pricing of public goods. Publ Choice 11:17–33
24. Dasgupta PS, Hammond PJ, Maskin ES (1979) The implementation of social choice rules: Some general results on incentive compatibility. Rev Econ Stud 46:185–216
25. Davis M, Maschler M (1965) The kernel of a cooperative game. Nav Res Logist Q 12:223–259
26. de Frutos MA (1998) Decreasing serial cost sharing under economies of scale. J Econ Theory 79:245–275
27. Demers A, Keshav S, Shenker S (1990) Analysis and simulation of a fair queueing algorithm. J Internetworking 1:3–26
28. Denault M (2001) Coherent allocation of risk capital. J Risk 4_7–21
29. Dewan S, Mendelson H (1990) User delay costs and internal pricing for a service facility. Manag Sci 36:1502–1517
30. Dutta B, Ray D (1989) A concept of egalitarianism under participation constraints. Econometrica 57:615–635
31. Dutta B, Ray D (1991) Constrained egalitarian allocations. Games Econ Behav 3:403–422
32. Flam SD, Jourani A (2003) Strategic behavior and partial cost sharing. Games Econ Behav 43:44–56
33. Fleurbaey M, Sprumont Y (2006) Sharing the cost of a public good without subsidies. Université de Montréal, Cahier
34. Friedman E (2002) Strategic properties of heterogeneous serial cost sharing. Math Soc Sci 44:145–154
35. Friedman E (2004) Paths and consistency in additive cost sharing. Int J Game Theory 32:501–518
36. Friedman E, Moulin H (1999) Three methods to share joint costs or surplus. J Econ Theory 87:275–312
37. Friedman E, Shenker S (1998) Learning and implementation on the Internet. Working paper 199821, Rutgers University, Piscataway
38. González-Rodríguez P, Herrero C (2004) Optimal sharing of surgical costs in the presence of queues. Math Methods Oper Res 59:435–446
39. Granot D, Huberman G (1984) On the core and nucleolus of minimum cost spanning tree games. Math Program 29:323–347

40. Green J, Laffont JJ (1977) Characterization of satisfactory mechanisms for the revelation of preferences for public goods. Econometrica 45:427–438
41. Groves T (1973) Incentives in teams. Econometrica 41:617–663
42. Haimanko O (2000) Partially symmetric values. Math Oper Res 25:573–590
43. Harsanyi J (1967) Games with incomplete information played by Bayesian players. Manag Sci 14:159–182
44. Hart S, Mas-Colell A (1989) Potential, value, and consistency. Econometrica 57:589–614
45. Haviv M (2001) The Aumann–Shapley price mechanism for allocating congestion costs. Oper Res Lett 29:211–215
46. Hougaard JL, Thorlund-Petersen L (2000) The stand-alone test and decreasing serial cost sharing. Econ Theory 16:355–362
47. Hougaard JL, Thorlund-Petersen L (2001) Mixed serial cost sharing. Math Soc Sci 41:51–68
48. Hougaard JL, Tind J (2007) Cost allocation and convex data envelopment. mimeo University of Copenhagen, Copenhagen
49. Henriet D, Moulin H (1996) Traffic-based cost allocation in a network. J RAND Econ 27:332–345
50. Iñarra E, Isategui JM (1993) The Shapley value and average convex games. Int J Game Theory 22:13–29
51. Israelsen D (1980) Collectives, communes, and incentives. J Comp Econ 4:99–124
52. Jackson MO (2001) A crash course in implementation theory. Soc Choice Welf 18:655–708
53. Joskow PL (1976) Contributions of the theory of marginal cost pricing. Bell J Econ 7:197–206
54. Kalai E (1977) Proportional solutions to bargaining situations: Interpersonal utility comparisons. Econometrica 45:1623–1630
55. Kaminski M (2000) 'Hydrolic' rationing. Math Soc Sci 40:131–155
56. Kolpin V, Wilbur D (2005) Bayesian serial cost sharing. Math Soc Sci 49:201–220
57. Koster M (2002) Concave and convex serial cost sharing. In: Borm P, Peters H (eds) Chapters in game theory. Kluwer, Dordrecht
58. Koster M (2005) Sharing variable returns of cooperation. CeNDEF Working Paper 05–06, University of Amsterdam, Amsterdam
59. Koster M (2006) Heterogeneous cost sharing, the directional serial rule. Math Methods Oper Res 64:429–444
60. Koster M (2007) The Moulin–Shenker rule. Soc Choice Welf 29:271–293
61. Koster M (2007) Consistent cost sharing. mimeo, University of Amsterdam
62. Koster M, Tijs S, Borm P (1998) Serial cost sharing methods for multi-commodity situations. Math Soc Sci 36:229–242
63. Koster M, Molina E, Sprumont Y, Tijs ST (2002) Sharing the cost of a network: core and core allocations. Int J Game Theory 30:567–599
64. Koster M, Reijnierse H, Voorneveld M (2003) Voluntary contributions to multiple public projects. J Publ Econ Theory 5:25–50
65. Koutsoupias E, Papadimitriou C (1999) Worst-case equilibria. In: 16th Annual Symposium on Theoretical Aspects of Computer Science, Trier. Springer, Berlin, pp 404–413
66. Legros P (1986) Allocating joint costs by means of the Nucleolus. Int J Game Theory 15:109–119
67. Leroux J (2004) Strategy-proofness and efficiency are incompatible in production economies. Econ Lett 85:335–340
68. Leroux J (2006) Profit sharing in unique Nash equilibrium: characterization in the two-agent case, Games and Economic Behavior 62:558–572
69. Littlechild SC, Owen G (1973) A simple expression for the Shapley value in a simple case. Manag Sci 20:370–372
70. Littlechild SC, Thompson GF (1977) Aircraft landing fees: A game theory approach. Bell J Econ 8:186–204
71. Maniquet F, Sprumont Y (1999) Efficient strategy-proof allocation functions in linear production economies. Econ Theory 14:583–595
72. Maniquet F, Sprumont Y (2004) Fair production and allocation of an excludable nonrival good. Econometrica 72:627–640
73. Maschler M (1990) Consistency. In: Ichiishi T, Neyman A, Tauman Y (eds) Game theory and applications. Academic Press, New York, pp 183–186
74. Maschler M (1992) The bargaining set, kernel and nucleolus. In: Aumann RJ, Hart S (eds) Handbook of Game Theory with Economic Applications, vol I. North–Holland, Amsterdam
75. Maschler M, Reijnierse H, Potters J (1996) Monotonicity properties of the nucleolus of standard tree games. Report 9556, Department of Mathematics, University of Nijmegen, Nijmegen
76. Maskin E, Sjöström T (2002) Implementation theory. In: Arrow KJ, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare, vol I. North–Holland, Amsterdam
77. Matsubayashi N, Umezawa M, Masuda Y, Nishino H (2005) Cost allocation problem arising in hub-spoke network systems. Europ J Oper Res 160:821–838
78. McLean RP, Pazgal A, Sharkey WW (2004) Potential, consistency, and cost allocation prices. Math Oper Res 29:602–623
79. Mirman L, Tauman Y (1982) Demand compatible equitable cost sharing prices. Math Oper Res 7:40–56
80. Monderer D, Shapley LS (1996) Potential games. Games Econ Behav 14:124–143
81. Moulin H (1987) Equal or proportional division of a surplus, and other methods. Int J Game Theory 16:161–186
82. Moulin H (1994) Serial cost-sharing of an excludable public good. Rev Econ Stud 61:305–325
83. Moulin H (1995) Cooperative microeconomics: A game–theoretic introduction. Prentice Hall, London
84. Moulin H (1995) On additive methods to share joint costs. Japan Econ Rev 46:303–332
85. Moulin H (1996) Cost sharing under increasing returns: A comparison of simple mechanisms. Games Econ Behav 13:225–251
86. Moulin H (1999) Incremental cost sharing: Characterization by coalition strategy-proofness. Soc Choice Welf 16:279–320
87. Moulin H (2000) Priority rules and other asymmetric rationing methods. Econometrica 68:643–684
88. Moulin H (2002) Axiomatic cost and surplus-sharing. In: Arrow KJ, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare. Handbooks in Economics, vol 19. North–Holland, Amsterdam, pp 289–357
89. Moulin H (2006) Efficient cost sharing with cheap residual claimant. mimeo Rice University, Housten

90. Moulin H (2007) The price of anarchy of serial, average and incremental cost sharing. mimeo Rice University, Housten
91. Moulin H, Shenker S (1992) Serial cost sharing. Econometrica 60:1009–1037
92. Moulin H, Shenker S (1994) Average cost pricing versus serial cost sharing; an axiomatic comparison. J Econ Theory 64:178–201
93. Moulin H, Shenker S (2001) Strategy-proof sharing of submodular cost: budget balance versus efficiency. Econ Theory 18:511–533
94. Moulin H, Sprumont Y (2005) On demand responsiveness in additive cost sharing. J Econ Theory 125:1–35
95. Moulin H, Sprumont Y (2006) Responsibility and cross-subsidization in cost sharing. Games Econ Behav 55:152–188
96. Moulin H, Vohra R (2003) Characterization of additive cost sharing methods. Econ Lett 80:399–407
97. Moulin H, Watts A (1997) Two versions of the tragedy of the commons. Econ Des 2:399–421
98. Mutuswami S (2004) Strategy proof cost sharing of a binary good and the egalitarian solution. Math Soc Sci 48:271–280
99. Myerson RR (1980) Conference structures and fair allocation rules. Int J Game Theory 9:169–182
100. Myerson RR (1991) Game theory: Analysis of conflict. Harvard University Press, Cambridge
101. Nash JF (1950) Equilibrium points in *n*-person games. Proc Natl Acad Sci 36:48–49
102. O'Neill B (1982) A problem of rights arbitration from the Talmud. Math Soc Sci 2:345–371
103. Osborne MJ (2004) An introduction to game theory. Oxford University Press, New York
104. Osborne MJ, Rubinstein A (1994) A Course in Game Theory. MIT Press, Cambridge
105. Peleg B, Sudhölter P (2004) Introduction to the theory of cooperative games, Series C: Theory and Decision Library Series. Kluwer, Amsterdam
106. Pérez-Castrillo D, Wettstein D (2006) An ordinal Shapley value for economic environments. J Econ Theory 127:296–308
107. Potters J, Sudhölter P (1999) Airport problems and consistent allocation rules. Math Soc Sci 38:83–102
108. Razzolini L, Reksulak M, Dorsey R (2004) An experimental evaluation of the serial cost sharing rule. Working paper 0402, VCU School of Business, Dept. of Economics, Richmond
109. Ritzberger K (2002) Foundations of non-cooperative game theory. Oxford University Press, Oxford
110. Rosenthal RW (1973) A class of games possessing pure-strategy Nash equilibria. J Econ Theory 2:65–67
111. Roth AE (ed) (1988) The Shapley value, essays in honor of Lloyd S Shapley. Cambridge University Press, Cambridge, pp 307–319
112. Samet D, Tauman Y, Zang I (1984) An application of the Aumann–Shapley prices for cost allocation in transportation problems. Math Oper Res 9:25–42
113. Sánchez SF (1997) Balanced contributions axiom in the solution of cooperative games. Games Econ Behav 20:161–168
114. Sandsmark M (1999) Production games under uncertainty. Comput Econ 14:237–253
115. Schmeidler D (1969) The Nucleolus of a characteristic function game. SIAM J Appl Math 17:1163–1170
116. Shapley LS (1953) A value for n-person games. Ann Math Study, vol 28. Princeton University Press, Princeton, pp 307–317
117. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 14:453–460
118. Shapley LS (1969) Utility comparison and the theory of games. In: Guilbaud GT (ed) La decision: Aggregation et dynamique des ordres de preference. Editions du Centre National de la Recherche Scientifique. Paris pp 251–263
119. Shapley LS (1971) Cores of convex games. Int J Game Theory 1:1–26
120. Sharkey W (1982) Suggestions for a game–theoretic approach to public utility pricing and cost allocation. Bell J Econ 13:57–68
121. Sharkey W (1995) Network models in economics. In: Ball MO et al (eds) Network routing. Handbook in Operations Research and Management Science, vol 8. North–Holland, Amsterdam
122. Shubik M (1962) Incentives, decentralized control, the assignment of joint cost, and internal pricing. Manag Sci 8:325–343
123. Skorin-Kapov D (2001) On cost allocation in hub-like networks. Ann Oper Res 106:63–78
124. Skorin-Kapov D, Skorin–Kapov J (2005) Threshold based discounting network: The cost allocation provided by the nucleolus. Eur J Oper Res 166:154–159
125. Sprumont Y (1998) Ordinal cost sharing. J Econ Theory 81:126–162
126. Sprumont Y (2000) Coherent cost sharing. Games Econ Behav 33:126–144
127. Sprumont Y (2005) On the discrete version of the Aumann–Shapley cost-sharing method. Econometrica 73:1693–1712
128. Sprumont Y, Ambec S (2002) Sharing a river. J Econ Theory 107:453–462
129. Sprumont Y, Moulin H (2005) Fair allocation of production externalities: Recent results. CIREQ working paper 28-2005, Montreal
130. Sudhölter P (1998) Axiomatizations of game theoretical solutions for one-output cost sharing problems. Games Econ Behav 24:42–71
131. Suijs J, Borm P, Hamers H, Koster M, Quant M (2005) Communication and cooperation in public network situations. Ann Oper Res 137:117–140
132. Tauman Y (1988) The Aumann–Shapley prices: A survey. In: Roth A (ed) The shapley value. Cambridge University Press, Cambridge, pp 279–304
133. Thomas LC (1992) Dividing credit-card costs fairly. IMA J Math Appl Bus & Ind 4:19–33
134. Thomson W (1996) Consistent allocation rules. Mimeo, Economics Department, University of Rochester, Rochester
135. Thomson W (2001) On the axiomatic method and its recent applications to game theory and resource allocation. Soc Choice Welf 18:327–386
136. Tijs SH, Driessen TSH (1986) Game theory and cost allocation problems. Manag Sci 32:1015–1028
137. Tijs SH, Koster M (1998) General aggregation of demand and cost sharing methods. Ann Oper Res 84:137–164
138. Timmer J, Borm P, Tijs S (2003) On three Shapley-like solutions for cooperative games with random payoffs. Int J Game Theory 32:595–613
139. van de Nouweland A, Tijs SH (1995) Cores and related solution concepts for multi-choice games. Math Methods Oper Res 41:289–311

140. von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
141. Watts A (2002) Uniqueness of equilibrium in cost sharing games. J Math Econ 37:47–70
142. Weber RJ (1988) Probabilistic values for games. In: Roth AE (ed) The Shapley value. Cambridge University Press, Cambridge
143. Young HP (1985) Producer incentives in cost allocation. Econometrica 53:757–765
144. Young HP (1985) Monotonic solutions of cooperative games. Int J Game Theory 14:65–72
145. Young HP (1985) Cost allocation: Methods, principles, applications. North-Holland, Amsterdam
146. Young HP (1988) Distributive justice in taxation. J Econ Theory 44:321–335
147. Young HP (1994) Cost allocation. In: Aumann RJ, Hart S (eds) Handbook of Game Theory, vol II. Elsevier, Amsterdam, pp 1193–1235
148. Young HP (1998) Cost allocation, demand revelation, and core implementation. Math Soc Sci 36:213–229

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of

Rowena Lohman
Cornell University, Ithaca, USA

Article Outline

Glossary

Aseismic Occurring without detectable radiated seismic energy.

Cascadia The region of the Pacific Northwest dominated by the Cascade Range and affected by subduction of the Juan de Fuca plate beneath North America.

Coseismic Occurring during an earthquake.

Elastic A form of behavior of a solid when subjected to stress. Elastic solids deform in response to stress by an amount proportional to a constant known as the "rigidity". When any applied stress is removed, an elastic solid recovers its original shape.

Geodesy The study of the shape and area of the Earth, including large-scale variations that affect the rotation dynamics of the planet of the whole, down to smaller length scales of earthquakes, landslides, etc.

Forward model A description of what a model of some process would predict about behavior of the system, e. g., how a given distribution of subsurface slip on a fault during an earthquake should affect observations of ground deformation at the surface.

GPS Global Positioning System. A network of satellites that transmit a signal that can be used by receivers (small transportable and/or permanent affixed to the ground) to infer three-dimensional positions.

InSAR Interferometric Synthetic Aperture Radar. The combination of Synthetic Aperture Radar imagery (generally acquired from airborne or satellite-based platforms) to infer changes in ground deformation, digital elevation models, variations in atmospheric water vapor, etc.

Interseismic deformation Occurs in the time period between earthquakes, usually associated with gradual increase in elastic stress to be released in future earthquakes.

Inverse theory The approach to determining the values for parameters of a given physical model that best describe observations of the system of interest.

Leveling The field of geodesy involved in the determination of variations in angle from horizontal between nearby fixed points on the Earth's surface, usually converted to changes in elevations.

Locked zone The portion of the fault zone that does not slip during the interseismic period, therefore accumulating stress and eventually rupturing coseismically.

Paleoseismology The study of individual earthquakes that occurred in the past, usually before the advent of instrumental recordings of seismic events.

Plate tectonics The theory governing how discrete plates on the Earth's surface move relative to each other over geologic time.

Postseismic deformation Deformation occurring in the hours to years following an earthquake.

Seismic cycle The combination of strain build-up and release that occurs on plate margins and along faults within plates, accommodated by processes within the coseismic, postseismic and interseismic time scales.

Seismogenic The region of a fault zone that is capable of producing earthquakes. Also refers to effects caused by an earthquake.

Subduction The process by which one tectonic plate descends beneath another, usually accompanied by volcanism and seismicity.

Triangulation The field of geodesy related to measuring horizontal angles and changes in angles between networks of fixed points.

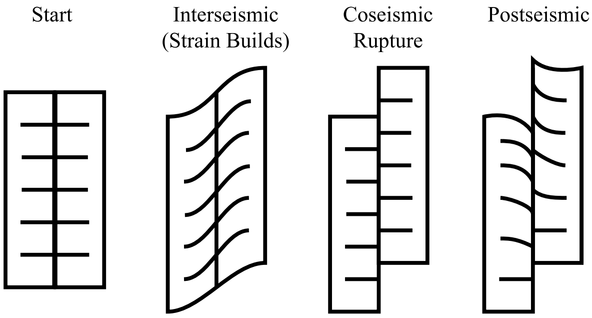
Trilateration The field of geodesy related to measuring distance and changes in distance between networks of fixed points.

Viscoelastic A material behavior that is a combination of viscous and elastic behavior, resulting in some permanent deformation when the material is subjected to changes in stress.

Viscosity A material property describing its ability to flow in response to an applied stress. A measure of the response of a material to a stress, resulting in permanent deformation. The deformation rate of a viscous material depends on both the viscosity and applied stress.

Definition of the Subject

The seismic cycle consists of the processes associated with the accumulation and release of stress on seismogenic faults. The cycle is commonly divided into 3 periods: the coseismic interval for events occurring during an earthquake, the postseismic interval after an earthquake, and the interseismic period in between earthquakes (Fig. 1). Some of this deformation during these different periods is related directly to the motion on the fault during the earthquake – the ground is translated in one direction or another, there is crushing of rock, rotation or heaving of blocks of earth, and landslides triggered by the shaking. There is also ground deformation that results because of secondary effects – movement of ground water within the crushed and strained rock, cascades of earthquakes triggered by stress changes during the first event, continued slip on the fault interface, as well as flow of deeper, more ductile layers of the crust and mantle in response to the changes in stress. By examining these different behaviors and studying crustal deformation we can learn more about the underlying cause.



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 1
Cartoon illustrating map view (view from overhead) of ground deformation during the seismic cycle for a strike-slip fault (after [73]). Similar models exist for other fault types

Earthquakes and other motion along fault zones are some of the ways that the Earth's crust accommodates far-field forcings due to plate tectonic motions. One of the prime features of interest as we study the seismic cycle is the magnitude of motion along these fault zones at different temporal and length scales. However, the only place where we can directly observe this motion is within the shallowest (< several meters) parts of the fault zone, through maps of features that are offset across the fault. The field of paleoseismology involves the search for information about previous earthquakes on a fault zone, often through trenches dug across the fault and the use of radiogenic dating to determine how frequently earthquakes have occurred in the past.

In order to infer what is occurring throughout the whole seismogenic zone (often the upper ∼15 km within the continental crust), we rely on an arsenal of tools that include seismology (the study of how seismic waves travel through the earth) and geodesy (the study of changes to the shape of the earth's surface). In this chapter, we will explore how we can draw conclusions about fault zone slip at depths far greater than are directly accessible to us, based on how the earth's surface deforms during, before and after earthquakes.

Introduction

Ground shaking due to earthquakes can be felt over most of the Earths' surface, both on land and underwater. Earthquakes are also associated with volcanic activity, landslides and tsunami – they are often concentrated near discrete tectonic plate boundaries but also occupy diffuse zones where the Earth's crust is deforming [12,57]. The seismic energy that is released during an earthquake passes through deep portions of the Earth and helps us to learn about material properties we could not illuminate any other way, including details of the structure of the Earth's crust, core and everything in between. However, the destructive cost of large earthquake requires that the primary goal of earthquake research is that of determining when and where damaging earthquakes will occur [99]. To achieve this goal, we first have to understand what happens during each earthquake, i. e., where it was located, how big it was, how it's location relates to the distribution of previous earthquakes, etc.

One of the most common methods that we use to study earthquakes relies on the feature most apparent to humans – the rapid and often destructive movements of the ground that occur during the earthquake. Seismology studies how the ground shakes during earthquakes and how we can use that information to better understand seismogenesis. However, this ground shaking is also always accompanied by some amount of permanent deformation of the ground. After all, the primary driving force behind most earthquakes is the slow motion of tectonic plates relative to each other. In this chapter we will not cover seismology, although combinations of seismic data and observations of geodetic displacements are often very powerful tools. Instead, we cover the geodetic observations of ground deformation during earthquakes and some of the methods that we use to interpret this deformation in terms of what actually happened on and around the fault zone.

Highlights of Earthquake Geodesy

Geodesy is defined as the branch of mathematics concerned with the shape and area of the Earth. In this chapter we will examine how the use of geodesy to quantify changes in the shape of the Earth can help us learn about earthquakes. We begin with a brief history of how the use of geodesy to study earthquakes has evolved, along with a description of some of the more interesting individual earthquakes and other seismic cycle behaviors that have been observed geodetically. We then explain the mechanics of how we use these measurements of surface deformation to understand processes deep beneath the Earth's surface.

Observations of Coseismic Displacements

The history of earthquake geodesy begins even before we really understood what earthquakes were. The key observations began when scientists began to associate earthquakes with observable deformation of the terrain. In Charles Lyell's Principles of Geology [46], he noted that earthquakes often accompany abrupt changes in the ground surface. The 1819 Ran of Cutch, India [64] and 1855 Wairarapa, New Zealand earthquakes were some of the first events where the accompanying ground deformation was observed. In 1835, during Charles Darwin's voyage on the *H.M.S. Beagle*, he experienced a large earthquake near Concepcion, Chile. During his reconnaissance of the area, he noted that the coastline had risen several meters in areas, exposing barnacles that had previously been underwater. He also found fossils hundreds of meters above sea level, indicating that numerous earthquakes had raised the cliffs over many millennia.

Several large earthquakes in the second half of the 19th century helped to advance the theory that earthquakes were caused by motion on faults. The 1872 Owens Valley, California, earthquake [19], the 1888 Amuri/Marlborough, New Zealand, earthquake [50], the 1891 Nobi, Japan, earthquake [36] and the 1893 Baluchistan earthquake [21] were each accompanied by visible deformation of the ground surface. These observations, and the fact that much of the deformation was consistent with regional, long-term topographic relief, helped counter arguments that earthquakes were primarily caused by volcanic activity [48].

Around the same time interval, the practice of surveying was coming into more widespread use, notably in India where Britain was involved in mapping the areas under its control. Surveyors installed permanent "monuments" or benchmarks, which they could then return to and re-survey at a later date. The May 17, 1892 Tapanuli, Sumatra, earthquake occurred during a triangulation survey by JJA Muller [58]. They noted changes in their surveyed angles between benchmarks that were consistent with two meters of deformation on a branch of the great Sumatran fault. In the foothills of the Himalaya, triangulation/leveling surveys measured deformation due to the 1897 Assam earthquake [64] and the 1905 Kangra earthquake [53]. In Italy, the 1915 Avezzano earthquake [94] was also spanned by early geodetic surveys.

In North America, the first great earthquake to be surveyed was the 1906 earthquake that destroyed most of the city of San Francisco, CA [40]. HF Reid [72] used three sets of triangulation surveys across San Andreas Fault to show that there had been approximately 3.2 meters of slip across the fault. During the 1920s and 1930s, there was a great deal of leveling work done in Japan which captured deformation associated with the 1927 Tango earthquake, and the 1944 and 1946 earthquakes associated with oceanic plate subduction in southwestern Japan [63,93]. Tide gauges along the coasts have also proved useful in earthquake studies, especially near subduction zones.

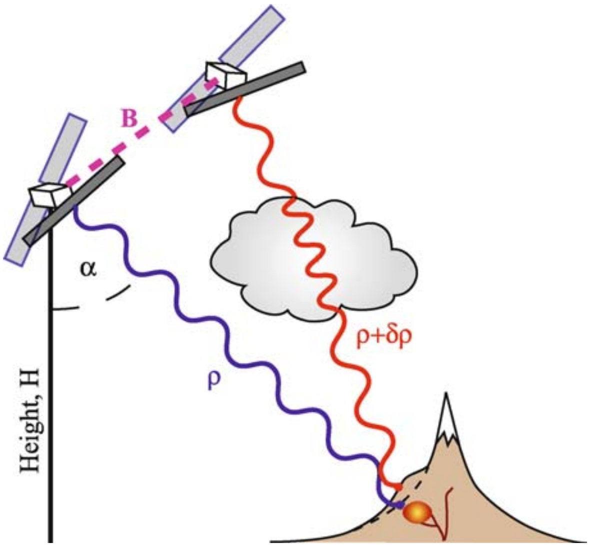
There was a boom in the use of the three main ground-based geodetic techniques (triangulation, trilateration and leveling) in the middle of the twentieth century, with observations of the 1940 Imperial Valley and 1962 Tehachapi earthquakes in California and the first observation of shallow interseismic creep on faults in the San Francisco Bay Area and in the Salton Trough/Imperial Valley region in Southern California. Next, the development of space-based geodesy began to allow for more precise and spatially extensive surveys of areas before and after earthquakes. Very Long Baseline Interferometry (VLBI) stations scattered around the world placed strong constraints on the relative motion of individual tectonic plates, as did the widespread use of Global Positioning System (GPS) [22,83].

GPS can be used either in continuous mode at permanent stations, which allows for high precision observations that can catch temporal changes in deformation, or in survey/campaign mode, where researchers take GPS receivers out to fixed benchmark stations in the field (sometimes previously studied using earlier surveying methods). Some of the first large earthquakes to be studied using GPS were the Ms 6.6 Superstition Hills, CA, earthquake in 1987 [39] and the 1989 Mw 7.1 Loma Prieta, CA earthquake [3]. The precise locations and descriptions of ground deformation allowed researchers to begin to examine how individual earthquakes fit in with the long-term topography, including the estimation of possible recurrence intervals [35,41].

InSAR The advent of Interferometric Synthetic Aperture Radar (InSAR), and it's application to the 1992 Landers, California, earthquake [17,49], ushered in a new era of geodesy with its unparalleled spatial density of deformation observations [23,76]. Synthetic Aperture Radar (SAR) images are acquired from a variety of platforms that send out radar signals, including satellites and airplanes (Fig. 2), and contain information about the phase and amplitude of the reflected return from the earth's surface.

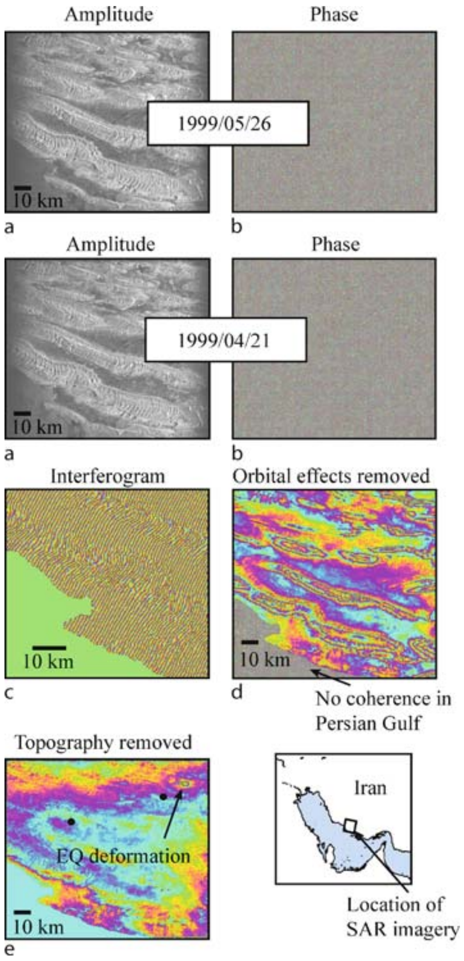
A SAR interferogram is the difference in phase between two SAR images, and reflects a variety of factors that change the path length between the satellite and the ground. After correcting for some of these factors, such as topography (Fig. 3), the resulting interferometric phase is a function of any ground deformation that occurred between the two SAR acquisitions, in addition to variations in atmospheric water vapor, the ionosphere, etc. [11,20,100]. The high spatial resolution of InSAR (pixels commonly 5 m × 20 m or smaller) combined with its large areal coverage (~100 km) produces images with so many pixels that they can become unwieldy to deal with computationally, especially when multiple interferograms are studied at once. Most researchers use various downsampling methods to reduce the number of data points considered without significantly reducing the amount of information retained [43].

The interferometric phase is only sensitive to ground deformation towards or away from the satellite, in a direction known as the satellite line of sight (LOS). Also, since the interferometric phase can only be measured as a fraction of the radar wavelength, an interferogram does not immediately give us an absolute measurement of the magnitude of ground deformation (Fig. 4). We need to add up the interferometric "fringes" to solve for the total deformation field. This "unwrapped" deformation field, along with information about the line-of-sight direction, can be combined in attempts to model the earthquake source parameters, or for models of deformation throughout other parts of the seismic cycle.



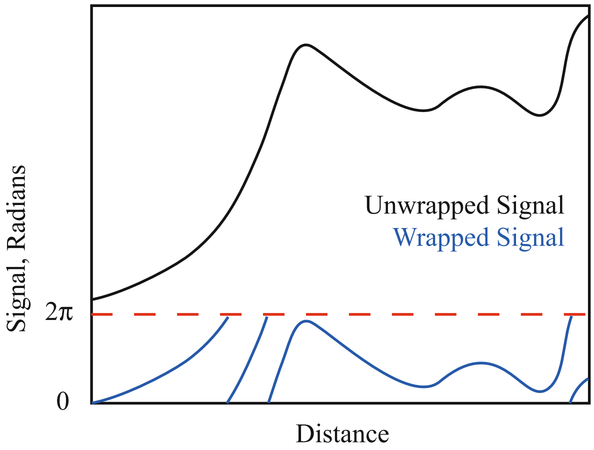
Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 2
Formation of a Synthetic Aperture Radar (SAR) interferogram: Two satellite radar images of the ground surface, separated by a spatial baseline, *B*, are combined to solve for the $\delta\rho$, the difference in line length (ρ) between the satellite and the ground along the satellite viewing direction (α). Knowledge of the satellite orbital path and the relative distance between the two image acquisitions (*B*) is necessary to convert $\delta\rho$ to elevation. If the two images are obtained simultaneously, or over a small time interval, the interferogram will only reflect topographic relief throughout the imaged area. However, if changes in ground surface elevation (e. g., the volcanic inflation shown in this figure between the *blue* and *red*), or changes in the atmosphere/ionosphere (e. g., water vapor content) occur between the two image acquisitions, this will also be reflected in the final interferogram

InSAR observations have improved the spatial complexity of these observations, allowing over a dozen large earthquakes to be studied in detail by a wide range of observers. Some noteworthy events that had various portions of their coseismic-postseismic activity covered by combinations of InSAR and GPS include the 1992 Landers, CA, earthquake [13,17,49], the 1995 Antofagasta, Chile, earthquake [70,77], the 1997 Hector Mine, CA earthquake [85], the 1999 Chi-Chi, Taiwan earthquake [8,98], the 2003 Bam, Iran earthquake [18], and the 2005 Nias–Simeulue earthquake [28]. In other cases, InSAR can help to determine the location of small- to moderate-sized earthquakes in areas with little other geophysical or field-based information [44].



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 3

Overview of steps in forming and interpreting an interferogram. An interferogram requires the combination of two sets of amplitude a and phase b observations from two separate image acquisitions (here, on two different dates separated by a month over an area along the Persian Gulf, Iran). While phase in the individual SAR images appear to be random noise, when the two images are combined, the phase changes vary coherently to form the interferogram in c. If we remove the effects of the satellite orbital geometry (the "curved earth effect"), we are left with d, which reflects both topography, ground surface deformation and any atmospheric changes between the two image acquisitions. The area in the *lower left* still looks like white noise due to the fact that water (here, the Persian Gulf) changes its reflectivity over short time scales and the phase does not remain coherent. If we remove the effects of topography e we are left with a map that clearly shows the effects of a small Mw 5 earthquake that occurred during this time interval (*black dots* show the seismically-determined locations for this event). Other features are due primarily to changes in atmospheric water vapor



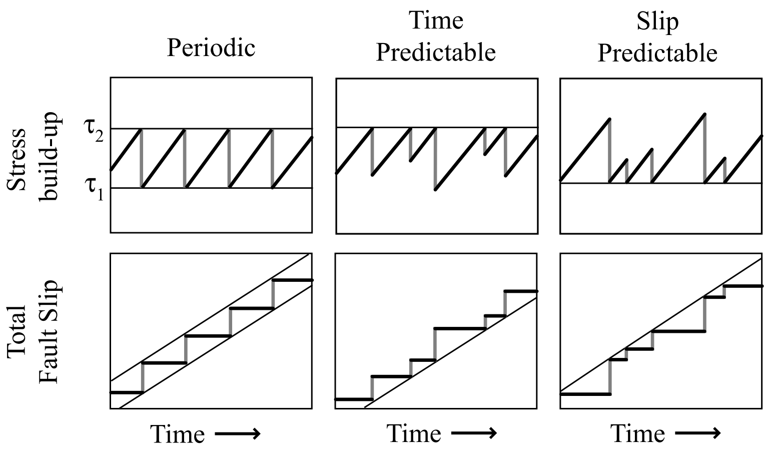
Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 4

Illustration of how the inherent 2π ambiguity affects the observed deformation. The "true" ground deformation vs. distance profile (*black* signal) would appear as the segmented *blue curve* ("*wrapped*" signal) when viewed using InSAR, since interferometry only retains information about the relative phase within a 2π cycle, not about the absolute phase, or number of cycles between the satellite and the ground. Reconstructing the original deformation field requires a process known as phase unwrapping, and there will also be an ambiguity as to the absolute value of the deformation field as a whole

Observations Throughout the Seismic Cycle

Early Models As the body of information about earthquakes grew, scientists were able to begin hypothesizing about the processes controlling their occurrence. Reid's observations of the 1906 San Francisco earthquake led him to propose his elastic rebound theory [73], where he hypothesized that the crust behaves like an elastic solid driven from the far field at a constant rate, which ruptures in earthquakes at periodic intervals to allow the two blocks to slide past each other (Fig. 1). In the interseismic period, the ground deforms smoothly in a manner that depends on the relative plate velocities, the thickness of the elastic plate, the elastic plate rigidity, etc. The size of the largest potential earthquake on a fault would depend on the length and depth of the elastic zone, and the timing until the next earthquake would depend on how much strain had built up since the last one.

This simple model, where each earthquake releases all of the built-up stress along the fault zone at a regular interval, does not seem to hold in the real world. On most faults, the magnitude of fault slip and rupture area appears to vary significantly between earthquakes (Fig. 5). Additionally, large amounts of accelerated deformation and other types of postseismic behavior are observed to



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 5

Cartoon illustrating theories governing the temporal distribution of large earthquakes [88], each assuming a constant increase in stress vs. time in the interseismic interval (*black lines*), and coseismic behavior (*gray*) that depends on the accumulated stress in various ways. In the periodic model [73], earthquakes occur at constant intervals and always have the same magnitude, releasing the same amount of built-up stress. In the time-predictable model, earthquakes always occur when a maximum stress, τ_2, is reached, but the slip in each event varies. The time of the next earthquake depends on how much stress was released in the previous event. In the slip-predictable model, earthquakes occur at varying times, but always release the amount of built-up stress down to the minimum level, τ_1. The amount of slip, therefore, depends on the time since the previous earthquake

occur immediately after earthquakes, indicating that stress release is not accommodated in a simple manner. Observations such as these can help us improve and expand on the models that we utilize to explain earthquake occurrence.

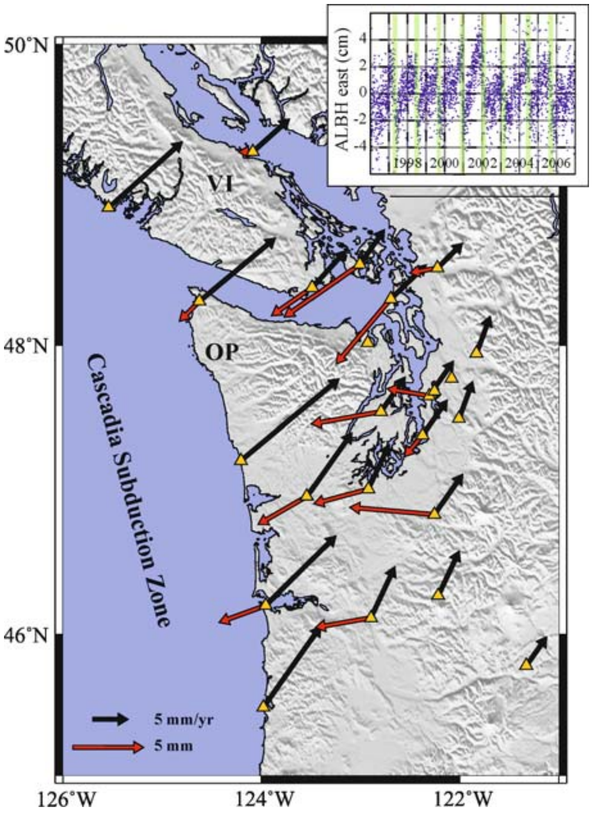
Postseismic Behavior Earthquakes may primarily release built-up strain due to plate motions, but they also produce stress increases in both the near-field within the crust and within the underlying mantle. Observations of postseismic deformation, which is driven by the preceding coseismic stress changes [25,79,92], span a wide range of behaviors that may be explained by equally large range of constitutive properties. For some shallow strike-slip earthquakes, the observed postseismic deformation is as large as the fault slip during the earthquake [38,86].

Observed postseismic behaviors include poroelastic deformation [13,33], where fluid flow following gradients in coseismic stress changes results in ground deformation, frictional afterslip [5,28,55,81,82] and viscoelastic relaxation of the lower crust [25,26,68,74,79]. Afterslip is sometimes triggered not only on the fault that caused the earthquake, but also on surrounding faults within the stress field [1,12]. Another very noticeable consequence of an earthquake is the series of aftershocks following it, which may be triggered by combinations of stress changes as the lower crust relaxes after an earthquake as well as those due to motion during the earthquake itself [90].

In cases where the coseismic slip distribution (and it's associated stress change) is well-constrained by combinations of geodetic and seismic data, we can explore models of crustal and mantle properties, or place bounds on laboratory-derived rock mechanics laws [7], in order to fit the postseismic response to the coseismic stress change. For instance, the distribution of afterslip places constraints on the frictional behavior of the fault zone [32,55]. However, any such study requires good understand of the processes occurring during the earthquake itself.

Interseismic Behavior Although it is usually less dramatic than deformation occurring during and immediately after earthquakes, interseismic deformation can also tell us a great deal about the fault zone. The depth of the "locked" or "coupled" zone that will eventually rupture seismically [24,29], the rate at which stress is accumulating along the fault zone [59,92], and even variations in crustal elastic properties when rocks of different types are brought into contact across the fault zone [14], can all be addressed by examination of interseismic deformation across a fault. In addition, data types that span a finite amount of time, such as InSAR or campaign GPS observations, will always contain some amount of interseismic strain that may need to be removed before studying measurements spanning an earthquake. The steadiest interseismic deformation is mainly due to flow of the lower crust and mantle beneath the elastic upper crust, but there are also observations of steady creep in the shallowest portions of several strike-slip fault zones [31,37,47,73].

Transient Behavior One of the most intriguing fault behaviors observed recently are isolated deformation events that are often not directly associated with an earthquake at all. These "slow" or "silent" earthquakes have been observed around the world in subduction zones, especially within the Cascadia subduction zone ([51,54], Fig. 6), Japan [27,56,65], Mexico [45], and Chile [64]. Along-strike variations in the amount of plate convergence that is taken up by these aseismic slip processes vs. large coseismic events may persist over very long timescales, as is indicated by the correlation of potentially aseismic "seismic gaps" in subduction zones and gravity anomalies [89,95]. Although most transient fault zone activity has been noticed along subduction zones, there are also observations of transient aseismic creep along the San Andreas Fault [60] and within the Imperial Valley, CA [42]. Further observations will help us understand how great of a contribution combinations of steady interseismic and transient aseismic behaviors make to the release of plate motion across many of the major tectonic boundaries around the world.



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 6

Episodic slow slip events in Cascadia. OP=Olympic Peninsula, VI=Vancouver Island. *Black arrows* indicate interseismic convergence, *red arrows* indicate motion during one of the slow slip events recorded by continuous GPS (*yellow triangles*). *Inset* shows detrended component of deformation recorded by stations ALBH on Vancouver Island, with more than a decade of regular slow slip events (*vertical green bars*). Figure after [54]

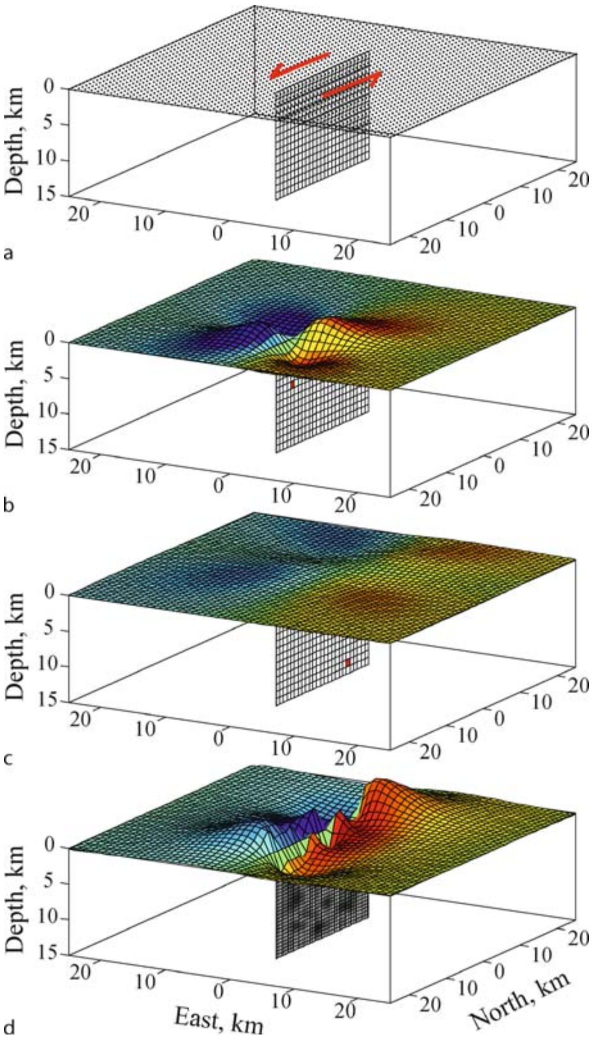
Modeling of Geodetic Observations

Overview of Inverse Theory

All of the observations discussed above involve measurements of surface deformation. We can learn nothing about processes occurring below the surface, within fault zones, without the ability to infer how those observations were generated. Inverse theory spans the range of problems relating to identifying which model of real-world behavior is consistent with observed data. Some of the simplest inverse problems include that of fitting a line to a series of data points, or even calculating the mean value of a set of observations. Good treatments of the field of inverse theory and it's applications to geophysical problems in general can be found in books by Parker [66] and Menke [52], as well as in key papers by Backus and Gilbert [4] and Tarantola and Valette [91].

Before a geophysical inversion can be performed (i. e., determining the size of an earthquake based on ground deformation) the "forward model" must be defined and understood. A forward model is a description of how we believe the system of interest, here, the Earth's crust, will behave in response to the process we are studying (i. e., fault slip). For inversions for fault slip, the forward model is our best guess at how the deformation field produced by motion on the fault surface will propagate through the solid earth to the ground surface where we can observe it. Laboratory, seismological and field observations support the idea that the crust behaves as an elastic solid on the short timescales associated with earthquakes and their immediate aftermath, although parts of the crust deform viscously at longer time scales under prolonged stress. The response of the elastic crust surrounding a plane that undergoes fault slip can be described mathematically [62]. Figure 7 shows the predicted ground deformation for strike-slip motion on fault "patches" at varying depths. The ground deformation response is linear, i. e., the deformation from multiple sources is just the sum of deformation from the individual fault patches. Therefore, we can predict the expected ground deformation from arbitrarily complicated fault zone geometries and slip distributions.

In an inverse problem, we consider the behavior of a forward model or family of forward models from a num-



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 7

Predicted ground surface deformation for a vertical, left-lateral strike-slip fault a, using Okada [62]. Fault slip on a shallow fault patch (*red*, panel b) produces a larger amount of deformation over a smaller area than does fault slip on a deeper region of the fault plane c. Here, *deformation of the grid* indicates vertical deformation and *color* indicates the amount of northward (*red*) or southward (*blue*) deformation. The amount of total deformation from a complicated fault slip distribution is achieved by summing the components from all parts of the fault d

ber of possible sources that could potentially explain our data, and we find the combination these sources that best fits the data. In this section, the family of forward models is represented by slip on a fault plane that has been divided up into a number of fault patches, and the best-fit model would be the distribution of fault slip magnitude along this fault plane that best matches the observed ground deformation.

Linear vs. Nonlinear Inverse problems can be divided up into two major families: Linear and Nonlinear. In a linear problem, we solve for the sum of forward models that best fits the data, without any size constraints on the individual contributions from each forward model (fault patch). One example of a linear problem is where we find the best-fit fault slip distribution on a fixed fault plane, without any constraints such as requiring that the fault slip has a positive or negative sign (i. e., right-lateral vs. left lateral). Linear problems are essentially extensions of the problem of fitting a line to a set of data, and are quite easy to solve [52].

Nonlinear problems exist when the functional shape and form of the predicted data varies with the parameter choice. An inversion for the location of the fault plane itself, where quantities such as the x, y, z, location or fault plane strike and dip are varied, is very nonlinear. The process of solving a nonlinear problem involves finding a model that best fits the data in some pre-defined way, usually by finding the model that produces the lowest sum of squared residuals with the data (least-squares fit). Nonlinear problems may have multi-dimensional misfit functions with poorly defined or multiple minima, and the only way to find the absolute or "global" minimum is by exhaustively searching the parameter space.

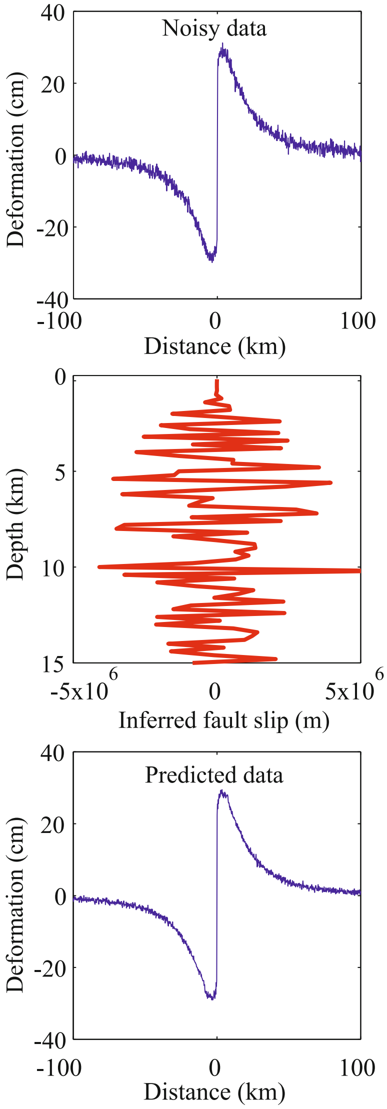
Methods for searching the parameter space in a nonlinear problem fall into two main camps. Gradient-based methods determine the slope and gradient of the misfit surface (often in many dimensions) and follow it downhill in an iterative way. These methods work well if the misfit surface has a single, well defined minimum. If the search begins near one of the local minima, it is possible that the true, global minimum will never be found. Global methods involve variations that range from simply sampling the parameter space very densely, to iterative methods that track many initial samples or families of samples in gradient-based searches. One method that performs very well in the ~9-dimensional parameter space of finding the best-fit fault plane and slip that fits data for a small earthquake is the Neighborhood Algorithm [44,78]. In the Neighborhood Algorithm (NA), the misfit values at many initial points are used to iteratively focus in on regions of the parameter space with lower misfit values. The NA method has the strength of being able to track and define multiple minima instead of just choosing one. A survey of several other nonlinear optimization methods applied to earthquake location problems can be found in Cervelli [6].

Smoothing The specific problem of inverting for fault slip on a particular fault plane has an interesting twist to it that is shared by many geophysical inverse problems. Just as an inversion for the best-fit line to data usually requires at least two data points, other inversions also effectively require the existence of more data than unknown quantities. For fault-slip problems the question does not boil down to simple numbers of data points vs. the number of fault patches – parts of the fault plane are better-resolved than others, and some data points contribute more information than is provided by their neighbors. For instance, if we had a thousand GPS receivers within 100 feet of the fault zone, but extended our target fault plane down to 100 km, the GPS data would be able to tell us little about what was going on at depth, even if we only divided the fault plane up into a few fault patches. Conversely, data points very close together and far from the fault contribute essentially the same information about what occurred – combining them can help to reduce noise, but it does not help us to isolate variations in fault slip along the fault.

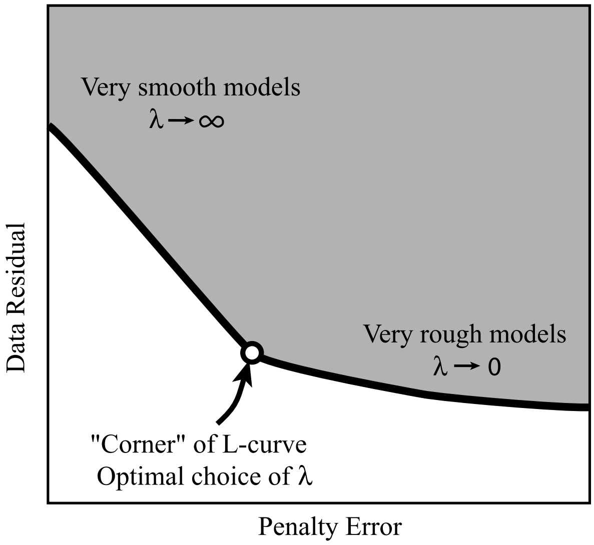
Because of these inherent variations in resolution, both in model and data, an inversion can run into problems if we simply solve for the best-fit fault slip distribution on a finely-divided fault plane. Since the forward models for two fault patches at great depth are very similar on the surface, the inversion cannot determine the relative strength of fault slip that should be assigned to each. The difference between [0 1] and [-1000 1001] is very small when you propagate it up to the surface, often smaller than the noise in the data (or even machine precision!). Therefore, the "best-fit" model would have unrealistically large variations that have nothing to do with what really occurred during the earthquake (Fig. 8). This effect is often known as "checkerboarding". While some of the features in the inferred slip distribution are related to the earthquake, much of the complexity is due to the attempt of the inversion to fit noise within the data.

There are several methods for dealing with this effect, all of which place some bounds on how large or spatially rough the variations in fault slip can be. The simplest method (although one of the hardest to optimize) is to parametrize the fault plane in a way that the fault patches are never so small that there are large tradeoffs in their predicted surface deformation [70]. The extreme of this would be to only use one fault patch. In most cases where we have a lot of data, this solution would not fit the data well nor tell us much about the earthquake.

Another method is to place a penalty during the least squares inversion on fault slip solutions that are very large (usually involving large variations) or that are spatially rough. These "regularized" inversions result in spatially smooth slip distributions that usually fit the data almost as well as the rough, unregularized inversions. Regularization always involves some choice of how much weight needs to be placed on the roughness penalty vs. the fit to the data, which can be a difficult procedure. Too large of a penalty weight and the slip distribution will be too smooth and will not fit the data (the logical extension of this is the single fault patch model discussed above). Too small of a weighting and the slip distribution will be arbitrarily rough, with unrealistic changes in sign throughout the fault zone.



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 8
Example of how noisy data can affect fault slip inversion. *Top:* Deformation from a vertical strike slip fault, with added noise. *Middle:* Best fit slip distribution inferred from noisy data. The large fluctuations mostly cancel each other out at the surface, producing (*bottom*) predicted data that matches both the underlying deformation signal and the noise



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 9
Typical L-curve. Each *point* on the curve corresponds to a slip distribution inferred using a different value of λ. In most cases, the "penalty error" will be a measure of model roughness

There are two main families of methods for choosing the appropriate smoothing weight (apart from random guessing, which often performs surprisingly well!). In general, as the amount of regularization is increased, the fault slip distribution becomes smoother and the fit to the data worsens. Ideally, we would choose the value where the proportion of the slip distribution that is just fitting data noise is as small as possible, without smoothing so much that we fit neither the noise nor the underlying signal due to the earthquake. The first method requires choosing the smoothing weight from a plot of data fit vs. model roughness, often referred to as an "L-curve" because of the characteristic shape of such curves (Fig. 9). The smoothing value at the corner of the "L" generally fits the data well without being "too rough". This type of parameter choice is slightly arbitrary and not easy to automate, but it does allow the researcher to include some intuition about characteristics of the earthquake.

The other family relies on the concept that a good choice of smoothing would be able to reproduce another, independent set of data spanning the earthquake fairly well. Too much smoothing and the slip model would fit neither the original nor the additional data sets well. If the choice of smoothing were too small, the resulting complex slip distribution would mainly be fitting noise in the original data set and would not, therefore, fit the independent noise in the second data set.

Of course, we rarely have the luxury of multiple data sets – if they do exist, we should use them for the main inversion! Data resampling procedures, known variously as the bootstrap, jacknife or cross-validation [10] are used to simulate the existence of multiple, independent data sets in cases where only one exists. Du et al. [9] compared cross-validation and other techniques for choosing smoothing parameters for geodetic data spanning the 1983 Borah Peak earthquake. Another powerful parameter choice method is the Akaike Bayesian Information Criterion (ABIC) that can be used to choose smoothing weights or other inversion characteristics [1,30,97].

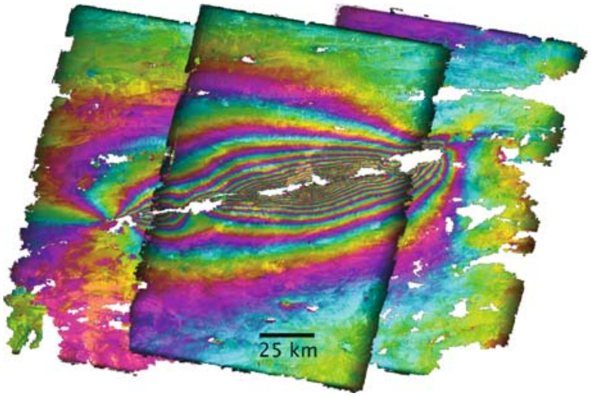
Noise Although finding the best-fit model to our data is certainly important, knowing how confident we are in that estimate is just as crucial. Studies of postseismic behavior, for instance, rely on good estimates of the coseismic slip distribution [25], and are improved further when we know what constraints we can place on that slip distribution. A variety of techniques exist for estimating these confidence limits – for linear problems, and when we know the character of the noise (magnitude, spatial correlation), it is quite simple to propagate data errors through to error bounds on the inversion results [52]. However, for nonlinear problems, and for cases where we are not quite sure how much of our signal is noise (usually the case with InSAR data), we need to rely on other methods for estimating the noise.

The same data resampling procedures described above can be used to generate multiple sets of inversion results that should reflect the noise structure of the data, even when it is not understood ahead of time. For nonlinear problems, such as the 3-D location of the best fit fault plane and earthquake mechanism for a particular earthquake, knowledge about the data noise can be used to construct multiple synthetic data sets that can then each be inverted using the nonlinear method of choice. Any conclusions about the problem (i. e., error bounds on fault slip) should be drawn from the ensemble as a whole [44].

Case Examples

Here we examine case examples for two earthquakes. In each case, we review the data and discuss characteristics of the inferred slip distributions.

Mw 7.6 Manyi, Tibet, Earthquake The 1997 strike-slip earthquake that occurred Tibet produced a very long rupture (170 km) with up to 7 meters of offset. It was such a long rupture that it takes three overlapping SAR tracks (Fig. 10) to cover its full length! Here, we show the effects



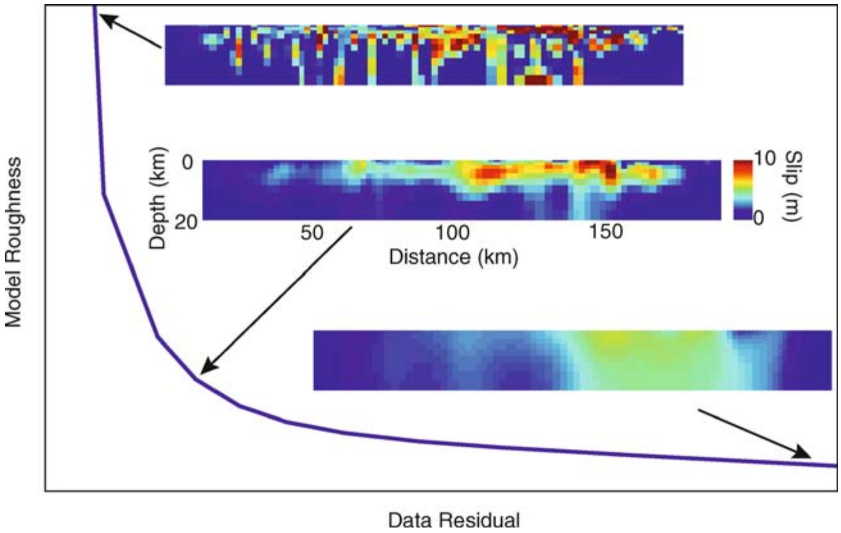
Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 10
Three overlapping interferograms spanning the 1997 Manyi earthquake

of variations in the spatial smoothing weights (λ) on an inversion of the deformation field for fault slip. Figure 11 illustrates the differences in inferred slip distributions for various magnitudes of λ, resulting in a spectrum between very smooth slip distributions that do not fit the data very well, to impossibly rough slip distributions that only fit the data marginally better than the "optimal" model (center panel). Our optimal model predicts up to 10 meters of fault slip with most slip occurring in the upper 10 km. Here, we chose the optimal value of λ using cross-validation, but the other methods discussed above can also be applied to this problem.

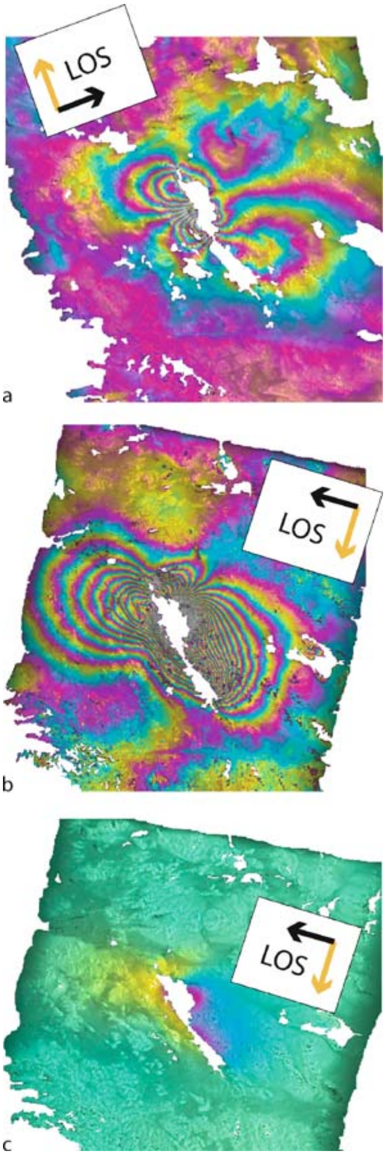
Asymmetries in the profiles of deformation across the fault during the Manyi earthquake have led some researchers to think that the elastic behavior of the crust may be different on one side of the fault vs. the other [67]. The magnitude of horizontal ground motion for a straight, vertical, strike-slip fault should be symmetric across the fault, if there are no other complications. However, variations in fault plane dip can affect the resulting ground deformation as much as variations in crustal properties, requiring that such studies carefully consider the fault geometry used in their inversions.

Mw 7.1 Hector Mine, CA, Earthquake The 1999 Hector Mine earthquake [16,85] occurred in the Mojave Desert of Southern California, and was widely studied due to it's proximity to large population centers and the San Andreas fault, it's very complete data coverage (GPS, InSAR, seismic data), and the fact that it occurred only a few years after another very large earthquake on a neighboring fault (1992 Landers earthquake). The combination of the two earthquakes within such a short time interval suggests that series of earthquakes may occur separated by long intervals of quiescence instead of events always occurring at semi-regular time intervals of strain buildup and release [61,69].

In Fig. 12, we show a subset of the available InSAR data spanning the Hector Mine earthquake. Both continuous and campaign GPS data were also recorded on either side of the Y-shaped rupture. In Fig. 13, we show the re-

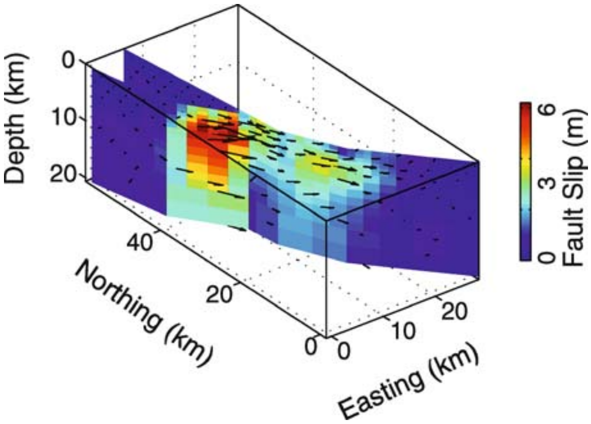


Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 11
L-curve for the Manyi EQ, showing slip distributions inferred using 3 different values of smoothing, each with the same color scale. Small λ (*top panel*) produces an unrealistically rough slip distribution, whereas very large λ (*bottom panel*) produces a smooth model that doesn't fit the data

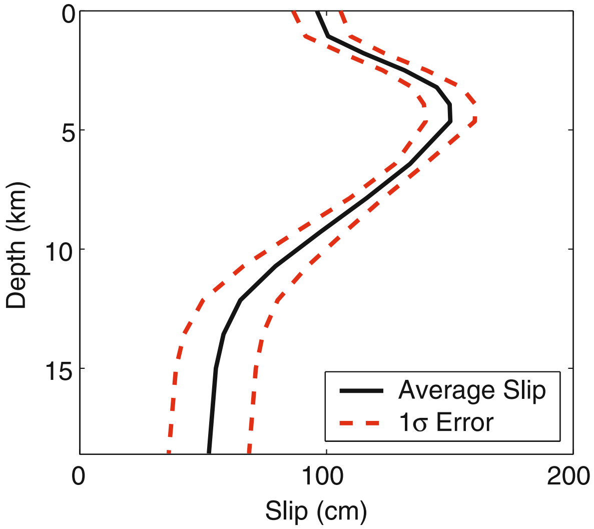


Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 12
Three representations of the Hector Mine EQ. a "Ascending" and b "descending" wrapped interferograms, where the satellite is moving in the direction indicated by the *yellow arrow* and viewing in the direction of the *black arrow*. Note how the same deformation field looks very different depending on the viewing direction. Color cycles correspond to ~3 cm of deformation. c Unwrapped version of interferogram in b. Now the color scale corresponds to 3 meters and spans the entire dynamic range of the data

sults of inverting the GPS and all the InSAR data for the best-fit fault slip distribution, with up to 6 meters of fault slip. Note how the maximum fault slip is not at the surface, but peaks at a few km depth. If we plot the average slip vs. depth (Fig. 14), the profile clearly shows how the shallow slip deficit is robust even given the amount of noise in the data.



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 13
3D representation of inferred fault slip that occurred during the Hector Mine earthquake



Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 14
Fault slip during the Hector Mine earthquake, averaged along strike and plotted vs. depth. *Red dashed lines* indicate the expected error bounds introduced by atmospheric noise

Future Directions

The spatial and temporal coverage of geodetic data sets such as InSAR and GPS is increasing to the point where we can observe new types of fault zone behavior and solve for characteristics on a scale and precision that could hardly have been imagined a few decades ago. As we increase our ability to understand and describe the kinematics of crustal deformation, we can begin to explore the dynamic processes driving seismic and volcanic deformation in tectonically active regions around the world. Three of the main fields that show the most promise in the near future are discussed below:

Transient Deformation and Slow, Silent Earthquakes

The episodic slow slip events that show up with amazing regularity (∼14 months) in Cascadia ([51,54], Fig. 6), and that are also observed in Japan [27,56,65] and Mexico [45], are some of the most exciting new types of fault zone behavior that scientists have observed in recent years. The idea that earthquakes contained energy that was released by slower ruptures was noticed using seismic data far earlier than it was ever observed geodetically (1960 Chilean earthquake [34]). The slow slip events are often associated with heightened levels of spatially-correlated, low-level seismicity, or seismic tremor [87]. Currently, the process that allows these slow earthquakes to initiate and propagate at the observed speeds is still unknown, although they may be occurring at depths where dehydration of subducted sediments becomes important.

Apart from their intrinsic interest (since we always seek to explain processes that we do not understand), these slow/silent events release a not insignificant amount of the accumulated strain across the fault boundaries. In some ways this is a boon to people living nearby – any plate tectonic motion that is accommodated aseismically means that there is that much less accumulated strain that could be released in a destructive earthquake. However, there are indications that aseismic slip events are often closely followed by earthquakes [42,75], indicating that the changes in stress during the slow event can push other, nearby, regions past the brink until they rupture coseismically. This, also, can be seen in a positive light – perhaps close monitoring of aseismic deformation along active plate boundaries can serve as an early warning system, or at least a signal to raise the forecasted earthquake hazard whenever heightened activity is observed.

Reduction or Modeling of Atmospheric Noise

One of the largest hurdles for researchers approaching InSAR data is the atmospheric noise present in all interferograms. GPS data is affected by this problem as well, but the atmospheric signal can be solved for or averaged out due to the long observation times allowed by GPS. An area of active research is the modeling and removing of the atmospheric noise signal from InSAR data, using a variety of tools that range from other satellite-based observations of atmospheric water vapor content to simplistic models of correlations between elevation and interferometric phase [96].

The first-order layering of the atmosphere results in profiles of water vapor that tend to decrease vs. elevation in a manner that often appears essentially linear when two SAR images are combined in an interferogram. However, lateral variations with distance from water bodies and gradients from one side of a mountain belt to the other, result in the fact that the appropriate elevation vs. signal can vary quite a bit across a typical SAR image. For some target problems, such as the location of a small earthquake or the fault slip distribution for an earthquake in a relatively flat area, these elevation-dependent signals will likely affect the inversion to only a limited extent. However, in studies where the signal of interest is correlated with elevation, such as fold growth or subsidence caused by lake loading, great care should be taken when removing overly simple models of atmospheric water vapor from the data.

A more promising, albeit still problematic, approach is the modeling of water vapor content based on observations from satellite-based platforms. Water vapor measurements are made by a variety of satellites, but two in particular have the spatial scale and temporal resolution that make them potentially very useful in InSAR applications. The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments on both the Terra and Aqua satellites acquire almost daily observations over most of the Earth's surface at fairly high resolution (1 km spacing). They are not acquired at exactly the same time as the SAR images, but they can still be used to track seasonal trends or to seek individual features that may persist between the SAR and MODIS image acquisition times. While these cannot be used to quantitatively remove the atmospheric signal, they can be very useful from a qualitative standpoint in determining whether a particular interferometric signal may or may not be tectonic in origin.

The Medium Resolution Imaging Spectrometer (MERIS) is physically located on the same satellite (ENVISAT) that acquires much of the SAR imagery used today. Since MERIS observations are made at essentially the same time as the SAR imagery, they see the same atmosphere and can not only be used in the same qualitative way as MODIS, but also show promise of allowing the actual removal of atmospheric contributions to the interferometric phase. Some difficulties lie in the fact that MERIS and MODIS measurements of atmospheric water vapor can only be made in cloud-free images, and that "double-bounce" effects of the signal bouncing off the base of both visible and invisible clouds can bias the observations and introduce spurious water vapor features. Still, these satellite-based methods show great promise. The most robust method for reducing the contribution from atmospheric noise to our inversions remains the use of multiple independent data sets, whenever available.

Data Assimilation

As mentioned in Sect. "Highlights of Earthquake Geodesy", the vast quantity of data now available to us can prove troublesome – if it takes hours to perform one forward model of a possible earthquake scenario with the available data, then the inverse problem quickly becomes unmanageable (or at least can begin to extend past the length of a normal graduate thesis). Parallel computing methods and the rapid decrease in cost for computing resources now makes the operation of large, multi-processor machines feasible within individual research departments. These large machines allow us to approach inverse problems with the "big hammer" of stochastic, or Monte Carlo, methods, which rely on the use of many randomly generated simulations of a system [6]. However, the expected continuing increase of both spatial and temporal coverage of deformation observations requires that we need to continue developing new tools that allow us to capitalize on data time series as well as individual snapshots of ground deformation.

Kalman filters and related methods are one very powerful tool now in widespread use, especially within the GPS community [51,84]. The incorporation of InSAR data into such methods is slightly more difficult, in part because of the large number of data points, but also because of the varying character of noise between interferograms and difficulties in dissociating the description of noise vs. modeling of the geophysical signal of interest. Data assimilation methods developed in the atmospheric sciences, which also deal with data sets of varying spatial and temporal scales and resolutions, are a potentially rich source of tools that the geodetic community can explore in the near future.

Bibliography

Primary Literature

1. Akaike H (1980) Bayesian statistics. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds) Likelihood and the Bayes procedure. University Press, Valencia, pp 143–166
2. Allen CR et al (1972) Displacements on the Imperial, Superstition Hills, and San Andreas faults triggered by the Borrego Mountain Earthquake. US Geol Surv Prof Pap 787:87–104
3. Arnadottir T, Segall P (1994) The 1989 Loma Prieta earthquake imaged from inversion of geodetic data. J Geophys Res 99:21835–21855
4. Backus G, Gilbert F (1970) Uniqueness in the inversion of inaccurate gross earth data. Phil Trans R Soc Lond 266:123–192
5. Burgmann R et al (2002) Time-dependent afterslip on and deep below the Izmit earthquake rupture. Bull Seism Soc Amer 92:126–137
6. Cervelli P et al (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan. J Geophys Res 106:11217–11237
7. Dieterich JH (1992) Earthquake nucleation on faults with rate- and state- dependent strength. Tectonophysics 211:115–134
8. Dominguez S, Avouac JP, Michel R (2003) Horizontal coseismic deformation of the 1999 Chi-Chi earthquake measured from SPOT satellite images: Implications for the seismic cycle along the western foothills of central Taiwan. J Geophys Res 108. doi: 10.1029/2001JB000951
9. Du Y, Aydin A, Segall P (1992) Comparison of various inversion techniques as applied to the determination of a geophysical deformation model for the 1983 Borah Peak earthquake. Bull Seism Soc Amer 82:1840–1866
10. Efron B, Tibshirani R (1993) An introduction to the bootstrap. In: Monographs on statistics and applied probability, vol 83. Chapman and Hall, London
11. Emardson TR, Simons M, Webb FH (2003) Neutral atmospheric delay in interferometric synthetic aperture radar applications: Statistical description and mitigation. J Geophys Res 108. doi: 10.1029/2002JB001781
12. England P, Jackson J (1989) Active deformation of the continents. Annu Rev Earth Planet Sci 17:197–226
13. Fialko Y (2004) Evidence of fluid-filled upper crust from observations of post-seismic deformation due to the 1992 Mw 7.3 Landers earthquake. J Geophys Res 109. doi: 10.1029/2004JB002985
14. Fialko Y (2006) Interseismic strain accumulation and the earthquake potential on the southern San Andreas fault system. Nature 441:968–971
15. Fialko Y et al (2002) Deformation on nearby faults induced by the 1999 Hector Mine Earthquake. Science 297:1858–1862
16. Fialko Y, Simons M, Agnew D (2001) The complete (3-D) surface displacement field in the epicentral area of the 1999 Mw 7.1 Hector Mine earthquake, California, from space geodetic observations. Geophys Res Lett 28:3063–3066
17. Freymueller J, King NE, Segall P (1994) The Co-seismic slip distribution of the Landers earthquake. Bull Seism Soc Amer 84:646–659
18. Funning GJ et al (2005) Surface displacements and source parameters of the 2003 Bam (Iran) earthquake from Envisat advanced synthetic aperture radar imagery. J Geophys Res 110. doi: 10.1029/2004JB003338
19. Gilbert GK (1890) Lake Bonneville. In: US Geol Surv Monograph, vol 1. Washington
20. Goldstein R (1995) Atmospheric limitations to repeat-track radar interferometry. Geophys Res Lett 22:2517–2520
21. Griesbach CL (1893) Notes on the earthquake in Baluchistan on the 20th December 1892. Geol Survey India Rec 26
22. Hager BH, King RW, Murray MH (1991) Measurement of crustal deformation using the global positioning system. Annu Rev Earth Planet Sci 19:351–382
23. Hanssen RA (2001) Radar interferometry: Data interpretation and error analysis. Kluwer, Dordrecht

24. Harris RA, Segall P (1987) Detection of a locked zone at depth on the Parkfield, California segment of the San Andreas Fault. J Geophys Res 92:7945–7962
25. Hearn EH (2002) Dynamics of Izmit earthquake postseismic deformation and loading of the Duzce earthquake hypocenter. Bull Seism Soc Amer 92:172–193
26. Hetland EA, Hager BH (2003) Postseismic relaxation across the central Nevada seismic belt. J Geophys Res 108. doi: 10.1029/2002JB002257
27. Hirose H et al (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. Geophys Res Lett 26:3237–3240
28. Hsu YJ et al (2006) Frictonal afterslip following the Mw 8.7, 2005 Nias-Simeulue earthquake, Indonesia. Science 312. doi: 10.1126/science.1126960
29. Ito T, Hashimoto M (2004) Spatiotemporal distribution of interplate coupling in southwest Japan from inversion of geodetic data. J Geophys Res 109. doi: 10.1029/2002JB002358
30. Jackson DD, Matsuura M (1985) A Bayesian approach to nonlinear inversion. J Geophys Res 90:581–591
31. Johanson IA, Burgmann R (2005) Creep and quakes on the northern transition zone of the San Andreas Fault from GPS and InSAR data. Geophys Res Lett 32. doi: 10.1029/2005GL023150
32. Johnson KM, Burgmann R, Larson K (2006) Frictional properties on the San Andreas fault near Parkfield, California, inferred from models of afterslip following the 2004 earthquake. Bull Seism Soc Amer 96:S321–S338
33. Jonsson S et al (2003) Post-earthquake ground movements correlated to pore-pressure transients. Nature 424:179–183
34. Kanamori H, Stewart GS (1972) A slow earthquake. Phys Earth Planet Int 18:167–175
35. King GCP, Stein RS, Rundle JB (1988) The growth of geological structures by repeated earthquakes 1: Conceptual framework. J Geophys Res 93:13307–13318
36. Koto B (1983) On the cause of the great earthquake in central Japan, 1891. J Coll Sci Imp Univ Japan 5:296–353
37. Langbein J, Gwyther RL, Hart RHG, Gladwin MT (1999) Slip-rate increase at Parkfield in 1993 detected by high-precision EDM and borehole tensor strainmeters Source. Geophys Res Lett 26(16):2529–2532
38. Langbein J, Murray JR, Snyder HA (2006) Coseismic and initial postseismic deformation from the 2004 Parkfield, California, Earthquake, observed by Global Positioning System, Electronic Distance Meter, Creep Meters, and Borehole Strainmeters. Bull Seism Soc Amer 96:304–320
39. Larsen S et al (1992) Global Positioning System measurements of deformations associated with the 1987 Superstition Hills earthquake – evidence for conjugate faulting. J Geophys Res 97:4885–4902
40. Lawson AC et al (1908) The California earthquake of April 18, 1906 – report of the state earthquake investigation committee. Carnegie Insitute, Washinton
41. Lin J, Stein RS (1989) Coseismic folding, earthquake recurrence and the 1987 source mechanism at Whittier Narrows, Los Angeles Basin, California. J Geophys Res 94:9614–9632
42. Lohman RB, McGuire JJ (2007) Earthquake swarms driven by aseismic creep in the Salton Trough, California. J Geophys Res 112. doi: 10.1029/2006JB004596
43. Lohman RB, Simons M (2005) Some thoughts on the use of InSAR data to constrain models of surface deformation: Noise structure and data downsampling. Geochem Geophys Geosyst 6. doi: 10.1029/2004GC000841
44. Lohman RB, Simons M (2005) Locations of selected small earthquakes in the Zagros mountains. Geochem Geophys Geosyst 6. doi: 10.1029/2004GC000849
45. Lowry AR et al (2001) Transient fault slip in Guerrero, southern Mexico. Geophys Res Lett 28:3753–3756
46. Lyell C (1837) Principles of Geology, 5th edn. Murray, London
47. Lyons S, Sandwell D (2003) Fault creep along the southern San Andreas from interferometric synthetic aperture radar, permanent scatterers and stacking. J Geophys Res 108. doi: 10.1029/2002JB001831
48. Mallet R (1862) The first principles of observational seismology. Chapman and Hall, London
49. Massonnet D, Rossi M, Carmona C, Adragna F, Pelzer G, Feigl K, Rabaute T (1993) The displacement field of the Landers earthquake mapped by radar interferometry. Nature 364: 138–142
50. McKay A (1890) On the earthquake of September 1888, in the Amuri and Marlborough districts of the South Island. NZ Geol Surv Rep Geol Explor 1885–1889 20:78–1007
51. McGuire J, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. Geophys J Int 155:778–788
52. Menke W (1989) Geophysical data analysis: Discrete inverse theory. Academic Press, London
53. Middlemiss CS (1910) The Kangra earthquake of 4th April, 1905. Geol Surv India Mem 37
54. Miller MM et al (2002) Periodic slow earthquakes from the Cascadia subduction zone. Science 295:2423
55. Miyazaki S et al (2004) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties. Geophys Res Lett 31. doi: 10.1029/2003GL019410
56. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. J Geophys Res 108. doi: 10.1029/2001JB000456
57. Molnar P, Tapponnier P (1975) Cenozoic tectonics of Asia: Effects of a continental collision. Science 189:419–425
58. Muller JJA (1895) De verplaasting van eenige traiangulatie pilaren in de residenti Tapanuli (Sumatra) tengevolge de aardbeving van 17 Mei 1892. Natuurwet Tijdscht Ned Indie 54: 299–307
59. Murray J, Segall P (2002) Testing time-predictable earthquake recurrence by direct measurement of strain accumulation and release. Nature 419:298–291
60. Nadeau RM, McEvilly TV (2004) Periodic pulsing of characteristic microearthquakes on the San Andreas Fault. Science 303:202–222
61. Nur A, Hagai R, Beroza G (1993) The nature of the Landers-Mojave earthquake line. Science 261:201–203
62. Okada Y (1985) Surface deformation due to shear and tensile faults in a half space. Bull Seism Soc Amer 75:1135–1154
63. Okudo T (1950) On the mode off the vertical land-deformation accompanying the great Nankaido earthquakes. Bull Geogr Surv Inst 2:37–59
64. Oldham RD (1928) The Cutch (Kacch) earthquake of 16th June 1819, with a revision of the great earthquake of 12th June 1897. Geol Surv India Mem 46:71–147

65. Ozawa S et al (2002) Detection and monitoring of ongoing aseismic slip in the Tokai region, central Japan. Science 298:1009–1012
66. Parker RL (1977) Understanding inverse theory. Annu Rev Earth Planet Sci 5:35–64
67. Peltzer G, Crampe F, King G (1999) Evidence of nonlinear elasticity of the crust from the Mw 7.6 Manyi (Tibet) earthquake. Science 286:272–276
68. Pollitz FF (2003) Transient rheology of the uppermost mantle beneath the Mojave desert, California. Earth Plan Sci Lett 215:89–104
69. Pollitz FF, Sacks IS (2002) Stress triggering of the 1999 Hector Mine earthquake by transient deformation following the 1992 Landers Earthquake. Bull Seism Soc Amer 92:1487–1496
70. Pritchard ME et al (2002) Co-seismic slip from the 1995 July 30 Mw=8.1 Antofagasta, Chile, earthquake as constrained by InSAR and GPS observations. Geophys J Int 150:362–376
71. Pritchard ME, Simons M (2006) An aseismic slip pulse in northern Chile and along-strike variations in seismogenic behavior. J Geophys Res 111. doi: 10.1029/2006JB004258
72. Reid HF (1910) The mechanics of the earthquake. In: Lawson AC (ed) The California earthquake of April 18, 1906. Carnegie Institute, Washington
73. Reid HG (1911) The elastic-rebound theory of earthquakes. Univ Calif Pub Bull 6:413–444
74. Reilinger R (1986) Evidence for postseismic viscoelastic relaxation following the 1959 M=7.5 Hebgen Lake, Montana, earthquake. J Geophys Res 91:9488–9494
75. Roeloffs EA (2006) Evidence for aseismic deformation rate changes prior to earthquakes. Annu Rev Earth Planet Sci 34:591–627
76. Rosen PA et al (2000) Synthetic aperture radar interferometry. Proc IEEE 88:333–382
77. Ruegg JC et al (1996) The Mw=8.1 Antofagasta (North Chile) earthquake July 30, 1995: First results from teleseismic and geodetic data. Geophys Res Lett 23:917–920
78. Sambridge M (1998) Geophysical inversion with a neighborhood algorithm – I: Searching a parameter space. Geophys J Int 138:479–494
79. Savage JC, Lisowski M, Svarc JL (1994) Postseismic deformation following the 1989 (M=7.1) Loma Prieta, California, earthquake. J Geophys Res 99:13757–13765
80. Schmidt DA et al (2005) Distribution of aseismic slip rate on the Hayward fault inferred from seismic and geodetic data. J Geophys Res 110. doi: 10.1029/2004JB003397
81. Scholz CH (1998) Earthquakes and friction laws. Nature 391: 37–42
82. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. J Geophys Res 105:S615–S634
83. Segall P, Davis JL (1997) GPS Applications for Geodynamics and Earthquake Studies. Annu Rev Earth Planet Sci 25:301–336
84. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. J Geophys Res 102:22391–22409
85. Simons M, Fialko Y, Rivera L (2002) Coseismic deformation from the 1999 Mw 7.1 Hector Mine, California, earthquake as inferred from InSAR and GPS observations. Bull Seism Soc Amer 92:1390–1402
86. Sharp RV et al (1982) Surface faulting in the Imperial Valley. In: Sharp RV, Lienkaemper JJ, Bonilla MG, Burke DB, Cox BF, Herd DG, Miller DM, Morton DM, Ponti DJ, Rymer MJ, Tinsley JC, Yount JC, Kahle JE, Hart EW, Sieh K (eds) The Imperial Valley, California, earthquake of October 15, 1979. US Geol Surv Prop Pap 1254:119–144
87. Shelley DR et al (2006) Low frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. Nature 442:188–191
88. Shimazaki K, Nakata T (1980) Time-Predictable Recurrence Model for Large Earthquakes. Geophys Res Lett 7(4):279–282
89. Song AT, Simons M (2003) Large trench-parallel gravity variations predict seismogenic behavior in subduction zones. Science 301:630–633
90. Stein R (1999) The role of stress transfer in earthquake occurrence. Nature 402:605–609
91. Tarantola A, Valette B (1982) Inverse problems = quest for information. J Geophys 50:159–170
92. Thatcher W (1984) The earthquake deformation cycle at the Nankai trough, southwest Japan. J Geophys Res 89:3087–3101
93. Tsuboi C (1932) Investigation on the deformation of the earth's crust in the Tango district connected with the Tango earthquake of 1927 (part 4). Bull Earthq Res Inst Tokyo Univ 10: 411–434
94. Ward S, Valensise GR (1989) Fault parameters and slip distribution of the 1915 Avezzano, Italy, earthquake derived from geodetic observations. Bull Seism Soc Amer 79:690–710
95. Wells RE et al (2003) Basin-centered asperities in great subduction zone earthquakes: A link between slip, subsidence and subduction erosion. J Geophys Res 108. doi: 10.1029/2002JB002072
96. Williams S, Bock Y, Fang P (1998) Integrated satellite interferometry: Tropospheric noise, GPS estimates and implications for interferometric synthetic aperture radar products. J Geophys Res 103:27051–27067
97. Yabuki T, Matsuura M (1992) Geodetic data inversion using a Bayesian information criterion for spatial distribution of fault slip. Geophys J Int 109:363–375
98. Yang M et al (2000) Geodetically observed surface displacements of the 1999 Chi-Chi, Taiwan earthquake. Earth Planet Space 52:403–413
99. Yeats RS, Sieh K, Allen CF (1997) The geology of earthquakes. Oxford University Press, New York
100. Zebker HA, Rosen PA, Hensley S (1997) Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps. J Geophys Res 102:7547–7563

Books and Reviews

Tse ST, Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. J Geophys Res 91:9452–9472

Cryosphere Models

ROGER G. BARRY
NSIDC, CIRES, University of Colorado, Boulder, USA

Article Outline

Glossary

Cryosphere All forms of terrestrial snow and ice.

Newtonian viscous body A body whose stress at each point is linearly proportional to its strain rate at that point.

Definition of the Subject

The cryosphere comprises all terrestrial forms of snow and ice – snow cover, floating ice, glaciers, ice sheets, frozen ground and permafrost. It is a critical element of the climate system because of its high reflectivity, its insulating effects on the land and ocean, and its storage of water on short and long time scales. Numerical models of components of the cryosphere have been developed over the last 30 years or so, and some elements of these are now incorporated in coupled climate models and earth system models.

Introduction

Currently there are no comprehensive models of the entire cryosphere. Rather there is a wide range of models of components of the cryosphere – snow cover, floating ice, glaciers, ice sheets, frozen ground and permafrost – and various components are treated with varying degrees of detail in coupled atmosphere-ocean-land models. Cryospheric processes are generally parametrized in such earth system models.

Models developed for each of the main cryospheric components are discussed in turn.

Snow Cover

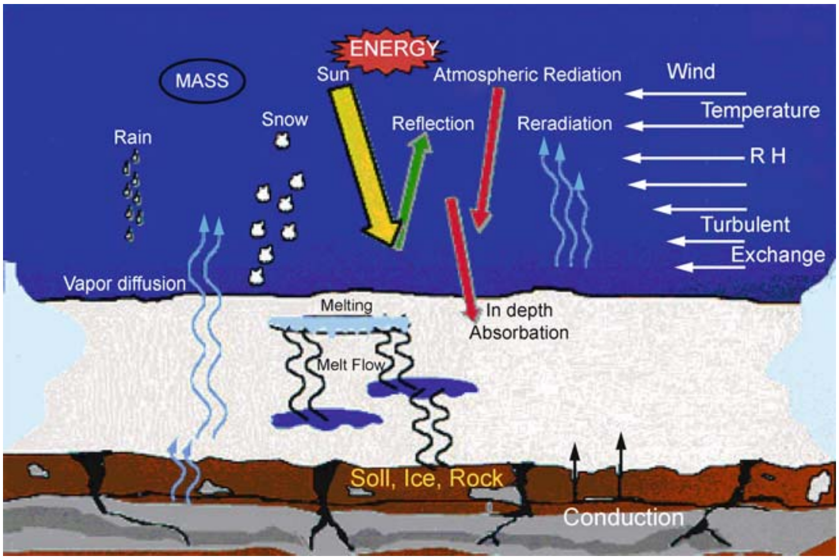
Snow cover is observed in situ at hydrometeorlogical stations, from daily depth measurements, (monthly) snow courses and in special automated networks such as the western United States Snow Telemetry (SNOTEL) network of snow pressure pillows. Its extent is also observed and mapped daily (since June 1999) over the Northern Hemisphere from operational satellites of the National Oceanic and Atmospheric Administration (NOAA) in the USA. Snow covers about 47 million km^2 at maximum in January and there is only a small area in South America in July. Hemispheric snow water equivalent estimates are routinely made from passive microwave data (1979-present) with a 25-km resolution (Armstrong et al. [4]).

There are numerous models of the formation and disappearance of snow cover. Many have a hydrological focus aimed at estimating seasonal runoff. Some use simple temperature degree-day formulations while others incorporate a full energy balance calculation. Dozier and Painter [14] examine the use of multispectral and hyperspectral remote sensing to estimate the snow's spectral albedo, along with other properties such as grain size, contaminants, temperature, liquid water content, and depth or water equivalent.

The U.S. Army Cold Regions Research and Engineering Laboratory Model SNTHERM (SNow THERmal Model) is a 1-D energy balance model for snow and soil that is forced by meteorologically determined surface fluxes [41]. It simulates in-snow properties and processes, such as heat conduction, water flow, melt, vapor flow, compaction, grain growth, and solar absorption (see Fig. 1). The output provides snow depth, profiles of snow temperature, water content, density, grain size, and surface fluxes of sensible heat and evaporation. Surface boundary conditions require: incoming solar and longwave radiation; wind speed, air temperature and humidity at some reference height; and precipitation. The model will estimate solar and longwave radiation from cloud cover, if data on these variables are not available. Lower boundary conditions include soil textural properties (currently clay or sand used as defaults), wetness and temperature profile.

A comparative study of three snow models with different complexities was carried out by Jin et al. [39] to assess how a physically detailed snow model can improve snow modeling within general circulation models. The three models were (a) SNTHERM; (b) a simplified three-layer model, Snow–Atmosphere–Soil Transfer (SAST), which includes only the ice and liquid-water phases; and (c) the snow submodel of the Biosphere –Atmosphere Transfer Scheme (BATS), which calculates snowmelt from the energy budget and snow temperature by the force–restore method. SNTHERM gave the best match to observations with the SAST simulation being close. BATS captured the major processes in the upper layers of a snow pack where solar radiation is the main energy source and gave satisfactory seasonal results.

CROCUS is a model of the Centre d'Etudes de la Neige, Grenoble [10]. It is a 1-D physical model that determines mass and energy balance for a snow cover and is used for operational avalanche forecasting. The snow cover is represented as a pile of layers parallel to the



Cryosphere Models, Figure 1
The snowpack energy balance as characterized by SNTHERM (US Army Corps of Engineers)

ground. Energy exchanges are projected orthogonally to the slope. The model describes the evolution of the internal state of the snow cover as a function of meteorological conditions. The variables describing the snow cover are temperature, density, liquid water content, and snow type of each layer. To match the natural layers, the thickness and number of layers are adjusted by the model. The model simulates the heat conduction, melting/refreezing of snow layers, settlement, metamorphism, and percolation. It simulates dry and wet snow metamorphism with experimental laws derived from laboratory data. Snow grains are characterized by their size and type. This allows an accurate albedo of the snow cover to be calculated.

Bartelt and Lehning [7] also present a 1-D physical model of the snow pack (SNOWPACK) with equations for heat transfer, water transport, vapor diffusion and mechanical deformation. New snow, snow drift and ablation are treated. The snow layers are treated in terms of height, density and microstructure (grain size, shale and bonding). The model is used for avalanche warnings in Switzerland.

Interception Models

A physically-based snowfall interception model that scales snowfall interception processes from branch to canopy is now available [25]. It takes account of the persistent presence and subsequent unloading of intercepted snow in cold climates. To investigate how snow is intercepted at the forest stand scale, measurements of wind speed, air temperature, above- and below-canopy snowfall, accumulation of snow on the ground and the load of snow intercepted by a suspended, weighed, full-size conifer were collected from spruce and pine stands in the southern boreal forest. Interception efficiency is found to be particularly sensitive to snowfall amount, canopy density and time since snowfall. Further work resulted in process-based algorithms describing the accumulation, unloading and sublimation of intercepted snow in forest canopies (Pomeroy et al. [70]). These algorithms are unique in that they scale up the physics of interception and sublimation from small scales, where they are well understood, to forest stand-scale calculations of intercepted snow sublimation.

Blowing Snow Models

Physically-based treatments of blowing snow and wind are used to develop a distributed model of blowing snow transport and sublimation over complex terrain for an Arctic tundra basin by Essery et al. [18]. A reasonable agreement with results from snow surveys is obtained when sublimation processes are included. Sublimation typically removes 15–45% of the seasonal snow cover. The model is able to reproduce the distributions of snow mass, classified by vegetation type and landform, which can be approximated by lognormal distributions. The representation used for the downwind development of blowing snow with changes in wind speed and surface characteristics is shown to have a moderating influence on snow redistribution.

Spatial fields of snow depth have power spectra in one and two dimensions that occur in two frequency intervals separated by a scale break between 7 and 45 m [84]. The break in scaling is controlled by the spatial distribution of vegetation height when wind redistribution is minimal and by the interaction of the wind with surface concavities and vegetation when wind redistribution is dominant.

In mountainous regions, wind plays a prominent role in determining snow accumulation patterns and turbulent heat exchanges, strongly affecting the timing and magnitude of snowmelt runoff. Winstral and Marks [90] use digital terrain analysis to quantify aspects of the upwind topography related to wind shelter and exposure. They develop a distributed time-series of snow accumulation rates and wind speeds used to force a distributed snow model. Terrain parameters were used to distribute rates of snow accumulation and wind speeds at an hourly time step for input to ISNOBAL, an energy and mass balance snow model. ISNOBAL forced with accumulation rates and wind fields generated from the terrain parametrizations accurately models the observed snow distribution (including the formation of drifts and scoured wind-exposed ridges) and snowmelt runoff. By contrast, ISNOBAL forced with spatially constant accumulation rates and wind speeds taken from the sheltered meteorological site at Reynolds Mountain in southwest Idaho, a typical snow-monitoring site, overestimated peak snowmelt inputs and tended to underestimate snowmelt inputs prior to the runoff peak.

Avalanche Models

Avalanches range in size from sluffs with a volume of $< 10\,\text{m}^3$ to extreme releases of 10^7–$10^8\,\text{m}^3$; corresponding impact pressures range from $< 10^3$–10^6 Pa. There are two main types – loose snow avalanches and slab avalanches. Commonly, they begin with the failure of snow layers with densities less than $300\,\text{kg m}^{-3}$. An avalanche path comprises a starting zone, the track, and a runout–deposition zone. Loose snow avalanches are initiated when the angle of repose is exceeded – about 45°. The angle increases as temperatures rise due to increased cohesion. Slush avalanches can occur on slopes <10°. Downslope propagation continues to the kinetic angle of repose at about 17° [52,69]. Slab avalanches occur when a cohesive slab is released over an extensive plane of weakness on slopes of 35–40°. Slab thicknesses are 0.1–4 m and have a mean density of $\sim 200\,\text{kg m}^{-3}$. Bed surface temperatures are near 0°C.

The variables of interest for forecasting are velocity, run-out distance and impact pressure. The acceleration of an avalanche is resisted by surface friction, air drag at the front and upper boundary, and ploughing at the advancing front and underneath surface. According to Perla [69], maximum velocities range from 20–$30\,\text{m s}^{-1}$ for path lengths up to 500 m and slope angles of 25–35°. The mean run-out length on 67 Colorado avalanche paths was 380 m (Bovis and Mears [9]). On occasion, the run-out may cross a valley floor and continue up the facing slope. Impact pressures are a maximum at 1–2 m above the surface and range in value from about 1–10×10^5 Pa (Perla [69]).

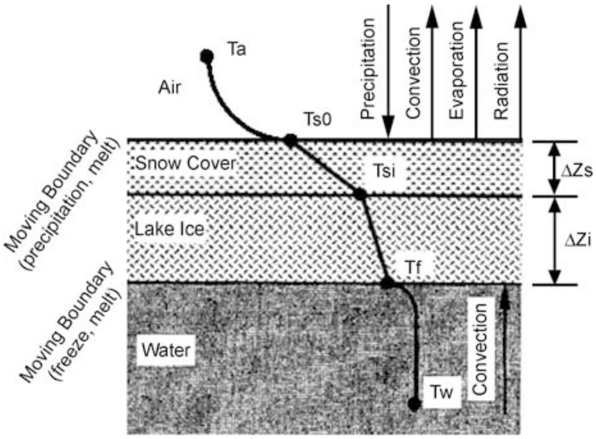
Land Surface Schemes in GCMs

Snow cover is treated in Land Surface Models (LSMs), but snow and ice albedo parametrizations differ widely in their complexity [5]. A Snow Model Intercomparison was conducted using 24 snow cover models developed in ten different countries [17]. The models differ as to being multilayer or not, the inclusion of a soil model, variable heat conductivity, variable snow density, and the treatment of liquid storage. Only four of the models met all five criteria.

27 atmospheric general circulation models (GCMs) were run under the auspices of the Atmospheric Model Intercomparison Project (AMIP). The AMIP models reproduce a seasonal cycle of snow extent similar to the observed cycle. However, GCMs tend to underestimate autumn and winter snow extent (especially over North America) and overestimate spring snow extent (especially over Eurasia). The majority of models displays less than half of the observed interannual variability. No temporal correlation is found between simulated and observed snow extent, even when only months with extremely high or low values are considered [20]. The second generation AMIP-II simulations gave better results [21].

Floating Ice

Lake ice formation is dependent on the density characteristics of fresh water, which reaches a maximum density at 4°C. As a water body cools in the autumn it becomes isothermal at 4°C. Further cooling of the surface allows a less dense layer to form and eventually frazil ice or sheet ice forms depending on the wind conditions. Snow accumulation on lake ice depresses the ice surface below the water level, causing the snow to become saturated and leading to the formation of white snow-ice (in contrast with the black lake water ice). In rivers the flow motion leads to frazil ice, which builds up into pancakes. A 1-D energy balance model of lake ice growth is described by Liston and Hall [47] that treats lake-ice freeze-up, break-up, total ice thickness and ice type (Fig. 2). The model is



Cryosphere Models, Figure 2
The elements of a model of lake ice [47]

forced by daily atmospheric data on precipitation, wind speed and air temperature.

Sea ice grows thermodynamically by freezing of sea water at near –1.8°C due to the salinity, by the accumulation of snow cover on its surface, and by dynamic processes such as ridging and rafting. It decays thermodynamically, by wave action and by export to areas of warmer ocean. Ice types include new ice, young ice, first year and multi-year ice (World Meteorological Organization [91]). Typical first year ice thickness in the spring in the Arctic is about 1.5–2 m; multi year unridged ice may be 3–4 m thick. Ridging produces keels that may extend to 20–30 m depth. Ice draft (below the sea surface) is measured by upward looking sonars that are moored or deployed on submarines in the Arctic. Ice extent and concentration – the fractional coverage of ice – are determined by aerial reconnaissance, and primarily by satellite remote sensing (optical, passive microwave, synthetic aperture radar, scatterometry and laser altimetry). Weekly or 10-day charts of ice conditions are produced by national operational ice services (see [94]).

Modeling sea ice in either a stand-alone model or a GCM involves the solution of the following equations [19]:

i. for momentum, to obtain the ice velocity fields;
ii. for thermodynamic processes to obtain net ice growth/melt; and
iii. conservation equations including deformation ands transport of ice, plus the thermodynamic sources and sinks.

Ice dynamics is based on five stresses: wind stress, water stress, internal ice stress, Coriolis force, and the stress from the tilt of the sea surface. The Coriolis force and the tilt term are an order of magnitude less than the other three terms. The air and water stresses assume a constant turning angle of 25° in the Arctic and –25° in the Antarctic [28]. Internal ice stress is highly variable depending on ice conditions. It can be negligible when the ice cover is not compact and there are "free-drift" conditions, but it can be the largest force when there is thick, compact ice cover. The force due to ice resistance to deformation involves the relationship between stress and strain rate, which is termed the rheology (Flato [19]). Early work assumed that stress is linearly dependent on strain rate as in a linear viscous fluid [12]. Pritchard et al. [71] used an elastic-plastic rheology where the stress is linearly dependent on strain up to a yield strength where failure occurs. Hibler [27] developed a viscous-plastic model with an elliptical yield curve; the pre-yield stress states are linearly related to the strain rate. Advances by Hunke and Dukowicz [35] address the response of the ice on the timescales associated with wind forcing through an elastic viscous-plastic (EVP) rheology. The model was modified so that it reduces to the viscous–plastic model at these timescales, whereas at shorter timescales the adjustment process takes place by a mathematically efficient elastic wave mechanism. Recently, Lagrangian [33,45] sea ice models using a granular rheology have been developed (Tremblay and Mysak [83]; Overland et al. [65]). They have advantages in that they model individual sea ice "floes", but are also computationally intensive and are still in their infancy.

Ice dynamics have been extensively treated by Hibler [27]. He couples the dynamics to the ice thickness characteristics by allowing the ice interaction to become stronger as the ice becomes thicker and/or contains a lower area percentage of thin ice. The dynamics in turn causes high/low oceanic heat losses in regions of ice divergence/convergence. The ice is considered to interact in a plastic manner with the plastic strength depending on the ice thickness and concentration. These in turn evolve according to continuity equations that include changes in ice mass and percent of open water due to advection, ice deformation and thermodynamic effects. Anisotropic dynamic behavior of sea ice has also been investigated [13,30], though such approaches are computationally intensive and currently are not commonly used in models. The standard model treats sea ice as a visco–plastic material that flows plastically under typical stress conditions but behaves as a linear viscous fluid where strain rates are small and the ice becomes nearly rigid. The standard viscous–plastic model has poor dynamic response to forcing on a daily timescale. Models do not generally account for high-frequency (sub-daily) inertial and tidal effects on dynamics, though research has shown that such effects can be important in the evolution of the ice cover [26,43]. The thermodynamics and dynamics are coupled through the ice thickness distribution. Essentially, deformation leads to pressure ridging and the formation of open water areas while thermodynamic processes act to ablate ridges and remove open water by ice formation in winter and create thinner ice/open water in summer. Thus, deformation acts to spread out the thickness distribution by promoting thick and thin ice categories while thermodynamic processes work towards a central ice thickness value [28].

Sea ice models typically feature processes of ice thermodynamics and dynamics although the earliest studies essentially used only thermodynamic ice growth and decay. The steady state Stefan relationship is written after Hibler and Flato [29]:

$$\rho_{\mathrm{I}} L \frac{\mathrm{d}H}{\mathrm{d}t} \approx \frac{k_{\mathrm{i}}}{H}(T_{\mathrm{m}} - T_{\mathrm{B}}) \, ,$$

where L = latent heat of fusion, T_{m} = melting point of the ice, T_{B} = upper boundary temperature of the ice, H = ice thickness, k_{i} = ice conductivity and ρ_{I} = ice density. Ice growth/melt at the underside is a result of the difference between the upward ocean heat flux and the heat conducted away from the ocean/ice interface into the ice. The first 1-D model of sea ice thermodynamics was developed by Maykut and Untersteiner (1971) [51]. A fuller treatment was made by Parkinson and Washington [66]. The model had four layers – ice, snow, ocean, and atmosphere – and 200 km horizontal resolution.

The incorporation of detailed thermodynamic processes includes the presence of snow on the sea ice, leads and polynyas, melt ponds, the effect of internal brine-pocket melting on surface ablation, the storage of sensible and latent heat inside the snow-ice system, and the transformation of snow into slush ice when the snow-ice interface sinks below the waterline due to the weight of snow. Models with enthalpy conservation improve the thermodynamic component of sea ice models [8]. These are starting to be included in larger-scale climate models.

An intermediate one-dimensional thermodynamic sea ice model developed by Ebert and Curry [15] includes leads and a surface albedo parametrization that interacts strongly with the state of the surface, and explicitly includes meltwater ponds (see Fig. 3). Four important positive feedback loops were identified: (1) the surface albedo feedback, (2) the conduction feedback, (3) the lead-solar flux feedback, and (4) the lead fraction feedback. The destabilizing effects of these positive feedbacks were mitigated by two strong negative feedbacks: (1) the outgoing longwave flux feedback, and (2) the turbulent flux feedback. A review of thermodynamic models is given by Steele and Flato [80].

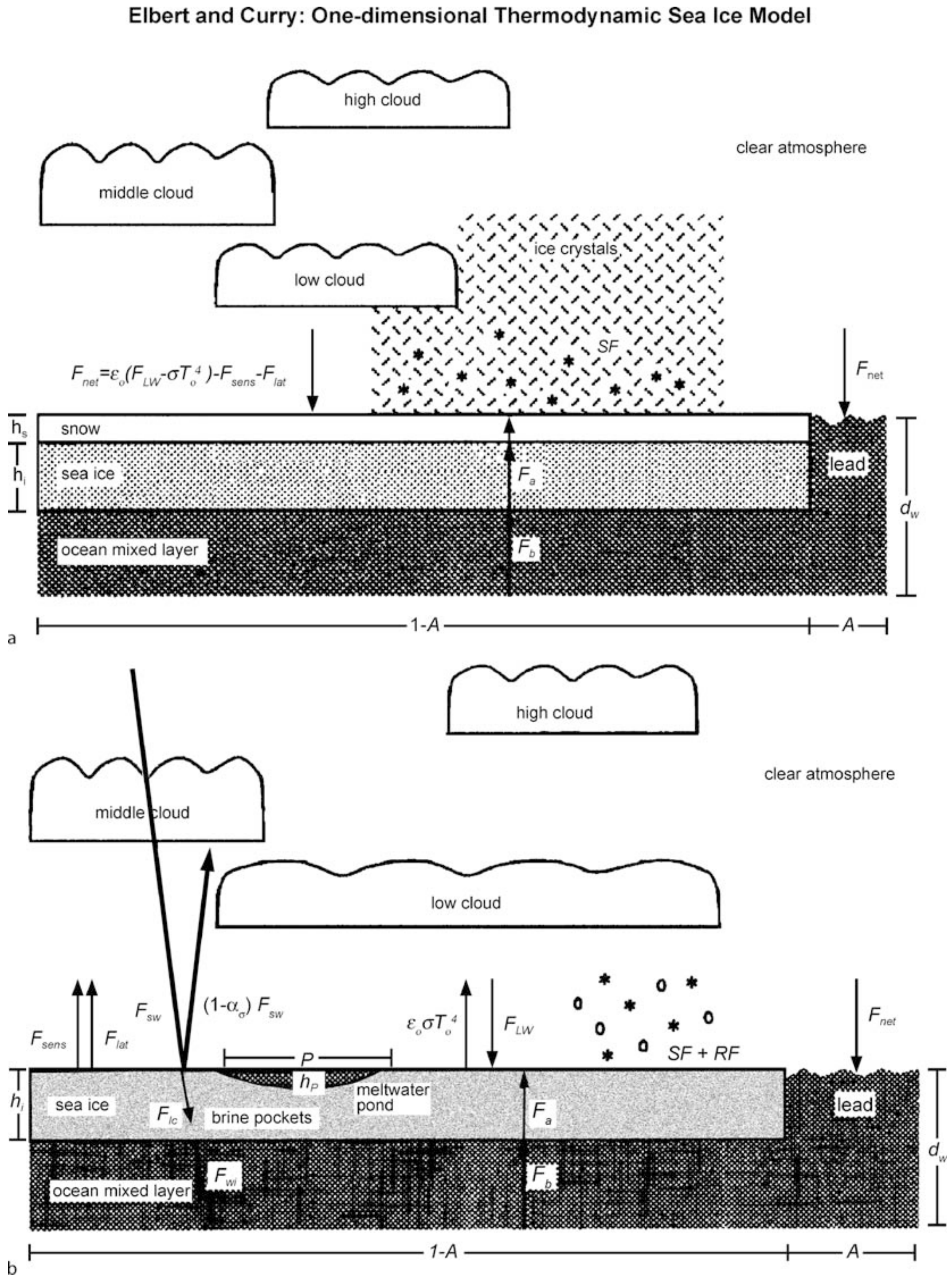
Conservation equations are needed for ice area (concentration) and ice volume (thickness).

$$\begin{aligned} \partial h/\partial t &= -\nabla \bullet (\boldsymbol{u} h) + S_h \\ \partial A/\partial t &= -\nabla \bullet (\boldsymbol{u} A) + S_A \, , \end{aligned}$$

where $\boldsymbol{u}$ is the ice velocity vector and S_h and S_A are source terms for mean ice thickness and concentration, respectively. The second equation must also have the constraint that the area $A \leq 1$. The ice strength is parametrized only in terms of ice thickness h and A [27].

An atmospheric GCM was coupled to a global 1-degree, 20-level ocean GCM with dynamic and thermodynamic sea ice by Washington and Meehl [86] and run with increasing atmospheric CO_2. The Coupled Model Intercomparison Project (CMIP) allows a comparison of predicted Arctic sea ice [53]. Of the 12 models, only seven include sea ice motion and only four of these have a prognostic solution to the momentum equation. Apart from errors and approximations in the sea ice representation, the models also suffer from errors in the atmospheric and oceanic forcing fields. While the northern hemisphere ice extent in winter is well simulated overall, the ice thickness does not capture the proper spatial distribution with thicker ice toward North America and Greenland and thinner ice in the Eurasian basin. The simulations for the southern hemisphere show a wider range of extents and thickness. Flato [19] examines the sea ice extent simulated by two GCMs for AD 1900–2100 with the 'business as usual' scenario of greenhouse gases and aerosol concentrations. Both show a progressive decrease in ice in both hemispheres although the two models differ significantly in the initial southern hemisphere ice extent.

Martin and Gerdes [50] make a comparison of sea ice drift results from different Arctic Ocean Model Intercomparison Project (AOMIP) sea ice-ocean coupled models and observations for 1979–2001. The models are capable of reproducing realistic drift pattern variability. However, one class of models has a realistic mode at drift speeds around 3 cm s^{-1} and a short tail toward higher speeds. Another class shows unrealistically a more even frequency distribution with large probability of drift speeds of 10 to 20 cm s^{-1}. Reasons for these differences lie in discrepancies of wind stress forcing as well as sea ice model characteristics and sea ice-ocean coupling. Hunke and Holland [36] underscore the sensitivity of Arctic sea ice and ocean to small changes in forcing parameters. A comparison of three sets of forcing data, all variants of



Cryosphere Models, Figure 3
The configuration of a one-dimensional thermodynamic model of sea ice **a** winter; **b** summer [15]

National Centers for Environmental Prediction (NCEP) forcing, give significant differences in ice thickness and ocean circulation using a global, coupled, sea ice-ocean model.

A study of GCMs used for the IPCC Fourth Assessment Report shows that while they produce reasonably similar ice extents in the Arctic, their equilibrium ice thickness values have a wide range due to differences in downwelling infrared radiation [16]. Holland et al. [32] found that in some scenarios of future CO_2 concentrations the sea ice cover can respond non-linearly with large decreases in extent within only 5–10 years, indicating that the current fitted to observations linear trends may not hold in the future. Stroeve et al. [81] showed that the IPPC models substantially underestimate the observed decline in Arctic sea ice extent compared to observations over the past 50 years. Hence their application in future scenarios is questionable.

Johnson et al. [40] examine the simulated sea ice concentration from nine ice-ocean numerical models in the AOMIP. The models have similar characteristics in winter (100% cover is produced), and most models reproduce an observed minimum in sea ice concentration for September 1990.

An assessment of coupled climate models with respect to the development of Arctic sea ice thickness during the 20th century is examined by Gerdes and Koeberle [22]. Model behavior is compared with results from an ocean–sea ice model using the AOMIP atmospheric forcing for the period 1948–2000. The hindcast exhibits virtually no trend in Arctic ice volume over its integration period 1948–2000. Most of the coupled climate models show a negative trend over the 20th century that accelerates towards the end of that century.

Glaciers

Glaciers are built up from snow that persists over many years. Initial densification leads to firn (densities of 400–830 kg m^{-3}) and at some depth, where the air passages between grains are sealed off (∼15–70 m according to wetness), to glacier ice with a density of 830–917 kg m^{-3} [67]. The glacier has upper accumulation and lower ablation areas, that are annually varying, and the ice slowly flows downhill towards the glacier terminus. Some glaciers occasionally display surges when the ice advances rapidly for a year or two and then stabilizes or retreats.

Glacier models consider either the mass balance and the rate of change of total mass, or the glacier dynamics and interactions between the ice and the bed. The flow velocity is modeled along the centerline of the glacier.

Glacier flow is determined from a relation between the shear strain rate (ε_{xy}) and shear stress (τ_{xy}) known as Glen's flow law [23]:

$$\overset{\bullet}{(\varepsilon_{xy})} = A\tau_{xy}^{n} ,$$

where $n \sim 3$. A depends on ice temperature, impurities and crystal orientation.

Recommended values of A decrease from 6.8×10^{15} s^{-1} kPa^{-3} at 0°C to 3.6×10^{-18} at −50°C (see Table 5.2 in [67]). Stress causes ice to deform by extension/compression, and shear leading to rotation.

Ice flows only by internal deformation when the bed is frozen, but where temperate conditions exist at the base, sliding becomes important. The sliding law relates basal velocity, shear stress, water pressure and the glacier bed characteristics. Weertman's [88] theory of sliding involved regelation and plastic deformation. Regelation operates over small bumps in the bed (<1 m dimension). All the ice is at pressure melting point. There is excess pressure on the upstream side of the bump so that the ice there is colder than on the downstream side. This causes heat to flow towards the upstream side through the bump and surrounding ice. The heat transferred melts ice on the upstream side and melt water flows around the bump, refreezing on the downstream side because it is colder than the ice there. Ice also deforms plastically. Near a bump, the longitudinal stress in the ice and, therefore, the strain are above average. The greater the distance over which the stress is enhanced, the greater is the ice velocity. This mechanism works best over larger bumps. Both processes are equally effective at the "controlling obstacle" size, about 1–10 cm.

Paterson [67] shows that the sliding velocity:

$$u = \text{constant}\,(\tau^{0.5}/R)^{4} ,$$

where R = roughness and τ = basal shear stress.

Water from surface ablation penetrates to the glacier bed and has been shown to lift the ice by as much as 40 cm on the Unteraargletscher, Switzerland (Iken et al. [38]). During rapid uplift events the glacier velocity increased 3–6 times. When the water pressure at the bed exceeds a certain value (the separation pressure) that depends on the bed roughness, cavities form in the lee of bumps. When the water pressure exceeds a second critical value, sliding becomes unstable.

A numerical ice flow model has recently been used to study the advance of tidewater glaciers into a deep fiord [56]. The results suggest that irrespective of the calving criterion and the accumulation rate in the catchment, the glacier cannot advance into deep water (>300 m) unless sedimentation at the glacier front is included.

Using a first-order theory of glacier dynamics, Oerlemans [61] related changes in glacier length to changes in air temperature. He constructed a temperature history for different parts of the world from 169 records of glacier length. The reconstructed warming in the first half of the 20th century is 0.5°C. The warming signals from glaciers at low and high elevations appear to be very similar.

Ice Sheets

There are currently two major ice sheets in Greenland and Antarctica, but during the Last Glacial Maximum about 20,000 years ago there were massive ice sheets over northern North America and Fennoscandia. The Antarctic ice sheet covers 12.4 million km^2 and reaches thicknesses in excess of 4500 m; it represents a sea level equivalent of 64.3 m. The Greenland ice sheet has an area of 1.8 million km^2 and a maximum thickness of 3200 m; it has a sea level equivalent of 7.5 m. Both ice sheets have a parabolic profile and gentle slopes (1–2°) away from the margins. Greenland has a suite of snow facies; from the interior outward these are: the dry-snow zone, percolation zone, wet-snow zone, superimposed ice zone, and the ablation area. The equilibrium line is between the ablation area and the higher zones. At its base, Greenland is close to sea level except for fringing coastal mountains, through which the ice reaches the sea in some 20 major outlet glaciers. Antarctica is mostly dry snow and is little affected by melting except at the margins and in the Antarctic Peninsula. In the Antarctic there are major ice shelves (e. g. the Ross Ice Shelf and the Filchner–Ronne Ice Shelf) that buttress large sections of the ice sheet (see Fig. 4) and extend around 44% of the coastline of the continent.

The major problems in ice sheet dynamics, following Paterson (see p. 238 in [66]), are (i) to calculate the distribution of ice thickness and velocity that will maintain a steady state, given the accumulation and ablation rates and surface temperature. The flow parameters, ice thickness at the ice divide, and geothermal heat flux must be specified. Flow lines, the time that ice takes to travel along them, and the age of the ice at different depths, can be calculated from the velocity distribution. (ii) to determine how the system will react to changes in accumulation, ablation or surface temperature.

The earliest ice sheet models assumed that ice deformed as a Newtonian viscous body. Orowan [64] and Nye [59] assumed that ice behaved as a perfectly plastic material, but Glen [23] established the relationship between strain rate and stress in ice as non-linear (see above).

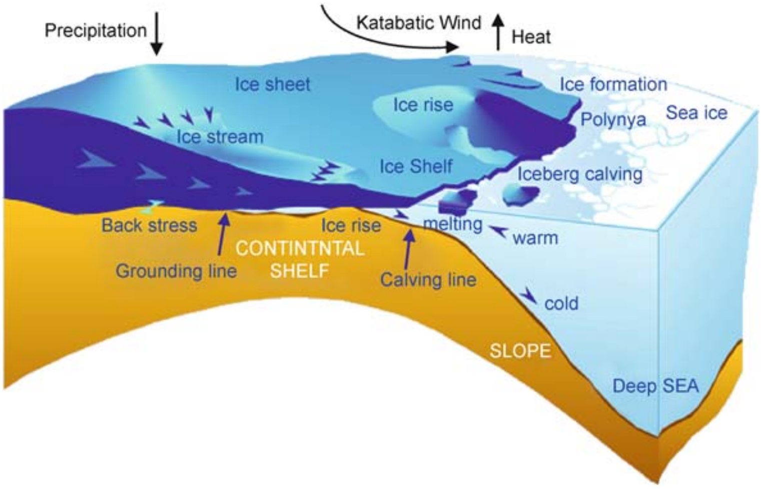
The surface profile of an ice sheet on a horizontal bed of half width L, thickness h and thickness at the centre H is:

$$h^2 = \frac{2\tau_0}{\rho g}(L - x) ,$$

where ρ = density, g = acceleration due to gravity, τ_0 = the basal shear stress (∼0–100 kPa), and $(L - x)$ is the distance from the edge measured along a flow line. The equation describes a parabola. The ice thickness at the centre is $H = (2\tau_0 L/\rho g)^{0.5}$.

The mass balance (B) can be expressed in a mass conservation equation as:

$$B = \frac{\partial q}{\partial x} : + \frac{\partial h}{\partial t} ,$$



Cryosphere Models, Figure 4
Schematic of an ice sheet and shelf showing the processes at work (after Wikipedia)

where h = ice thickness, t = time, and the flux $q = h\,\boldsymbol{u}$, where $\boldsymbol{u}$ is the velocity averaged over the ice thickness (see p. 246 in Paterson [67]).

If the flow term $\partial q/\partial x$ is small, then the surface elevation will vary in response to the local accumulation/ablation, which will determine the profile. 'Balance velocities' are steady state velocities that are calculated from accumulation rate and ice thickness. Paterson [67] shows that for a 1000 km radius circular ice sheet, of perfectly plastic ice with a yield stress of 199 kPa, an accumulation rate of 150 mm of ice/year, and ablation by iceberg calving at the margin, the ice in the centre would be 4700 m thick. Balance velocities would increase from 1,5 m a^{-1} at 100 km from the centre to 45 m a^{-1} at 900 km and the travel time for ice to move from the center to the edge would be 150,000 years.

In Greenland and Antarctica much of the ice transport is accomplished by fast flowing ice streams – regions where the ice flow is much faster than on either side. Most occupy deep channels with beds below sea level and terminate either as a floating glacier tongue (an outlet glacier) or become part of an ice shelf. Morgan et al. [54] indicate that while ice streams and outlet glaciers account for only 13% of the coastline of Antarctica, they drain about 90% of the accumulation of the interior.

A hierarchy of land ice models is presented by van der Veen and Payne [85]. The simple lamellar flow model, involves a balance between driving stress and basal drag. The surface and bed topography must be nearly level for lamellar flow, which is a good approximation to conditions in the interior of an ice sheet. In cases where an ice stream is bounded by a rock wall or stagnant ice on one or both sides, lateral drag needs to be incorporated. The proportion of driving stress that is supported by drag at the bed is termed the shape factor [60]; it is less than one for narrow ice streams. An important issue in Antarctica is the interaction between the ice sheet and ice shelves. The peripheral ice shelves are thought to exert a back stress that stabilizes the inland ice sheet where it is grounded below sea level, as in most of West Antarctica [82]. Where the bedrock slopes down towards the ice sheet interior, the grounding line is unstable; If the grounding line initially retreats, the ice at the grounding line becomes thicker due to the bedrock slope, and the creep thinning (thinning associated with along-flow gradients in the ice velocity) increases causing the grounding line to retreat further – a positive feedback. The dynamics of ice sheet grounding lines is examined by Schoof [76]. A boundary layer theory for ice flux through the transition zone shows that the flux increases sharply with ice thickness at the grounding line. He finds that marine ice sheets have well-defimed, discrete equilibrium profiles, and steady grounding lines cannot be stable on reverse bed slopes. Also, marine ice sheets with overdeepened beds may undergo hysteresis with variations in sea level, accumulation rate, bed slipperiness and ice viscosity.

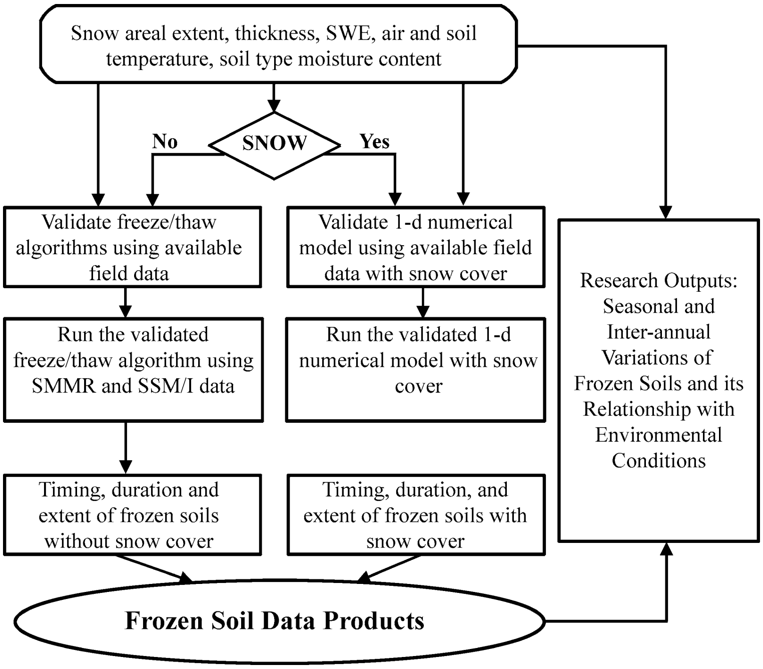
Two general types of an ice sheet model have been developed. One is prognostic, based on the original work by Budd et al. [11]; the other category is diagnostic, addressing specific aspects of ice sheet processes. Prognostic models involve four sets of equations (van der Veen and Payne [85]). These are: (i) diagnostic equations for the horizontal velocity components as functions of local ice geometry and ice rheology (Glen's law); (ii) prognostic equations for the evolution of internal ice temperature, given appropriate boundary conditions at the upper and lower ice surfaces; (iii) a diagnostic equation for ice vertical velocity via the divergence of the horizontal velocity; and (iv) a prognostic equation for ice thickness based on the snow accumulation, snow/ice melt and the divergence/convergence of horizontal ice flow. The effects of bedrock depression under the changing weight of the ice load must also be taken into account. Such models have been used to reconstruct ice sheet history over glacial cycles, as well as to assess the responses to future climate change.

Diagnostic models do not address time evolution of the ice sheet and treat the internal stress regime in much greater detail, particularly the contributions of longitudinal and lateral stresses. Recently, models have been developed that do not assume negligible vertical shear. Huybrechts and de Wolde [38] have combined prognostic model elements with a detailed diagnostic model to study the multi-century behaviour of the Antarctic and Greenland ice sheets. A fully dynamic 3-D thermo-mechanical ice sheet model was coupled to a two-dimensional climate model.

A model validation exercise was undertaken by the European Ice Sheet Modeling Initiative (EISMINT). Payne et al. [68] examined the effects of thermo-mechanical coupling while MacAyeal et al. [49] test ice shelf models for the Ross Ice Shelf. Overall, the models agreed in the main features that were simulated.

Frozen Ground and Permafrost

The surface layers of soil and rock may be seasonally or perennially frozen. Perennially frozen ground or permafrost is frozen for at least two successive summers. The ground need not contain ice but may be rock below 0°C. Ground ice may be segregated, in veins, or massive in occurrence. The spatial extent of permafrost ranges from


Snow areal extent, thickness, SWE, air and soil temperature, soil type moisture content
SNOW
No
Yes
Validate freeze/thaw algorithms using available field data
Validate 1-d numerical model using available field data with snow cover
Run the validated freeze/thaw algorithm using SMMR and SSM/I data
Run the validated 1-d numerical model with snow cover
Timing, duration and extent of frozen soils without snow cover
Timing, duration, and extent of frozen soils with snow cover
Research Outputs: Seasonal and Inter-annual Variations of Frozen Soils and its Relationship with Environmental Conditions
Frozen Soil Data Products

Cryosphere Models, Figure 5
The flowchart of a model processing the effects of snow cover on frozen soil, used to study the timing, duration, number of days, and areal extent of near-surface soil freeze/thaw status [92]

continuous (>90% of the surface is underlain), to discontinuous (50–90% of the surface), sporadic (10–50%) and isolated (<10%). Continuous permafrost is associated with mean annual temperatures below about –7°C. Its thickness ranges from a few meters up to 1500 m in Yakutia. Subsea permafrost also occurs offshore in the Eurasian shelf seas and in the Beaufort Sea; it is a relic of previous glacial intervals when the sea bed was exposed by sea level lowering of up to 135 m under low temperature conditions.

Ground temperatures are largely determined by heat conduction, although in areas of seasonal freezing and discontinuous permafrost localized circulation of groundwater may need to be considered. The thermal properties of the ground vary with the mineral composition, organic content, moisture content (as vapor, water and ice), and temperature, as well as the overlying vegetation and snow cover. A frozen soil algorithm has been developed [92] to detect the near-surface soil freeze/thaw cycle over snow-free and snow-covered land in the United States (see Fig. 5).

The conductive heat transfer is given [89] as:

$$G = -K(\mathrm{d}T/\mathrm{d}z)$$

where K = the thermal conductivity ($W\ m^{-1}\ K^{-1}$).

For steady state conditions, the temperature at depth T_z is written:

$$T_z = T_s + (G/K)z$$

where T_s = surface temperature.

The heat conduction equation is:

$$\partial T/\partial t = \kappa \partial^2 T/\partial z^2$$

where κ = thermal diffusivity ($\mathrm{m}^2\ \mathrm{s}^{-1}$), the coefficient of heat diffusion. To simulate soil freezing and thawing processes, soil water phase change has to be considered (for details, see Lunardini [48]).

Permafrost models can be broadly classified as either equilibrium or process-based transient models according to their underlying methodology. Equilibrium models are based on empirical and semi-empirical relationships between permafrost occurrence and topoclimatic factors (altitude, slope, aspect) and mean air temperature, freezing and thawing indices, and snow cover. They are often used to predict the lateral "boundaries" of permafrost distribution, or to estimate "average" geocryological parameters. thawing indices, snow cover and solar radiation, for example [77]. A 'frost index' model was developed by Nelson and Outcalt [55] for the Arctic and has been applied to mountain areas [31]. A wide spectrum of explicit equilibrium permafrost models is available to estimate the thickness of the active layer. The simplest approaches are based on several variations of the analytical Stefan solution to the heat conduction problem with phase change. These methods have been used to estimate regional-scale active layer thickness [56,79]. Kudryavtsev et al. [42] developed a more comprehensive equilibrium permafrost model with analytical solutions that has been adapted and used with Geographical Information System (GIS) technology to estimate active layer thickness at regional [74,78] and circum-arctic [2] scales. The advantages of equilibrium models are their relative simplicity and low data requirements. The major drawback of such models is their inability to resolve seasonal and inter-annual variability [77], which is frequently required for ecological and hydrological studies in the Arctic. These limitations have led to recent spatial adaptations of process-based transient numerical models.

Process-oriented transient models detail one-dimensional heat transfer in soils with phase change driven by either surface temperature [24] or the energy balance components [46]. They account for the major physical processes governing development of the ground thermal regime, simulate soil freezing/thawing processes, and provide insight into the response of soil thermal regime to changes in environmental conditions. These models can provide good results for simulating active layer thickness and permafrost temperatures when driven with known boundary conditions and forcing parameters measured at site-specific locations [73,93]. Recently, such one-dimensional heat transfer models with phase change have been used to simulate regional-scale soil thermal regime [62,63]. However, their adaptation from point-specific to regional-scale is not a straightforward process. It requires simplification, careful selection of climate forcing data, and treatment of surface and subsurface parameters with largely unknown distributions over the modeled domain.

Most GCMs do not treat permafrost dynamics. However, Nikolsky et al. [58] show that in the Community Land Model (CLM3) GCM improvements can be made to the representation of permafrost dynamics and their climate feedbacks. They do this by increasing the total soil depth by adding new layers, incorporating a surface organic soil layer, and modifying the model's numerical scheme to include unfrozen water dynamics and more realistic treatment of the model phase changes between ice and water.

The Community Climate System Model (CCSM) has a 5-layer snow model over a 10 layer 3.4 m deep soil model that treats thermal and hydrologic frozen soil processes. A projection made for the 21st century shows severe degradation of the permafrost in the Northern Hemisphere.

Some models address only the active layer that is the top layer of soil that thaws during the summer and freezes again during the autumn. The thaw depth can be analyzed by the Stefan solution for heat transfer in a medium with phase change (Anisimov et al. [3]):

$$z = [(2n\lambda t T)/(\rho w L)]^{0.5}$$

where z = active layer thickness (m), n = the ratio of seasonal ground surface and air temperature degree-day sums, λ = thermal conductivity of thawed soil (W m^{-1} K^{-1}), t = warm season duration > 0°C (s), T = mean warm season temperature (°C), ρ = soil density (kg m^{-3}), w = relative water content (decimal proportion) and L = latent heat of fusion (J kg^{-1}).

A dynamic 3-D terrain model is currently being developed and tested in Svalbard (Humlum [34]). The model takes topographic data, terrain surface characteristics (geomorphology and vegetation) and meteorological variables (air temperature, wind speed and direction, and cloud cover) as input and provides output on phenomena such as terrain surface net radiation balance, snow cover thickness and duration, glacier mass, active layer thickness, stable permafrost thickness and the amount of summer melt water discharge.

A stochastic model was developed by Anisimov et al. [3] and used to calculate the probability density function of active-layer thickness (ALT). Equations for the mean, variance, and higher moments of ALT were derived by applying stochastic averaging to a semi-empirical model of seasonal thawing. The stochastic model was applied in a case study in the Kuparuk River basin, north-central Alaska.

Shiklomanov et al. [77] compare three models of active layer thickness (ALT) for northern Alaska. One model (NSIDC) is very accurate in the topographically homogeneous Coastal Plain but overestimates (ALT) in the Brooks Range Foothills. The UAF-GIPL 2.0 model reproduced site-specific active layer values well but overestimated ALT on the Coastal Plain. Large differences in ALT fields mainly result from differences in model approaches for characterizing largely unknown spatial distribution of surface (vegetation, snow) and subsurface (soil properties and moisture) conditions.

Data set limitations are a major problem (Anisimov et al. [3]). A permafrost model, forced with available climate data sets, was used to calculate the large-scale characteristics of permafrost in northern Eurasia. Zonal-mean air and ground temperatures, depth of seasonal thawing, and area occupied by near-surface permafrost in Eurasia north of 45° N were analyzed. The 0.5–1.0 °C difference in zonal-mean air temperature between the data sets translates into a 10–20% uncertainty in estimates of near-surface permafrost area, which is comparable to the extent of changes projected for the following several decades.

GCMs have been used to simulate changes in permafrost conditions with global warming. Anisimov and Nelson [1] were the first to study this. Most recently, Saito et al. [74] use a coupled global climate model at high horizontal resolution (0.5° land mesh) with a five-layer, 4.0 m deep soil to evaluate changes in the distribution of frozen ground and subsurface hydrothermal regimes under global warming. Two types of frozen ground were classified according to monthly soil temperatures: "permafrost" for regions with a maximum active layer thickness less than 4 m and "seasonally frozen ground." Approximately 60% of present-day permafrost would degrade into seasonally frozen ground by 2100 in the circum-Arctic basins.

Future Directions

In the next 5–10 years we can expect to see more comprehensive treatment of cryospheric processes in climate models. Already, steps are being taken to incorporate ice sheet processes and to enhance the treatment of frozen ground and permafrost. Increased model resolution will place new demands for cryospheric data sets for boundary conditions and as validation data. Mountain glaciers, as well as lake ice, will need to be incorporated especially in regional climate models.

Bibliography

Primary Literature

1. Anisimov OA, Nelson FE (1996) Permafrost distribution in the Northern Hemisphere under scenarios of climate change. Glob Planet Chang 14:59–72
2. Anisimov OA, Shiklomanov NI, Nelson FE (1997) Global warming and active-layer thickness: results from transient general circulation models. Glob Planet Chang 15:61–77
3. Anisimov OA, Shiklomanov NI, Nelson FE (2002) Variability of seasonal thaw depth in permafrost regions: a stochastic modeling approach. Ecol Model 153:217–227
4. Armstrong RL, Brodzik MJ, Knowles K, Savoie M (2005) Global monthly EASE-Grid snow water equivalent climatology. Digital media. National Snow and Ice Data Center, Boulder
5. Barry RG (1996) The parameterization of surface albedo for sea ice and its snow cover. Progr Phys Geog 20:61–77
6. Barry RG (2002) The role of snow and ice in the global climate system: A review. Polar Geog 24:235–246
7. Bartelt P, Lehning M (2002) A physical SNOWPACK model for the Swiss Avalanche Warning Services. Part I: Numerical model. Cold Reg Sci Technol 35(3):123–145
8. Bitz CM, Lipscomb WH (1999) An energy-conserving thermodynamic model of sea ice. J Geophys Res 105:15669–15677
9. Bovis MJ, Mears AI (1976) Statistical prediction of snow avalanche runout from terrain variables in Colorado. Arct Alp Res 8:115–120
10. Brun E, David P, Sudul M, Brunot G (1992) A numerical model to simulate snow cover stratigraphy for operational avalanche forecasting. J Glaciol 38:13–22
11. Budd WF, Jenssen D, Radok U (1971) Derived physical charcteristics of the Antarctic ice sheet. ANARE Interim Report Series A (IV) Glaciology Pobl
12. Campbell WJ (1965) The wind-driven circulation of ice and water in a polar ocean. J Geophys Res 70:3279–3301
13. Coon MD, Knoke GS, Echert DS, Pritchard RS(1998) The architecture of anisotropic elastic-plastic sea ice mechanics constitutive law. J Geophys Res 103(C10):21915–21925
14. Dozier J, Painter TH (2004) Multispectral and hyperspectral remote sensing of alpine snow properties. Annu Rev Earth Planet Sci 32:465–494
15. Ebert EE, Curry JA (1993) An intermediate one-dimensional thermodynamic sea ice model for investigating ice-atmosphere interactions. J Geophys Res 98(C6):10085–10110
16. Eisenman I, Untersteiner N, Wettlaufer JS (2007) On the reliability of simulated Arctic sea ice in global climate models. Geophys Res Lett 34:L10501, doi:10.1029/2007GL029914
17. Essery R, Yang Z-L (2001) An overview of models participating in the snow model intercomparison project (SnowMIP). In: 8th Scientific Assembly of IAMAS, Innsbruck. http://www.cnrm.meteo.fr/snowmip/. Accessed 22 Aug 2008
18. Essery R, Long L, Pomeroy JW (1999) A distributed model of blowing snow over complex terrain. Hydrol Process 13:2423–2438
19. Flato GM (2004) Sea-ice modelling. In: Bamber JL, Payne AJ (eds) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, pp 367–390
20. Frei A, Robinson DA (1995) Evaluation of snow extent and its variability in the Atmospheric Model Intercomparison Project. J Geophys Res 103(D8):8859–8871
21. Frei A, Miller JA, Robinson DA (2003) Improved simulations of snow extent in the second phase of the Atmospheric Model Intercomparison Project (AMIP-2). J Geophys Res 108(D12):4369, doi:10.1029/2002JD003030
22. Gerdes R, Koeberle C (2007) Comparison of Arctic sea ice thickness variability in IPCC Climate of the 20th Century experiments and in ocean–sea ice hindcasts. J Geophys Res 112(C4)C04S13
23. Glen J (1955) The creep of polycrystalline ice. Proc Roy Soc Lond A228:519–538
24. Goodrich LE (1982) The influence of snow cover on the ground thermal regime. Can Geotech J 19:421–432
25. Hedstrom N, Pomeroy JW (1998) Measurements and modelling of snow interception in the boreal forest Hydrol. Processes 12:1611–1525
26. Heil P, Hibler WD III (2002) Modeling the high-frequency component of Arctic sea ice drift and deformation. J Phys Oceanogr 32:3039–3057

27. Hibler WD III (1979) A dynamic-thermodynamic sea ice model. J Phys Oceanogr 9:815–846
28. Hibler WD III (2004) Modelling the dynamic response of sea ice. In: Bamber JL, Payne AJ (eds) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, pp 227–334
29. Hibler WD III, Flato GM (1992): Sea ice models. In: Trenberth K (ed) Climate System Modeling. Cambridge University Press, New York, pp 413–436
30. Hibler WD III, Schulson EM (2000) On modeling the anisotropic failure and flow of flawed sea ice. J Geophys Res 105(C7):17105–17120
31. Hoelzle M, Mittaz C, Etzelmueller B, Haeberli W (2001) Surface energy fluxes and distribution models of permafrost in European mountain areas: An overview of current developments. Permafr Periglac Process 12:53–68
32. Holland MM, Bitz CM, Tremblay H (2006) Future abrupt reductions in the summer Arctic sea ice. Geophys Res Lett 33:L23503. doi:10.1029/2006GL028024
33. Hopkins MA (1996) On the mesoscale interaction of lead ice and floes. J Geophys Res 101:18315–18326
34. Humlum O (2007) Modeling energy balance, surface temperatures, active layer depth and permafrost thickness around Longyeardalen, Svalbard. http://www.unis.no/research/geology/Geo_research/Ole/Modelling.htm. Accessed 22 Aug 2008
35. Hunke EC, Dukowicz JK (1997) An elastic–viscous–plastic model for sea ice dynamics. J Phys Oceanogr 27:1849–1867
36. Hunke EC, Holland MM (2007) Global atmospheric forcing data for Arctic ice-ocean modeling. J Geophys Res 112:C04S14
37. Huybrechts P, de Wolde J (1999) The dynamic response of the Greenland and Antarctic ice sheets to multiple-century climatic warming. J Climate 12:2169–2188
38. Iken A, Roethlisberger H, Flotron A, Haeberli W (1983) The uplift of the Unteraargletscher at the beginning ot the melt season – a consequence of water storage at the bed. J Glaciol 30:15–25
39. Jin J, Gao X, Yang Z-L, Bales RC, Sorooshian S, Dickinson RE, Sun SF, Wu GX (1999) Comparative analyses of physically based snowmelt models for climate simulations. J Climate 12:2643–2657
40. Johnson M, Gaffigan S, Hunke E, Gerdes R (2007) A comparison of Arctic Ocean sea ice concentration among the coordinated AOMIP model experiments. J Geophys Res 112:C04S11
41. Jordan R (1991) A one-dimensional temperature model for a snow cover. Technical documentation for SNTHERM Special Technical Report 91-16. US Army Cold Regions Research and Engineering Laboratory, Hanover
42. Kudryavtsev VA et al (1974) Fundamentals of frost forecasting in geological engineering investigations. Nauka, Moscow (in Russian). English translation US Armt Cold Regions Res Engr Lan, Hannover, Draft translation 1977
43. Kwok R, Cunningham GF, Hibler III WD (2003) Sub-daily sea ice motion and deformation from RADARSAT observations. Geophys Res Lett 30(23):2218 doi:10.1029/2003GL018723
44. Lawrence DM, Slater AG (2007) A projection of severe near-surface permafrost degradation during the 21st century. Geophys Res Lett 32:L24401
45. Lindsay RW, Stern HL (2005) A new Lagrangian model of Arctic sea ice. J Phys Oceanogr 34:272–283
46. Ling F, Zhang T-J (2004) A numerical model for surface energy balance and thermal regime of the active layer and permafrost containing unfrozen water. Cold Regions Sci Technol 38:1–15
47. Liston GE, Hall DK (1995) An energy-balance model of lake-ice evolution. J Glaciol 41(138):373–382
48. Lunardini V (1988) Freezing of soil with an unfrozen water content and variable thermal properties. US Army Cold Regions Res Engineering Lab, Hanover, p 31
49. MacAyeal DR et al (1996) An ice-shelf model test based on the Ross ice shelf. Antarct Ann Glaciol 23:46–51
50. Martin Y, Gerdes R (2007) Sea ice drift variability in Arctic Ocean Model Intercomparison Project models and observations. J Geophys Res 112(C4):C04S10
51. Maykut G, Untersteiner N (1971) Some results from a time-dependent thermodynamic mode; of sea ice. J Geophys Res 76:1550–75
52. McClung D, Schaerer P(2006) The Avalanche Handbook. The Mountaineers, Seattle
53. Meehl GA, Boer GA, Covet C, Latif M, Stouffer RJ (1997) Intercomparison makes for a better climate model. EOS 78:445–446
54. Morgan VI, Jacka TH, Akermasn GJ, Clarke AL (1982) Outlet glacier and mass budget studies in Enderby, Kemp and Mac-Robertson Lands, Antarctica. Ann Glaciol 3L:204–210
55. Nelson FE, Outcalt DSI (1987) A computational method for prediction and regionalization of permafrrost. Arct Alp Res 19:279–88
56. Nelson FE et al (1997) Estimating Active-Layer Thickness over a Large Region: Kuparuk River Basin, Alaska, USA. Arct Alp Res 29:367–378
57. Nick EM, van derr Veen CJ, Oerlemans J (2007) Controls on advance of tidewater glaciers: Results from numerical modeling applied to Columbia Glavier. J Geophys Res 112:G03S24
58. Nicolsky DJ, Romanovsky VE, Alexeev VA, Lawrence DM (2007) Improved modeling of permafrost dynamics in a GCM Land Surface Scheme. Geophys Res Lett 34(8):L08591
59. Nye J (1951) The flow of glaciers and ice sheets as a problem in plasticity. Proc Roy Soc Lond A 207:554–572
60. Nye J (1965) The flow of a glacier in a channel of rectangular, elliptic or parabolic cross-section. J Glaciol 5:661–690
61. Oerlemans J (2005) Extracting a climate signal from 169 glacier records. Science 308:675–677
62. Oelke C et al (2003) Regional-scale modeling of soil freeze/thaw over the Arctic drainage basin. J Geophys Res 108(D10):4314
63. Oelke C, Zhang T-J (2004) A model study of circum-Arctic soil temperatures. Permafr Periglac Process 15:103–121
64. Orowan E (1949) Remarks at the joint meeting of the British Glaciological Society, the British Rheologists Club and the Institute of Metals. J Glaciol 1:231–236
65. Overland JE, McNutt SL, Salo S, Groves J, Li S (1998) Arctic sea ice as a granular plastic. J Geophys Res 104(C10):21845–21867
66. Parkinson CL, Washington WM (1979) A large-scale numerical model of sea ice. J Geophys Res 84:311–337
67. Paterson WSB (1994) The physics of glaciers. Pergamon, Elsevier Science, New York, p 480
68. Payne AJ et al (2000) Results from the EISMINT Phase 2 simplofoed geometry experiments: the effevts of thermomechanical coupling. J Glaciol 46(153):227–238
69. Perla RI (1980) Avalanche release, motion, and impact. In: Colbeck SC (ed) Dynamics of snow and ice masses. Academic Press, New York, pp 397–462

70. Pomeroy JW, Parviainen J, Hedstrom N, Gray DM (1998) Coupled modelling of forest snow interception and sublimation. Hydrol Process 12:2317–2337
71. Pritchard RS, Coon M, McPhee MG, Leavitt E (1977) Winter ice dynamics in the nearshore Beaufort Sea. AIDJEX Bull.37, Applied Physics Lab, University of Washington, Seattle, pp 37–93
72. Raymond CF (1980) Temperate valley glaciers. In: Colbeck SC (ed) Dynamics of snow and ice masses. New York. Academic Press, pp 79–139
73. Romanovsky VE, Osterkamp TE, Duzbury NS (1997) An evaluation of three numerical models used in simulations of the active layer and permafrost temperature regimes. Cold Regions Sci Technol 26:195–201
74. Saito K, Kimoto M, Zhang T, Takata K, Emori S (2007) Evaluating a high-resolution climate model: Simulated hydrothermal regimes in frozen ground regions and their change under the global warming scenario. J Geophys Res 112:F02S11
75. Sazonava TS, Romanovsky V (2003) A model for regional-scale estimation of temporal and spatial variability of active layer thickness and mean annual ground temperatures. Permafr Periglac Proc 14:125–139
76. Schoof C (2007) Ice sheet grounding line dynamics: Steady states, stability, and hysteresis. J Geophys Res 112:F03S28
77. Shiklomanov NI et al (2007) Comparison of model-produced active layer fields: Results for northern Alaska. J Geophys Res 112(F2):F02S10
78. Shiklomanov NI, Nelson FE (1999) Analytic representation of the active layer thickness field, Kuparuk River Basin, Alaska. Ecol Model 123:105–125
79. Shiklomanov NI, Nelson FE (2002) Active-layer mapping at regional scales: a 13-year spatial time series for the Kuparuk region, north-central Alaska. Permafrost Periglac Proc 13:219–230
80. Steele M, Flato GM (2000) Sea ice growth and modeling: A survey. In: Lewis EL et al (eds) The freshwater budget of the Arctic. Kluwer, Dordrecht, pp 549–587
81. Stroeve J et al (2007) Arctic sea ice decline: Faster than forecast. Geophys Res Lett 34:L09501, doi:10.1029/2007GL029703
82. Thomas RH (1979) The dynamics of marine ice sheets. J Glaciol 24:167–177
83. Tremblay L-B, Mysak LA (1997) Modeling sea ice as a granular material, including the dilatancy effect. J Phys Oceanogr 27:2342–2360
84. Trujillo E, Ramirez JA, Elder KJ (2007) Topographic, meteorologic and canopy controls on the scaling characteristics if the spatial distribution of snow depth fields. Water Resour Res 43:W07409
85. van der Veen CJ, Payne AJ (2004) Modelling land-ice dynamics. In: Bamber JL, Payne AJ (eds) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, pp 169–225
86. Washington WM, Meehl GA (1996) High-latitude climate change in a global coupled ocean-atmosphere-sea ice model with increased atmospheric CO_2. J Geophys Res 101(D8):12795–12802
87. Washington WM, Semtner AJ, Parkinson C, Morrison L (1976) On the development of a seasonal change sea-ice model. J Oceanogr 6:679–685
88. Weertman J (1957) On the sliding of glaciers. J Glaciol 5:287–303
89. Williams PJ, Smith MW (1989} The frozen earth. Cambridge University Press, Cambridge, p 306
90. Winstral A, Marks D (2002) Simulating wind fields and snow redistribution using terrain-based parameters to model snow accumulation and melt over a semi-arid mountain catchment. Hydrol Process 16:3585–3603
91. World Meteorological Organization (2007) WMO sea ice nomenclature, no 269. WMO, Geneva
92. Zhang T-J, Armstrong RL, Smith J (2003) Investigation of the near-surface soil freeze-thaw cycle in the contiguous United States: Algorithm development and validation. J Geophys Res 108(D22):8860, GCP 21-1 – 21-14
93. Zhang T-J et al (2005) Spatial and temporal variability in active layer thickness over the Russian Arctic drainage basin. J Geophys Res 110:D16101
94. Joint Commission on Oceanography and Marine Meteorology (2007) http://www.ipy-ice-portal.org/. Accessed 22 Aug 2008

Books and Reviews

Bamber JL, Payne AJ (eds) (2004) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, p 644

Curvelets and Ridgelets

JALAL FADILI[1], JEAN-LUC STARCK[2]
[1] GREYC CNRS UMR 6072, ENSICAEN,
Caen Cedex, France
[2] Laboratoire Astrophysique des Interactions
Multi-échelles UMR 7158, CEA/DSM-CNRS-Universite
Paris Diderot, SEDI-SAP, Gif-sur-Yvette Cedex, France

Article Outline

Glossary

WT1D The one-dimensional Wavelet Transform as defined in [53]. See also ▶ Numerical Issues When Using Wavelets.

WT2D The two-dimensional Wavelet Transform.

Discrete ridgelet trasnform (DRT) The discrete implementation of the continuous Ridgelet transform.

Fast slant stack (FSS) An algebraically exact Radon transform of data on a Cartesian grid.

First generation discrete curvelet transform (DCTG1) The discrete curvelet transform constructed based on the discrete ridgelet transform.

Second generationdiscrete curvelet transformx (DCTG2) The discrete curvelet transform constructed based on appropriate bandpass filtering in the Fourier domain.

Anisotropic elements By anistropic, we mean basis elements with elongated effective support; i. e. length > width.

Parabolic scaling law A basis element obeys the parabolic scaling law if its effective support is such that width ≈ length^2.

Definition of the Subject

Despite the fact that wavelets have had a wide impact in image processing, they fail to efficiently represent objects with highly anisotropic elements such as lines or curvilinear structures (e. g. edges). The reason is that wavelets are non-geometrical and do not exploit the regularity of the edge curve.

The Ridgelet and the Curvelet [16,17] transforms were developed as an answer to the weakness of the separable wavelet transform in sparsely representing what appears to be simple building atoms in an image, that is lines, curves and edges. Curvelets and ridgelets take the form of basis elements which exhibit high directional sensitivity and are highly anisotropic [9,18,32,68]. These very recent geometric image representations are built upon ideas of multiscale analysis and geometry. They have had an important success in a wide range of image processing applications including denoising [42,64,68], deconvolution [38,74], contrast enhancement [73], texture analysis [2], detection [44], watermarking [78], component separation [70,71], inpainting [37,39] or blind source separation [6,7]. Curvelets have also proven useful in diverse fields beyond the traditional image processing application. Let's cite for example seismic imaging [34,42,43], astronomical imaging [48,66,69], scientific computing and analysis of partial differential equations [13,14]. Another reason for the success of ridgelets and curvelets is the availability of fast transform algorithms which are available in non-commercial software packages following the philosophy of reproducible research, see [4,75].

Introduction

Sparse Geometrical Image Representation

Multiscale methods have become very popular, especially with the development of the wavelets in the last decade. Background texts on the wavelet transform include [23,53,72]. An overview of implementation and practical issues of the wavelet transform can also be found in ► Numerical Issues When Using Wavelets.

Despite the success of the classical wavelet viewpoint, it was argued that the traditional wavelets present some strong limitations that question their effectiveness in higher-dimension than 1 [16,17]. Wavelets rely on a dictionary of roughly isotropic elements occurring at all scales and locations, do not describe well highly anisotropic elements, and contain only a fixed number of directional elements, independent of scale. Following this reasoning, new constructions have been proposed such as the ridgelets [9,16] and the curvelets [17,18,32,68]. Ridgelets and curvelets are special members of the family of multiscale orientation-selective transforms, which has recently led to a flurry of research activity in the field of computational and applied harmonic analysis. Many other constructions belonging to this family have been investigated in the literature, and go by the name contourlets [27], directionlets [76], bandlets [49,62], grouplets [54], shearlets [47], dual-tree wavelets and wavelet packets [40,46], etc.

Throughout this paper, the term 'sparsity' is used and intended in a weak sense. We are aware that practical images and signals may not be supported in a transform domain on a set of relatively small size (sparse set). Instead, they may only be compressible (nearly sparse) in some transform domain. Hence, with a slight abuse of terminology, we will say that a representation is sparse for an image within a certain class, if it provides a compact description of such an image.

Notations

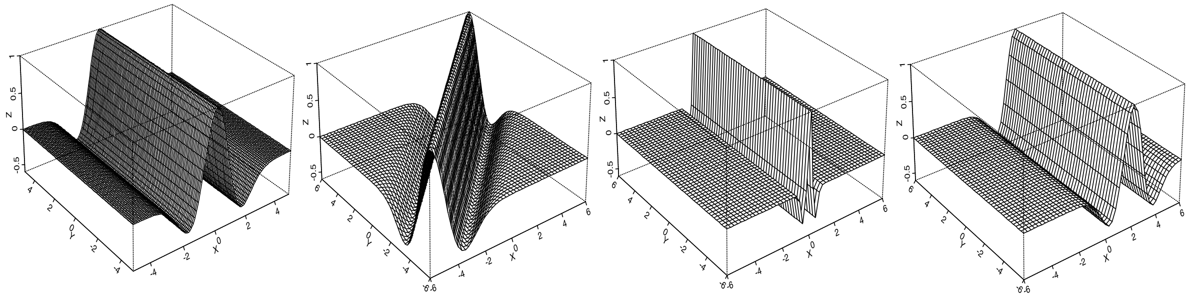
We work throughout in two dimensions with spatial variable $\mathbf{x} \in \mathbb{R}^2$ and $\boldsymbol{\nu}$ a continuous frequency-domain variable. Parentheses (., .) are used for continuous-domain function evaluations, and brackets [., .] for discrete-domain array indices. The hat ˆ notation will be used for the Fourier transform.

Ridgelets

The Continuous Ridgelet Transform

The two-dimensional continuous ridgelet transform in $\mathbb{R}^2$ can be defined as follows [10]. We pick a smooth univariate function $\psi : \mathbb{R} \to \mathbb{R}$ with sufficient decay and satisfying the admissibility condition

$$\int |\hat{\psi}(\nu)|^2/|\nu|^2 \, d\nu < \infty \,, \qquad (1)$$



Curvelets and Ridgelets, Figure 1

Few Ridgelets examples – The second to fourth graphs are obtained after simple geometric manipulations of the first ridgelet, namely rotation, rescaling, and shifting

which holds if, say, ψ has a vanishing mean $\int \psi(t)\mathrm{d}t = 0$. We will suppose a special normalization about ψ so that $\int_0^\infty |\hat{\psi}(\boldsymbol{\nu})|^2 \boldsymbol{\nu}^{-2}\mathrm{d}\boldsymbol{\nu} = 1$.

For each scale $a > 0$, each position $b \in \mathbb{R}$ and each orientation $\theta \in [0, 2\pi)$, we define the bivariate *ridgelet* $\psi_{a,b,\theta} : \mathbb{R}^2 \to \mathbb{R}$ by

$$\begin{aligned}\psi_{a,b,\theta}(\mathbf{x}) &= \psi_{a,b,\theta}(x_1, x_2)\\ &= a^{-1/2} \cdot \psi((x_1 \cos\theta + x_2 \sin\theta - b)/a)\,;\end{aligned} \tag{2}$$

A ridgelet is constant along lines $x_1 \cos\theta + x_2 \sin\theta = \text{const}$. Transverse to these ridges it is a wavelet. Figure 1 depicts few examples of ridgelets. The second to fourth panels are obtained after simple geometric manipulations of the ridgelet (left panel), namely rotation, rescaling, and shifting.

Given an integrable bivariate function $f(\mathbf{x})$, we define its ridgelet coefficients by

$$\mathcal{R}_f(a, b, \theta) := \langle f, \psi_{a,b,\theta}\rangle = \int_{\mathbb{R}^2} f(\mathbf{x})\overline{\psi}_{a,b,\theta}(\mathbf{x})\,\mathrm{d}\mathbf{x}\,.$$

We have the exact reconstruction formula

$$f(\mathbf{x}) = \int_0^{2\pi}\int_{-\infty}^{\infty}\int_0^{\infty} \mathcal{R}_f(a, b, \theta)\psi_{a,b,\theta}(\mathbf{x})\frac{\mathrm{d}a}{a^3}\mathrm{d}b\frac{\mathrm{d}\theta}{4\pi} \tag{3}$$

valid almost everywhere for functions which are both integrable and square integrable. This formula is stable and one can prove a Parseval relation [16].

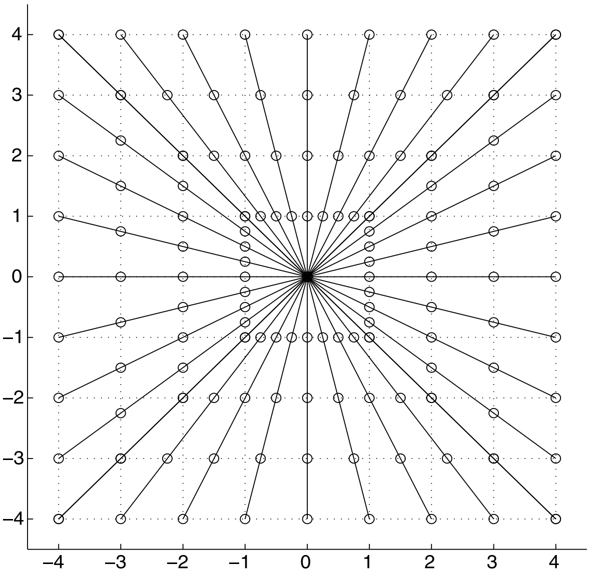
Ridgelet analysis may be constructed as wavelet analysis in the Radon domain. The rationale behind this is that the Radon transform translates singularities along lines into point singularities, for which the wavelet transform is known to provide a sparse representation. Recall that the Radon transform of an object f is the collection of line integrals indexed by $(\theta, t) \in [0, 2\pi) \times \mathbb{R}$ given by

$$Rf(\theta, t) = \int_{\mathbb{R}^2} f(x_1, x_2)\delta(x_1 \cos\theta + x_2 \sin\theta - t)\,\mathrm{d}x_1\mathrm{d}x_2\,, \tag{4}$$

where δ is the Dirac distribution. Then the ridgelet transform is precisely the application of a 1-dimensional wavelet transform to the slices of the Radon transform where the angular variable θ is constant and t is varying. Thus, the basic strategy for calculating the continuous ridgelet transform is first to compute the Radon transform $Rf(t, \theta)$ and second, to apply a one-dimensional wavelet transform to the slices $Rf(\cdot, \theta)$. Several digital ridgelet transforms (DRTs) have been proposed, and we will describe three of them in this section, based on different implementations of the Radon transform.

The RectoPolar Ridgelet Transform A fast implementation of the Radon transform can be proposed in the Fourier domain, based on the projection-slice-theorem. First the 2D-FFT of the given image is computed. Then the resulting function in the frequency domain is to be used to evaluate the frequency values in a polar grid of rays passing through the origin and spread uniformly in angle. This conversion from Cartesian to Polar grid could be obtained by interpolation, and this process is well known by the name gridding in tomography. Given the polar grid samples, the number of rays corresponds to the number of projections, and the number of samples on each ray corresponds to the number of shifts per such angle. Applying one dimensional inverse Fourier transform for each ray, the Radon projections are obtained.

The above described process is known to be inaccurate due to the sensitivity to the interpolation involved. This implies that for a better accuracy, the first 2D-FFT employed should be done with high-redundancy.



Curvelets and Ridgelets, Figure 2
Illustration of the pseudo-polar grid in the frequency domain for an *n* by *n* image ($n = 8$)

An alternative solution for the Fourier-based Radon transform exists, where the polar grid is replaced with a pseudo-polar one. The geometry of this new grid is illustrated in Fig. 2. Concentric circles of linearly growing radius in the polar grid are replaced by concentric squares of linearly growing sides. The rays are spread uniformly not in angle but in slope. These two changes give a grid vaguely resembling the polar one, but for this grid a direct FFT can be implemented with no interpolation. When applying now 1D-FFT for the rays, we get a variant of the Radon transform, where the projection angles are not spaced uniformly.

For the pseudo-polar FFT to be stable, it was shown that it should contain at least twice as many samples, compared to the original image we started with. A by-product of this construction is the fact that the transform is organized as a 2D array with rows containing the projections as a function of the angle. Thus, processing the Radon transform in one axis is easily implemented. More details can be found in [68].

One-Dimensional Wavelet Transform To complete the ridgelet transform, we must take a one-dimensional wavelet transform (WT1D) along the radial variable in Radon space. We now discuss the choice of the digital WT1D.

Experience has shown that compactly-supported wavelets can lead to many visual artifacts when used in conjunction with nonlinear processing, such as hard-thresholding of individual wavelet coefficients, particularly for decimated wavelet schemes used at critical sampling. Also, because of the lack of localization of such compactly-supported wavelets in the frequency domain, fluctuations in coarse-scale wavelet coefficients can introduce fine-scale fluctuations. A frequency-domain approach must be taken, where the discrete Fourier transform is reconstructed from the inverse Radon transform. These considerations lead to use band-limited wavelet, whose support is compact in the Fourier domain rather than the time-domain [28,29,68]. In [68], a specific over-complete wavelet transform [67,72] has been used. The wavelet transform algorithm is based on a scaling function ϕ such that $\hat{\phi}$ vanishes outside of the interval $[-\nu_c, \nu_c]$. We define the Fourier transform of the scaling function as a re-normalized B_3-spline

$$\hat{\phi}(\nu) = \frac{3}{2} B_3(4\nu),$$

and $\hat{\psi}$ as the difference between two consecutive resolutions

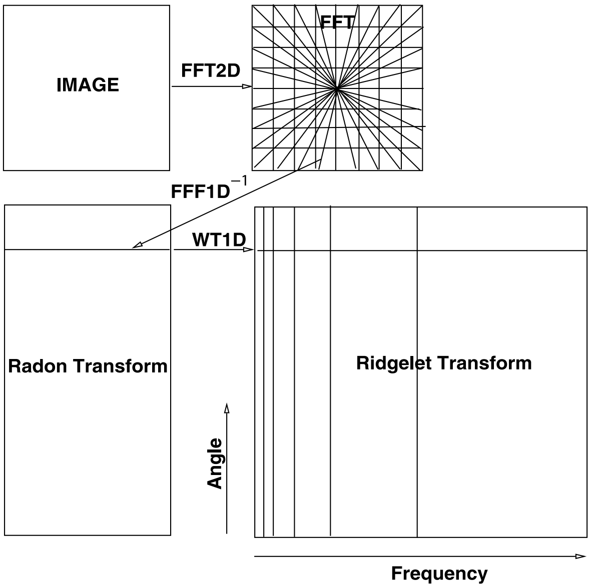
$$\hat{\psi}(2\nu) = \hat{\phi}(\nu) - \hat{\phi}(2\nu).$$

Because $\hat{\psi}$ is compactly supported, the sampling theorem shows than one can easily build a pyramid of $n + n/2 + \cdots + 1 = 2n$ elements, see [72] for details.

This WT1D transform enjoys the following useful properties:

- The wavelet coefficients are directly calculated in the Fourier space. In the context of the ridgelet transform, this allows avoiding the computation of the one-dimensional inverse Fourier transform along each radial line.
- Each sub-band is sampled above the Nyquist rate, hence, avoiding aliasing –a phenomenon typically encountered by critically sampled orthogonal wavelet transforms [65].
- The reconstruction is trivial. The wavelet coefficients simply need to be co-added to reconstruct the input signal at any given point. In our application, this implies that the ridgelet coefficients simply need to be co-added to reconstruct Fourier coefficients.

This wavelet transform introduces an extra redundancy factor. However, we note that the goal in this implementation is not data compression or efficient coding. Rather, this implementation would be useful to the practitioner whose focuses on data analysis, for which it



Curvelets and Ridgelets, Figure 3
Discrete ridgelet transform flowchart. Each of the 2*n* radial lines in the Fourier domain is processed separately. The 1-D inverse FFT is calculated along each radial line followed by a 1-D nonorthogonal wavelet transform. In practice, the one-dimensional wavelet coefficients are directly calculated in the Fourier space

is well-known that over-completeness through (almost) translation-invariance can provide substantial advantages.

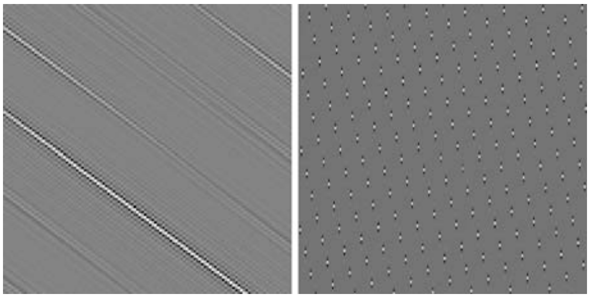
Assembling all above ingredients together gives the flowchart of the discrete ridgelet transform (DRT) depicted in Fig. 3. The DRT of an image of size $n \times n$ is an image of size $2n \times 2n$, introducing a redundancy factor equal to 4.

We note that, because this transform is made of a chain of steps, each one of which is invertible, the whole transform is invertible, and so has the exact reconstruction property. For the same reason, the reconstruction is stable under perturbations of the coefficients.

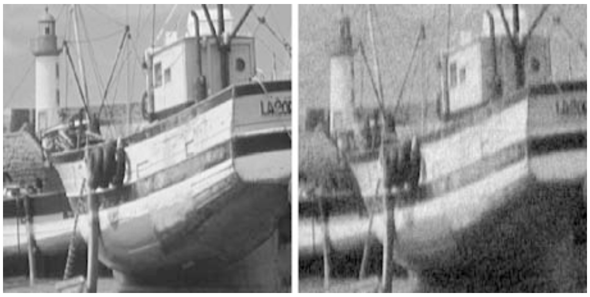
Last but not least, this discrete transform is computationally attractive. Indeed, the algorithm we presented here has low complexity since it runs in $O(n^2 \log n)$ flops for an $n \times n$ image.

The Orthonormal Finite Ridgelet Transform

The orthonormal finite ridgelet transform (OFRT) has been proposed [26] for image compression and filtering. This transform is based on the finite Radon transform [55] and a 1D orthogonal wavelet transform. It is not redundant and reversible. It would have been a great alternative to the previously described ridgelet transform if the OFRT were not based on an awkward definition of a line. In fact, a line in the OFRT is defined algebraically rather that geometrically, and so the points on a 'line' can be arbitrarily and randomly spread out in the spatial domain. Figure 4 shows the back-projection of a ridgelet coefficient by the FFT-based ridgelet transform (left) and by the OFRT (right). It is clear that the backprojection of the OFRT is nothing like a ridge function.



Curvelets and Ridgelets, Figure 4
The backprojection of a ridgelet coefficient by the FFT-based ridgelet transform (*left*), and by the OFRT (*right*)



Curvelets and Ridgelets, Figure 5
Part of original noise-free `Boat` image (*left*), and reconstruction after hard thresholding its OFRT coefficients (*right*)

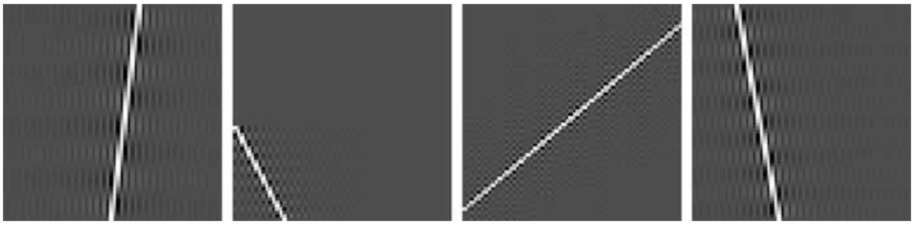
Because of this specific definition of a line, the thresholding of the OFRT coefficients produces strong artifacts. Figure 5 shows a part of the original image `Boat`, and its reconstruction after hard thresholding the OFRT of the noise-free `Boat`. The resulting image is not smoothed as one would expect, but rather a noise has been added to the noise-free image as part of the filtering !

Finally, the OFRT presents another limitation: the image size must be a prime number. This last point is however not too restrictive, because we generally use a spatial partitioning when denoising the data, and a prime number block size can be used. The OFRT is interesting from the conceptual point of view, but still requires work before it can be used for real applications such as denoising.

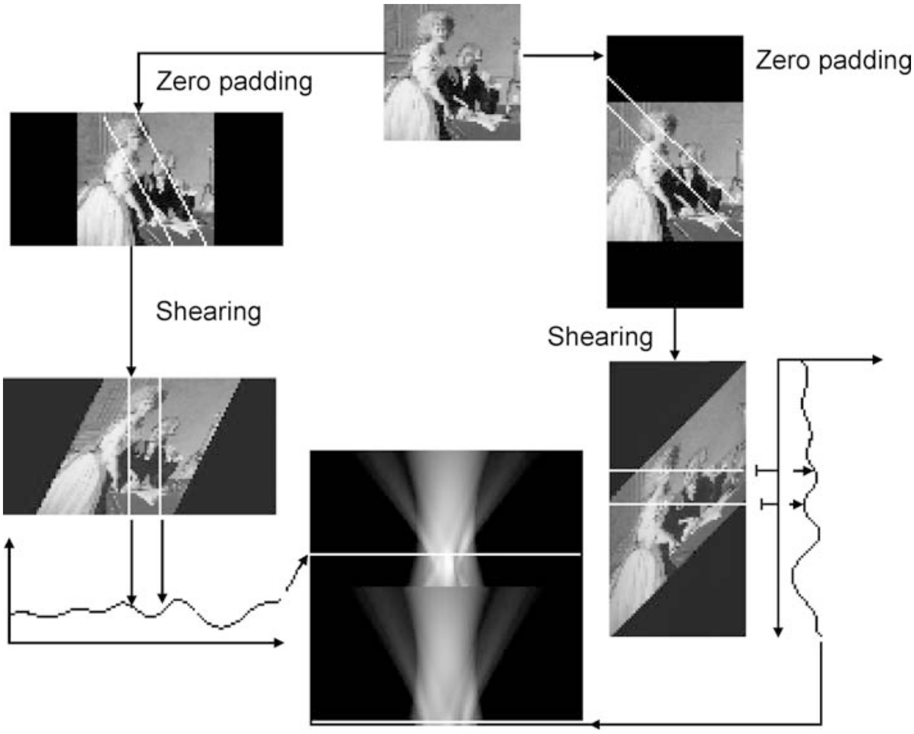
The Fast Slant Stack Ridgelet Transform

The Fast Slant Stack (FSS) [3] is a Radon transform of data on a Cartesian grid, which is algebraically exact and geometrically more accurate and faithful than the previously described methods. The back-projection of a point in Radon space is exactly a ridge function in the spatial domain (see Fig. 6). The transformation of an $n \times n$ image is a $2n \times 2n$ image. n line integrals with angle between $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$ are calculated from the zero padded image on the y-axis, and n line integrals with angle between $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$ are computed by zero padding the image on the x-axis. For a given angle inside $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$, $2n$ line integrals are calculated by first shearing the zero-padded image, and then integrating the pixel values along all horizontal lines (resp. vertical lines for angles in $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$). The shearing is performed one column at a time (resp. one line at a time) by using the 1D FFT. Figure 7 shows an example of the image shearing step with two different angles ($5\frac{\pi}{4}$ and $-\frac{\pi}{4}$). A DRT based on the FSS transform has been proposed in [33]. The connection between the FSS and the Linogram has been investigated in [3]. A FSS algorithm is also proposed in [3], based on the 2D Fast Pseudo-polar Fourier transform which evaluates the 2-D Fourier transform on a non-Cartesian (pseudo-polar) grid, operating in $O(n^2 \log n)$ flops.

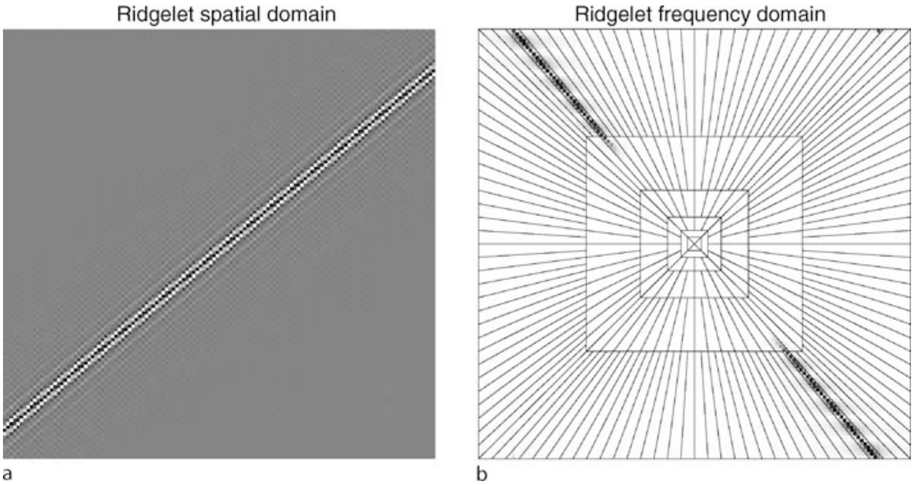
Figure 8 left exemplifies a ridgelet in the spatial domain obtained from the DRT based on FSS implementation. Its



Curvelets and Ridgelets, Figure 6
Backprojection of a point at four different locations in the Radon space using the FSS algorithm



Curvelets and Ridgelets, Figure 7
Slant Stack Transform of an image



Curvelets and Ridgelets, Figure 8
a Example of a ridgelet obtained by the Fast Slant Stack implementation. b Its FFT superimposed on the DRT frequency tiling

Fourier transform is shown on Fig. 8 right superimposed on the DRT frequency tiling [33]. The Fourier transform of the discrete ridgelet lives in an angular wedge. More precisely, the Fourier transform of a discrete ridgelet at scale j lives within a dyadic square of size $\sim 2^j$.

Local Ridgelet Transforms

The ridgelet transform is optimal for finding global lines of the size of the image. To detect line segments, a partitioning must be introduced [9]. The image can be decomposed into overlapping blocks of side-length b pixels in such a way that the overlap between two vertically adjacent blocks is a rectangular array of size b by $b/2$; we use overlap to avoid blocking artifacts. For an n by n image, we count $2n/b$ such blocks in each direction, and thus the redundancy factor grows by a factor of 4.

The partitioning introduces redundancy, as a pixel belongs to 4 neighboring blocks. We present two competing strategies to perform the analysis and synthesis:

1. The block values are weighted by a spatial window w (analysis) in such a way that the co-addition of all blocks reproduce exactly the original pixel value (synthesis).
2. The block values are those of the image pixel values (analysis) but are weighted when the image is reconstructed (synthesis).

Experiments have shown that the second approach leads to better results especially for restoration problems, see [68] for details. We calculate a pixel value, $f[i_1, i_2]$ from its four corresponding block values of half-size $m = b/2$, namely, $B_1[k_1, l_1]$, $B_2[k_2, l_1]$, $B_3[k_1, l_2]$ and $B_4[k_2, l_2]$ with $k_1, l_1 > b/2$ and $k_2 = k_1 - m, l_2 = l_1 - m$, in the following way:

$$\begin{aligned} f_1 &= w(k_2/m)B_1[k_1, l_1] + w(1 - k_2/m)B_2[k_2, l_1] \\ f_2 &= w(k_2/m]B_3[k_1, l_2] + w(1 - k_2/m)B_4[k_2, l_2] \\ f[i_1, i_2] &= w(l_2/m)f_1 + w(1 - l_2/m)f_2 . \end{aligned} \tag{5}$$

where $w(x) = \cos^2(\pi x/2)$ is the window. Of course, one might select any other smooth, non-increasing function satisfying, $w(0) = 1, w(1) = 0, w'(0) = 0$ and obeying the symmetry property $w(x) + w(1 - x) = 1$.

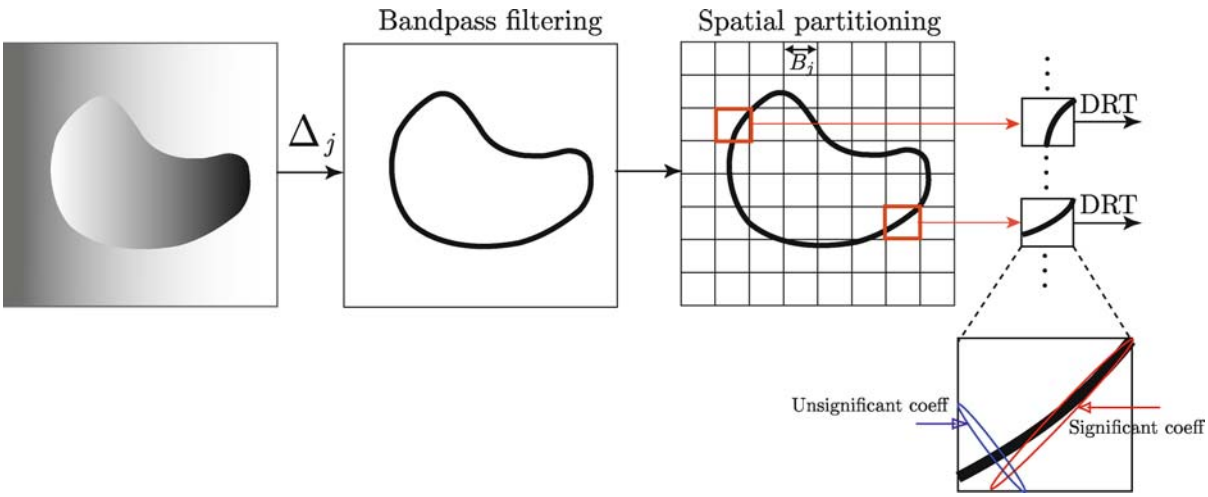
Sparse Representation by Ridgelets

The continuous ridgelet transform provides sparse representation of both smooth functions (in the Sobolev space W_2^2) and of perfectly straight lines [11,31]. We have just seen that there are also various DRTs, i. e. expansions with countable discrete collection of generating elements, which correspond either to frames or orthobases. It has been shown for these schemes that the DRT achieves near-optimal M-term approximation – that is the non-linear approximation of f using the M highest ridgelet coefficients in magnitude - to smooth images with discontinuities along straight lines [16,31]. In summary, ridgelets provide sparse presentation for piecewise smooth images away from global straight edges.

Curvelets

The First Generation Curvelet Transform

In image processing, edges are curved rather than straight lines and ridgelets are not able to efficiently represent such



Curvelets and Ridgelets, Figure 9
Local ridgelet transform on bandpass filtered image. At fine scales, curved edges are almost straight lines

images. However, one can still deploy the ridgelet machinery in a localized way, at fine scales, where curved edges are almost straight lines (see Fig. 9). This is the idea underlying the first generation curvelets (termed here CurveletG1) [18].

First Generation Curvelets Construction The CurveletG1 transform [18,32,68] opens the possibility to analyze an image with different block sizes, but with a single transform. The idea is to first decompose the image into a set of wavelet bands, and to analyze each band by a local ridgelet transform as illustrated on Fig. 9. The block size can be changed at each scale level. Roughly speaking, different levels of the multiscale ridgelet pyramid are used to represent different sub-bands of a filter bank output. At the same time, this sub-band decomposition imposes a relationship between the width and length of the *important* frame elements so that they are anisotropic and obey approximately the parabolic scaling law width $\approx$ length^2.

The First Generation Discrete Curvelet Transform (DCTG1) of a continuum function $f(\mathbf{x})$ makes use of a dyadic sequence of scales, and a bank of filters with the property that the bandpass filter Δ_j is concentrated near the frequencies $[2^{2j}, 2^{2j+2}]$, i. e.

$$\Delta_j(f) = \Psi_{2j} * f, \quad \widehat{\Psi}_{2j}(\boldsymbol{\nu}) = \widehat{\Psi}(2^{-2j}\boldsymbol{\nu}) .$$

In wavelet theory, one uses a decomposition into dyadic sub-bands $[2^j, 2^{j+1}]$. In contrast, the sub-bands used in the discrete curvelet transform of continuum functions has the nonstandard form $[2^{2j}, 2^{2j+2}]$. This is nonstandard feature of the DCTG1 well worth remembering (this is where the approximate parabolic scaling law comes into play).

The DCTG1 decomposition is the sequence of the following steps:

- *Sub-band Decomposition.* The object f is decomposed into sub-bands.
- *Smooth Partitioning.* Each sub-band is smoothly windowed into “squares” of an appropriate scale (of side-length $\sim 2^{-j}$).
- *Ridgelet Analysis.* Each square is analyzed via the DRT.

In this definition, the two dyadic sub-bands $[2^{2j}, 2^{2j+1}]$ and $[2^{2j+1}, 2^{2j+2}]$ are merged before applying the ridgelet transform.

Digital Implementation It seems that the isotropic “à trous” wavelet transform (▶ Numerical Issues When Using Wavelets), [72] is especially well-adapted to the needs of the digital curvelet transform. The algorithm decomposes an n by n image $f[i_1, i_2]$ as a superposition of the form

$$f[i_1, i_2] = c_J[i_1, i_2] + \sum_{j=1}^{J} w_j[i_1, i_2],$$

where c_J is a coarse or smooth version of the original image f and w_j represents ‘the details of f’ at scale 2^{-j}. Thus, the algorithm outputs $J + 1$ sub-band arrays of size $n \times n$.
A sketch of the DCTG1 implementation is given in Algorithm 1.
The side-length of the localizing windows is doubled *at every other* dyadic sub-band, hence maintaining the fundamental property of the curvelet transform which says

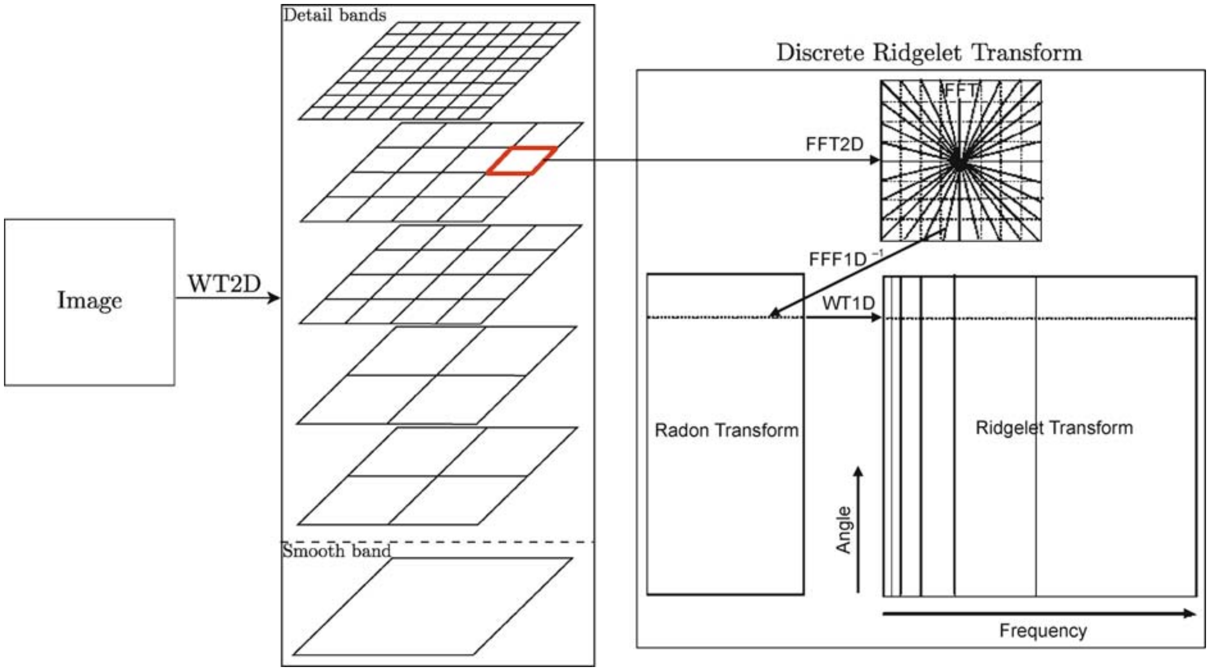
```
Require: Input $n \times n$ image $f[i_1, i_2]$, type of DRT (see above).
Apply the à trous isotropic WT2D with $J$ scales,
Set $B_1 = B_{\min}$,
for $j = 1, \ldots, J$ do
    Partition the sub-band $w_j$ with a block size $B_j$ and apply the DRT to each block,
    if $j$ modulo $2 = 1$ then
        $B_{j+1} = 2B_j$,
    else
        $B_{j+1} = B_j$.
    end if
end for
```

Curvelets and Ridgelets, Algorithm 1
DCTG1

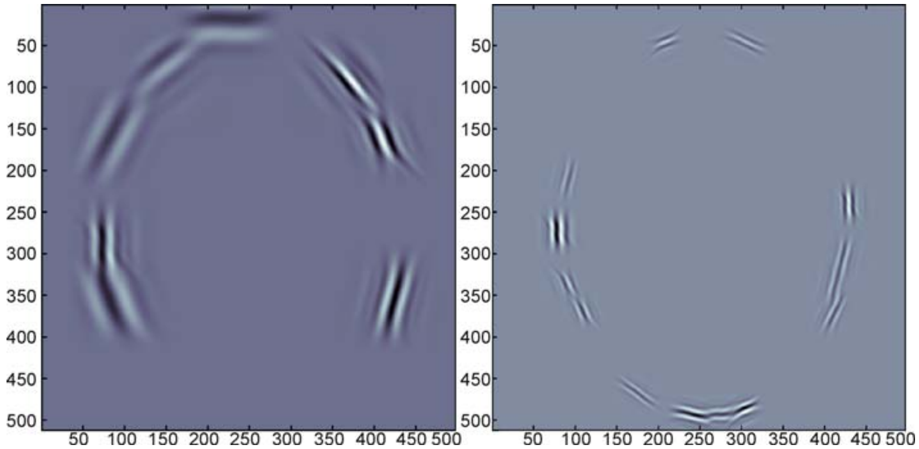
that elements of length about $2^{-j/2}$ serve for the analysis and synthesis of the jth sub-band $[2^j, 2^{j+1}]$. Note also that the coarse description of the image c_J is left intact. In the results shown in this paper, we used the default value $B_{\min} = 16$ pixels in our implementation. Figure 10 gives an overview of the organization of the DCTG1 algorithm.

This implementation of the DCTG1 is also redundant. The redundancy factor is equal to $16J + 1$ whenever J scales are employed. The DCTG1 algorithm enjoys exact reconstruction and stability, as each step of the analysis (decomposition) algorithm is itself invertible. One can show that the computational complexity of the DCTG1 algorithm we described here based on the DRT of Fig. 3 is $O(n^2(\log n)^2)$ for an $n \times n$ image.
Figure 11 shows a few curvelets at different scales, orientations and locations.

Sparse Representation by First Generation Curvelets The CurveletG1 elements can form either a frame or a tight frame for $L_2(\mathbb{R}^2)$ [17], depending on the WT2D used and the DRT implementation (rectopolar or FSS Radon transform). The frame elements are anisotropic by construction and become successively more anisotropic at progressively higher scales. These curvelets also exhibit directional sensitivity and display oscillatory components across the 'ridge'. A central motivation leading to the curvelet construction was the problem of non-adaptively representing piecewise smooth (e. g. C^2) images f which have discontinuity along a C^2 curve. Such a model is the so-called cartoon model of (non-textured) images. With the CurveletG1 tight frame construction, it was shown in [17] that for such f, the M-term non-linear approximations f_M of f obey,



Curvelets and Ridgelets, Figure 10
First Generation Discrete Curvelet Transform (DCTG1) flowchart. The figure illustrates the decomposition of the original image into sub-bands followed by the spatial partitioning of each sub-band. The ridgelet transform is then applied to each block



Curvelets and Ridgelets, Figure 11
A few first generation curvelets

for each $\kappa > 0$,

$$\|f - f_M\|_2 \leq C_\kappa M^{-2+\kappa} , \quad M \to +\infty .$$

The M-term approximations in the CurveletG1 are almost rate optimal, much better than M-term Fourier or wavelet approximations for such images, see [53].

The Second Generation Curvelet Transform

Despite these interesting properties, the CurveletG1 construction presents some drawbacks. First, the construction involves a complicated seven-index structure among which we have parameters for scale, location and orientation. In addition, the parabolic scaling ratio width $\approx$ length^2 is not completely true (see Subsect. "First Generation Curvelets Construction"). In fact, CurveletG1 assumes a wide range of aspect ratios. These facts make mathematical and quantitative analysis especially delicate. Second, the spatial partitioning of the CurveletG1 transform uses overlapping windows to avoid blocking effects. This leads to an increase of the redundancy of the DCTG1. The computational cost of the DCTG1 algorithm may also be a limitation for large-scale data, especially if the FSS-based DRT implementation is used.

In contrast, the second generation curvelets (CurveletG2) [15,20] exhibit a much simpler and natural indexing structure with three parameters: scale, orientation (angle) and location, hence simplifying mathematical analysis. The CurveletG2 transform also implements a tight frame expansion [20] and has a much lower redundancy. Unlike the DCTG1, the discrete CurveletG2 implementation will not use ridgelets yielding a faster algorithm [15,20].

Second Generation Curvelets Construction

Continuous Coronization The second generation curvelets are defined at scale 2^{-j}, orientation θ_l and position $\mathbf{x}_\mathbf{k}^{j,l} = R_{\theta_l}^{-1}(2^{-j}k_1, 2^{-j/2}k_2)$ by translation and rotation of a mother curvelet φ_j as

$$\varphi_{j,l,\mathbf{k}}(\mathbf{x}) = \varphi_j(R_{\theta_l}(\mathbf{x} - \mathbf{x}_\mathbf{k}^{j,l})) , \tag{6}$$

where R_{θ_l} is the rotation by θ_l radians. θ_l is the equi-spaced sequence of rotation angles $\theta_l = 2\pi 2^{-\lfloor j/2 \rfloor} l$, with integer l such that $0 \leq \theta_l \leq 2\pi$ (note that the number of orientations varies as $1/\sqrt{\text{scale}}$). $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$ is the sequence of translation parameters. The waveform φ_j is defined by means of its Fourier transform $\hat{\varphi}_j(\boldsymbol{\nu})$, written in polar coordinates in the Fourier domain

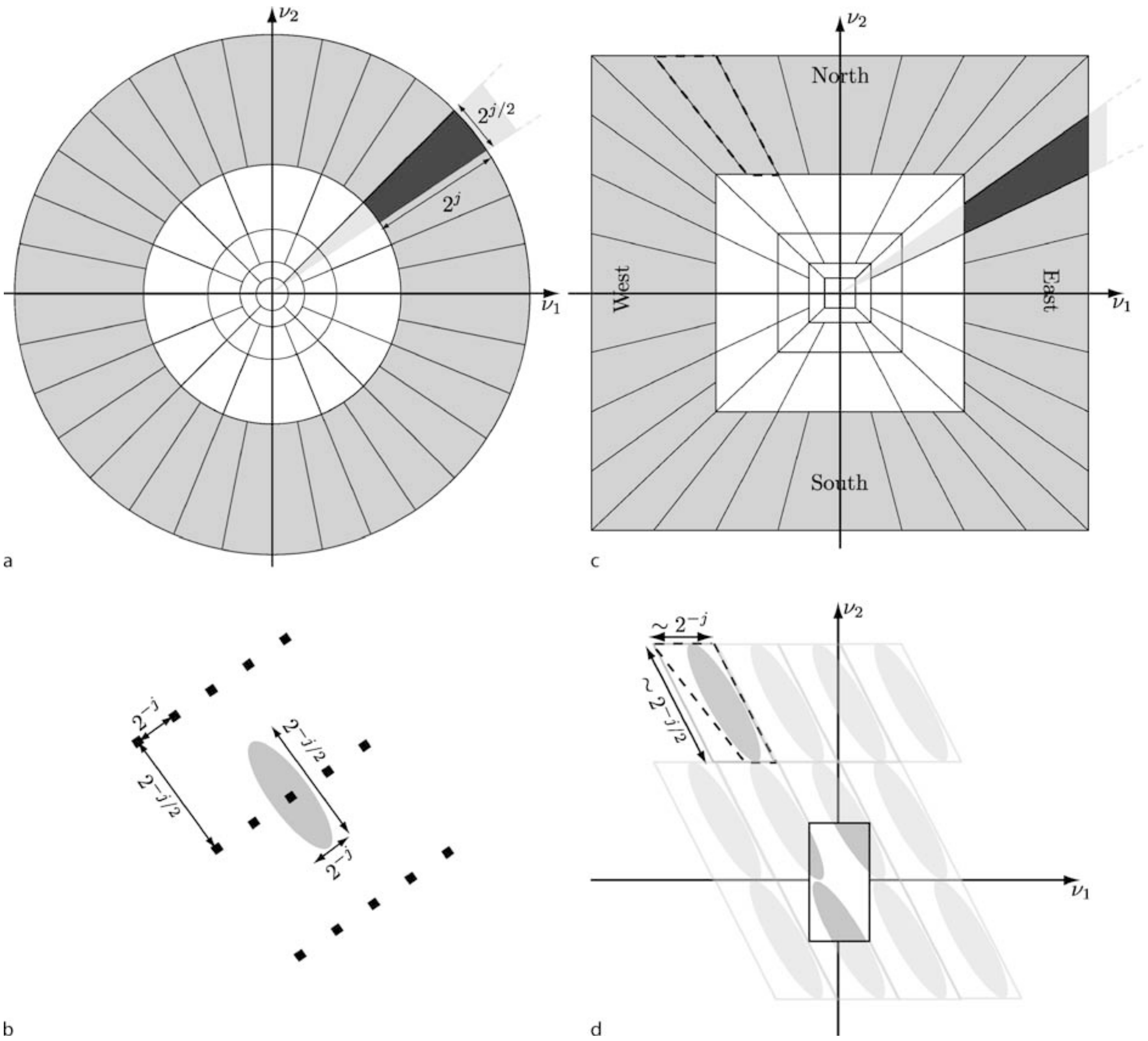
$$\hat{\varphi}_j(r, \theta) = 2^{-3j/4} \hat{w}(2^{-j} r) \hat{v} \left(\frac{2^{\lfloor j/2 \rfloor} \theta}{2\pi} \right) . \tag{7}$$

The support of $\hat{\varphi}_j$ is a polar *parabolic wedge* defined by the support of $\hat{w}$ and $\hat{v}$, the radial and angular windows (both smooth, nonnegative and real-valued), applied with scale-dependent window widths in each direction. $\hat{w}$ and $\hat{v}$ must also satisfy the partition of unity property [15]. See the frequency tiling in Fig. 12a.

In continuous frequency $\boldsymbol{\nu}$, the CurveletG2 coefficients of data $f(\mathbf{x})$ are defined as the inner product

$$c_{j,l,\mathbf{k}} := \langle f \varphi_{j,l,\mathbf{k}} \rangle = \int_{\mathbb{R}^2} \hat{f}(\boldsymbol{\nu}) \hat{\varphi}_j(R_{\theta_l} \boldsymbol{\nu}) e^{i \mathbf{x}_\mathbf{k}^{j,l} \cdot \boldsymbol{\nu}} \mathrm{d}\boldsymbol{\nu} . \tag{8}$$

This construction implies a few properties: (i) the CurveletG2 defines a tight frame of $L_2(\mathbb{R}^2)$, (ii) the effective length and width of these curvelets obey the parabolic



Curvelets and Ridgelets, Figure 12

a Continuous curvelet frequency tiling. The *dark gray* area represents a wedge obtained as the product of the radial window (annulus shown in *gray*) and the angular window (*light gray*). b The Cartesian grid in space associated to the construction in a whose spacing also obeys the parabolic scaling by duality. c Discrete curvelet frequency tiling. The window $\hat{u}_{j,l}$ isolates the frequency near trapezoidal wedge such as the one shown in dark gray. d The wrapping transformation. The *dashed line* shows the same trapezoidal wedge as in c. The parallelogram contains this wedge and hence the support of the curvelet. After periodization, the wrapped Fourier samples can be collected in the rectangle centered at the origin

scaling relation $(2^{-j} = \text{width}) = (\text{length} = 2^{-j/2})^2$, (iii) the curvelets exhibit an oscillating behavior in the direction perpendicular to their orientation. Curvelets as just constructed are complex-valued. It is easy to obtain real-valued curvelets by working on the symmetrized version $\hat{\varphi}_j(r,\theta) + \hat{\varphi}_j(r,\theta+\pi)$.

Discrete Coronization The discrete transform takes as input data defined on a Cartesian grid and outputs a collection of coefficients. The continuous-space definition of the CurveletG2 uses coronae and rotations that are not especially adapted to Cartesian arrays. It is then convenient to replace these concepts by their Cartesian counterparts. That is concentric squares (instead of concentric circles) and shears (instead of rotations), see Fig. 12c.

The Cartesian equivalent to the radial window $\hat{w}_j(\boldsymbol{\nu}) = \hat{w}(2^{-j}\boldsymbol{\nu})$ would be a bandpass frequency-localized window which can be derived from the difference of separable low-pass windows $H_j(\boldsymbol{\nu}) = \hat{h}(2^{-j}\nu_1)\hat{h}(2^{-j}\nu_2)$ (h is a 1D low-pass filter):

$$\hat{w}_j(\boldsymbol{\nu}) = \sqrt{H_{j+1}^2(\boldsymbol{\nu}) - H_j^2(\boldsymbol{\nu})},\ \forall j \geq 0\,, \quad \text{and} \quad \hat{w}_0(\boldsymbol{\nu}) = \hat{h}(\nu_1)\hat{h}(\nu_2)\,.$$

Another possible choice is to select these windows inspired by the construction of Meyer wavelets [20,57]. See [15] for more details about the construction of the Cartesian $\hat{w}_j$'s.

Let's now examine the angular localization. Each Cartesian coronae has four quadrants: East, North, West and South. Each quadrant is separated into $2^{\lfloor j/2 \rfloor}$ orientations (wedges) with the same areas. Take for example the East quadrant ($-\pi/4 \leq \theta_l < \pi/4$). For the West quadrant, we would proceed by symmetry around the origin, and for the North and South quadrant by exchanging the roles of ν_1 and ν_2. Define the angular window for the lth direction as

$$\hat{v}_{j,l}(\boldsymbol{\nu}) = \hat{v}\left(2^{\lfloor j/2 \rfloor}\frac{\nu_2 - \nu_1 \tan\theta_l}{\nu_1}\right), \tag{9}$$

with the sequence of equi-spaced slopes (and not angles) $\tan\theta_l = 2^{-\lfloor j/2 \rfloor} l$, for $l = -2^{\lfloor j/2 \rfloor}, \ldots, 2^{\lfloor j/2 \rfloor} - 1$. We can now define the window which is the Cartesian analog of $\hat{\varphi}_j$ above,

$$\hat{u}_{j,l}(\boldsymbol{\nu}) = \hat{w}_j(\boldsymbol{\nu})\hat{v}_{j,l}(\boldsymbol{\nu}) = \hat{w}_j(\boldsymbol{\nu})\hat{v}_{j,0}(S_{\theta_l}\boldsymbol{\nu}), \tag{10}$$

where S_{θ_l} is the shear matrix. From this definition, it can be seen that $\hat{u}_{j,l}$ is supported near the trapezoidal wedge $\{\boldsymbol{\nu} = (\nu_1, \nu_2) | 2^j \leq \nu_1 \leq 2^{j+1}, -2^{-j/2} \leq \nu_2/\nu_1 - \tan\theta_l \leq 2^{-j/2}\}$. The collection of $\hat{u}_{j,l}(\boldsymbol{\nu})$ gives rise to the frequency tiling shown in Fig. 12c. From $\hat{u}_{j,l}(\boldsymbol{\nu})$, the digital CurveletG2 construction suggests Cartesian curvelets that are translated and sheared versions of a mother Cartesian curvelet $\hat{\varphi}_j^D(\boldsymbol{\nu}) = \hat{u}_{j,0}(\boldsymbol{\nu}), \varphi_{j,l,\mathbf{k}}^D(\mathbf{x}) = 2^{3j/4}\varphi_j^D\left(S_{\theta_l}^T\mathbf{x} - \mathbf{m}\right)$ where $\mathbf{m} = (k_1 2^{-j}, k_2 2^{-j/2})$.

Digital Implementation The goal here is to find a digital implementation of the Second Generation Discrete Curvelet Transform (DCTG2), whose coefficients are now given by

$$c_{j,l,\mathbf{k}} := \left\langle f\varphi_{j,l,\mathbf{k}}^D \right\rangle = \int_{\mathbb{R}^2} \hat{f}(\boldsymbol{\nu})\hat{\varphi}_j^D(S_{\theta_l}^{-1}\boldsymbol{\nu})e^{iS_{\theta_l}^{-T}\mathbf{m}\cdot\boldsymbol{\nu}}\mathrm{d}\boldsymbol{\nu}\,. \tag{11}$$

To evaluate this formula with discrete data, one may think of (i) using the 2D FFT to get $\hat{f}$, (ii) form the windowed frequency data $\hat{f}\hat{u}_{j,l}$ and (iii) apply the the inverse Fourier transform. But this necessitates to evaluate the FFT at the sheared grid $S_{\theta_l}^{-T}\mathbf{m}$, for which the classical FFT algorithm is not valid. Two implementations were then proposed [15], essentially differing in their way of handling the grid:

- A tilted grid mostly aligned with the axes of $\hat{u}_{j,l}(\boldsymbol{\nu})$ which leads to the Unequi-Spaced FFT (USFFT)-based DCTG2. This implementation uses a nonstandard interpolation. Furthermore, the inverse transform uses conjugate gradient iteration to invert the interpolation step. This will have the drawback of a higher computational burden compared to the wrapping-based implementation that we will discuss hereafter. We will not elaborate more about the USFFT implementation as we never use it in practice. The interested reader may refer to [15] for further details and analysis.
- A grid aligned with the input Cartesian grid which leads to the wrapping-based DCTG2.

The wrapping-based DCTG2 makes a simpler choice of the spatial grid to translate the curvelets. The curvelet coefficients are essentially the same as in (11), except that $S_{\theta_l}^{-T}\mathbf{m}$ is replaced by $\mathbf{m}$ with values on a rectangular grid. But again, a difficulty rises because the window $\hat{u}_{j,l}$ does not fit in a rectangle of size $2^j \times 2^{j/2}$ to which an inverse FFT could be applied. The *wrapping* trick consists in periodizing the windowed frequency data $\hat{f}\hat{u}_{j,l}$, and reindexing the samples array by wrapping around a $\sim 2^j \times 2^{j/2}$ rectangle centered at the origin, see Fig. 12d to get a gist of the wrapping idea.

The wrapping-based DCTG2 algorithm can be summarized as in Algorithm 2.

The DCTG2 implementation can assign either wavelets or curvelets at the finest scale. In the CurveLab toolbox [75], the default choice is set to wavelets at the finest scale, but this can be easily modified directly in the code.

```
Require: Input $n \times n$ image $f[i_1, i_2]$, coarsest decomposition scale, curvelets or wavelets at the finest scale.
Apply the 2D FFT and obtain Fourier samples $\hat{f}[i_1, i_2]$.
for each scale $j$ and angle $l$ do
    Form the product $\hat{f}[i_1, i_2]\hat{u}_{j,l}[i_1, i_2]$.
    Wrap this product around the origin.
    Apply the inverse 2D FFT to the wrapped data to get discrete DCTG2 coefficients.
end for
```

Curvelets and Ridgelets, Algorithm 2
DCTG2 via wrapping

We would like to apologize to the expert reader as many technical details are (deliberately) missing here on the CurveletG2 construction. For instance, low-pass coarse component, window overlapping, windows over junctions between quadrants. This paper is intended to give an overview of these recent multi-scale transforms, and the genuinely interested reader may refer to the original papers of Candès, Donoho, Starck and co-workers for further details (see bibliography).

The computational complexity of the wrapping-based DCTG2 analysis and reconstruction algorithms is that of the FFT $O(n^2 \log n)$, and in practice, the computation time is that of 6 to 10 2D FFTs [15]. This is a faster algorithm compared to the DCTG1. The DCTG2 fast algorithm has participated to make the use of the curvelet transform more attractive in many applicative fields (see Sect. "Stylized Applications" for some of them). The DCTG2, as it is implemented in the CurveLab toolbox [75], has reasonable redundancy, at most ~ 7.8 (much higher in 3D) if curvelets are used at the finest scale. This redundancy can even be reduced down to 4 (and 8 in 3D) if we replace in this implementation the Meyer wavelet construction, which introduces a redundancy factor of 4, by another wavelet pyramidal construction, similar to the one presented in Sect. "One–dimensional Wavelet Transform" which has a redundancy less than 2 in any dimension. Our experiments have shown that this modification does not modify the results in denoising experiments. DCTG2 redundancy is anyway much smaller than the DCTG1 one which is $16J + 1$. As stated earlier, the DCTG2 coefficients are complex-valued, but a real-valued DCTG2 with the same redundancy factor can be easily obtained by properly combining coefficients at orientations θ_l and $\theta_l + \pi$.

The DCTG2 can be extended to higher dimensions [21]. In the same vein as wavelets on the interval [53], the DCGT2 has been recently adapted to handle image boundaries by mirror extension instead of periodization [25]. The latter modification can have immediate implications in image processing applications where the contrast difference at opposite image boundaries may be an issue (see e. g. the denoising experiment discussion reported in Sect. "Stylized Applications").

We would like to make a connection with other multiscale directional transforms directly linked to curvelets. The contourlets tight frame of Do and Vetterli [27] implements the CurveletG2 idea directly on a discrete grid using a perfect reconstruction filter bank procedure. In [51], the authors proposed a modification of the contourlets with a directional filter bank that provides a frequency partitioning which is close to the curvelets but with no redundancy. Durand in [35] recently introduced families of non-adaptive directional wavelets with various frequency tilings, including that of curvelets. Such families are non-redundant and form orthonormal bases for $L_2(\mathbb{R}^2)$, and have an implementation derived from a single nonseparable filter bank structure with nonuniform sampling.

Sparse Representation by Second Generation Curvelets It has been shown by Candès and Donoho [20] that with the CurveletG2 tight frame construction, the M-term nonlinear approximation error of C^2 images except at discontinuities along C^2 curves obey

$$\|f - f_M\|_2 \leq CM^{-2}(\log M)^3 \ .$$

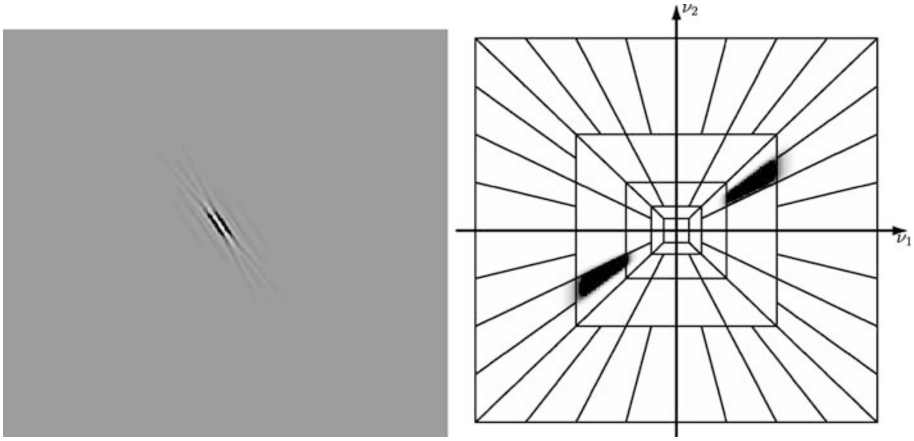
This is an asymptotically optimal convergence rate (up to the $(\log M)^3$ factor), and holds uniformly over the $C^2 - C^2$ class of functions. This is a remarkable result since the CurveletG2 representation is non-adaptative. However, the simplicity due to the non-adaptivity of curvelets has a cost: curvelet approximations loose their near optimal properties when the image is composed of edges which are not exactly C^2. Additionally, if the edges are C^α-regular with $\alpha > 2$, then the curvelets convergence rate exponent remain 2. Other adaptive geometric representations such as bandlets are specifically designed to reach the optimal decay rate $O(M^{-\alpha})$ [49,62].

Stylized Applications

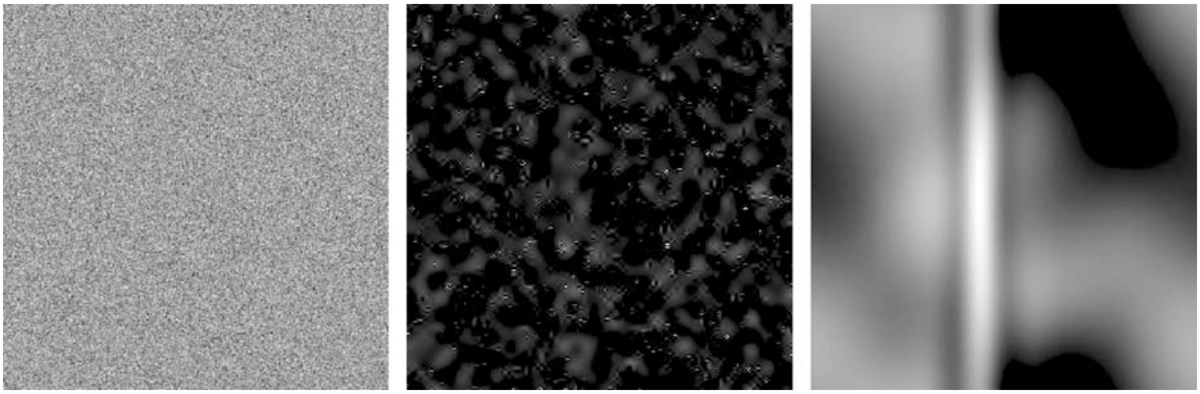
Denoising

Elongated Feature Recovery The ability of ridgelets to sparsely represent piecewise smooth images away from discontinuities along lines has an immediate implication on statistical estimation. Consider a piecewise smooth image f away from line singularities embedded in an additive white noise of standard deviation σ. The ridgelet-based thresholding estimator is nearly optimal for recovering such functions, with a mean-square error (MSE) decay rate almost as good as the minimax rate [12].

To illustrate these theoretical facts, we simulate a vertical band embedded in white Gaussian noise with large σ. Figure 14 (top left) represents such a noisy image. The parameters are as follows: the pixel width of the band is 20 and the signal-to-noise ratio (SNR) is set to 0.1. Note that it is not possible to distinguish the band by eye. The wavelet transform (undecimated wavelet transform) is also incapable of detecting the presence of this object; roughly speaking, wavelet coefficients correspond to weighted averages over approximately isotropic neighborhoods (at different scales) and those wavelets clearly do not correlate very well with the very elongated structure (pattern) of the object to be detected.



Curvelets and Ridgelets, Figure 13
An example of second generation real curvelet. *Left*: curvelet in spatial domain. *Right*: its Fourier transform



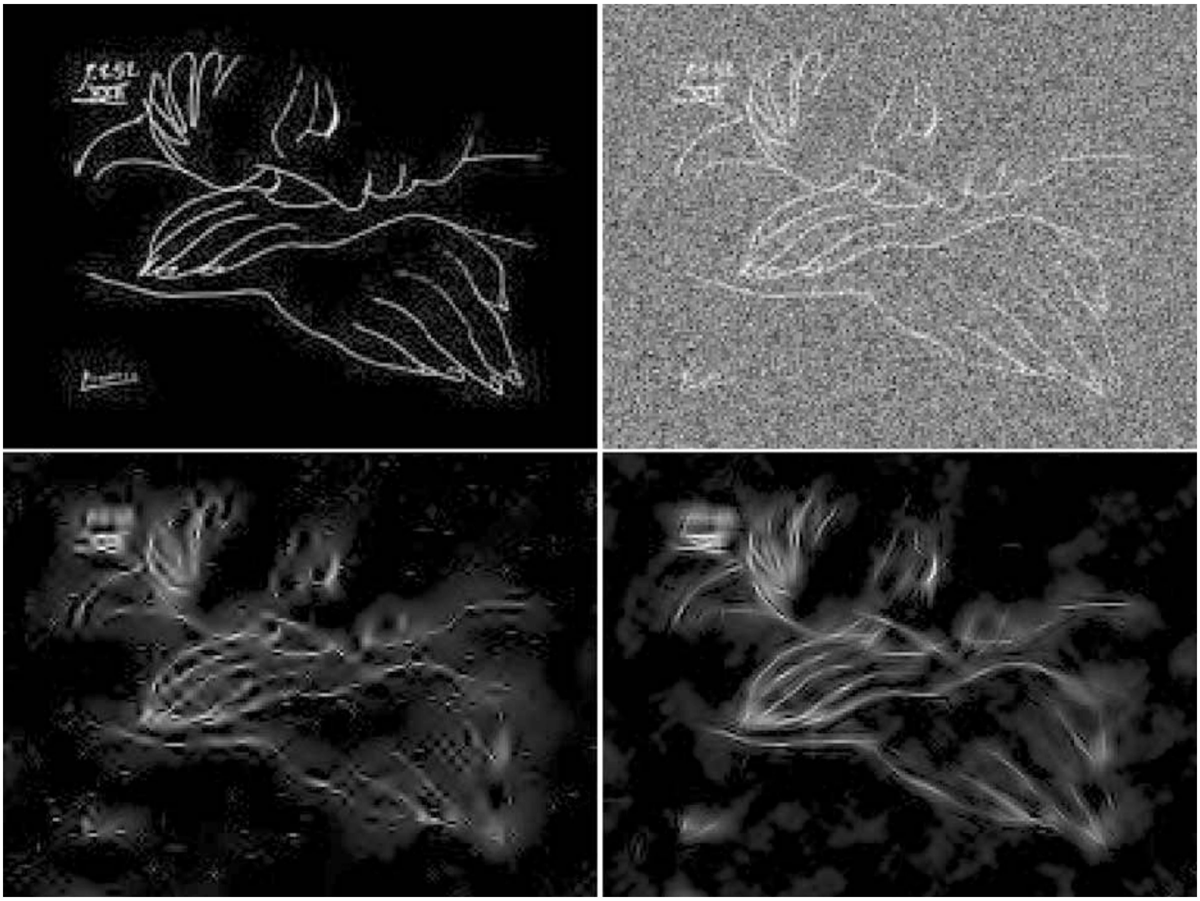
Curvelets and Ridgelets, Figure 14
Original image containing a vertical band embedded in white noise with relatively large amplitude (*left*). Denoised image using the undecimated wavelet transform (*middle*). Denoised image using the DRT based on the rectopolar Radon transform (*right*)

Curve Recovery Consider now the problem of recovering a piecewise C^2 function f apart from a discontinuity along a C^2 edge. Again, a simple strategy based on thresholding curvelet tight frame coefficients yields an estimator that achieves a MSE almost of the order $O(\sigma^{4/3})$ uniformly over the $C^2 - C^2$ class of functions [19]. This is the optimal rate of convergence as the minimax rate for that class scales as $\sigma^{4/3}$ [19]. Comparatively, wavelet thresholding methods only achieves a MSE of order $O(\sigma)$ and no better. We also note that the statistical optimality of the curvelet thresholding extends to a large class of ill-posed linear inverse problems [19].

In the experiment of Fig. 15, we have added a white Gaussian noise to "War and Peace", a drawing from Picasso which contains many curved features. Figure 15 bottom left and right shows respectively the restored images by the undecimated wavelet transform and the DCTG1. Curves are more sharply recovered with the DCTG1.

In a second experiment, we compared the denoising performance of several digital implementations of the curvelet transform; namely the DCTG1 with the rectopolar DRT, the DCTG1 with the FSS-based DRT and the wrapping-based DCTG2. The results are shown in Fig. 16, where the original 512×512 `Peppers` image was corrupted by a Gaussian white noise $\sigma = 20$ (PSNR = 22dB).

Although the FSS-based DRT is more accurate than the rectopolar DRT, the denoising improvement of the former (PSNR=31.31dB) is only 0.18 dB better than the latter (PSNR=31.13dB) on `Peppers`. The difference is almost undistinguishable by eye, but the computation time is 20 higher for the DCTG1 with the FSS DRT. Consequently, it appears that there is a little benefit of using the



Curvelets and Ridgelets, Figure 15

The Picasso picture `War and Peace` (*top left*), the same image contaminated with a Gaussian white noise (*top right*). The restored images using the undecimated wavelet transform (*bottom left*) and the DCTG1 (*bottom right*)

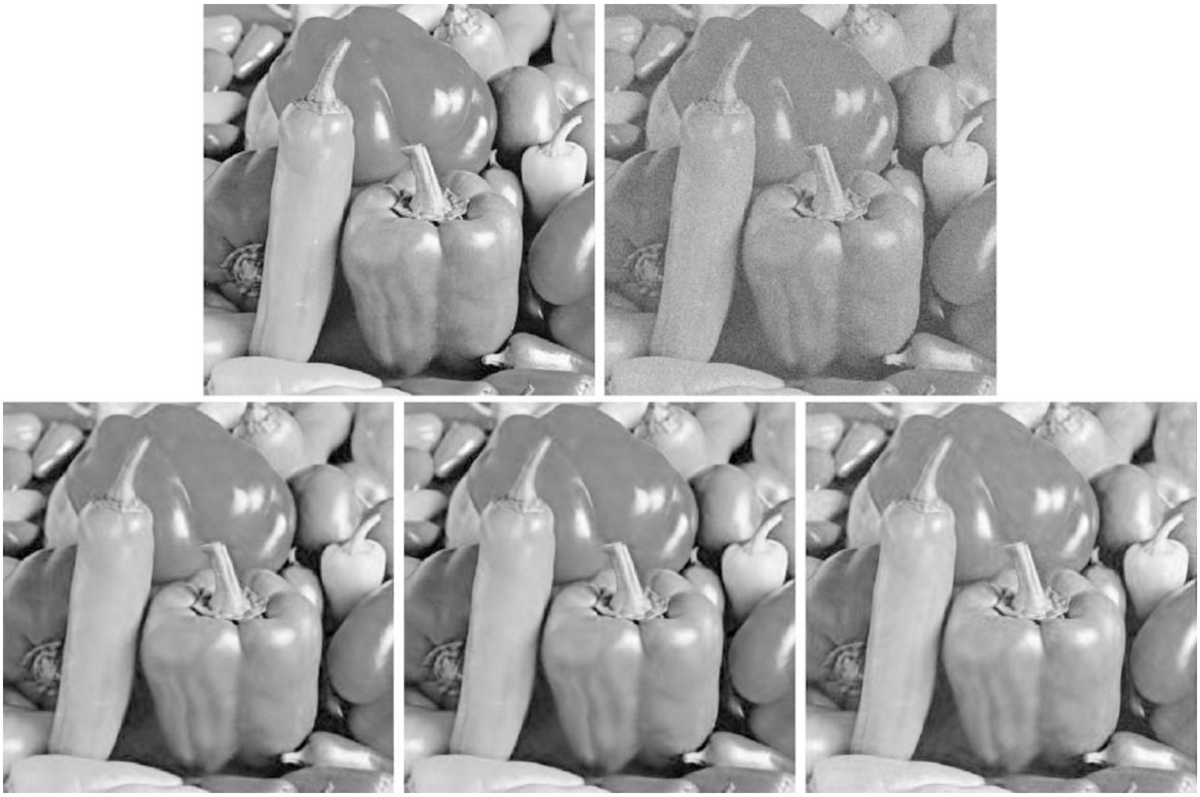
FSS DRT in the DCTG1 for restoration applications. Denoising using the DCTG2 with the wrapping implementation gives a PSNR=30.62 dB which is ~ 0.7 dB less than the DCTG1. But this is the price to pay for a lower redundancy and a much faster transform algorithm. Moreover, the DCTG2 exhibits some artifacts which look like 'curvelet ghosts'. This is a consequence of the fact that the DCTG2 makes a central use of the FFT which has the side effect of treating the image boundaries by periodization.

Linear Inverse Problems

Many problems in image processing can be cast as inverting the linear degradation equation $y = Hf + \varepsilon$, where f is the image to recover, y the observed image and ε is a white noise of variance $\sigma^2 < +\infty$. The linear mapping H is generally ill-behaved which entails ill-posedness of the inverse problem. Typical examples of linear inverse problems include image deconvolution where H is the convolution operator, or image inpainting (recovery of missing data) where H is a binary mask.

In the last few years, some authors have attacked the problem of solving linear inverse problems under the umbrella of sparse representations and variational formulations, e.g. for deconvolution [19,24,38,41,74] and inpainting [37,39]. Typically, in this setting, the recovery of f is stated as an optimization problem with a sparsity-promoting regularization on the representation coefficients of f, e.g. its wavelet or curvelet coefficients. See [37,38,39,74] for more details.

In Fig. 17 first row, we depict an example of deconvolution on `Barbara` using the algorithm described in [38] with the DCTG2 curvelet transform. The original, degraded (blurred with an exponential kernel and noisy) and restored images are respectively shown on the left, middle



Curvelets and Ridgelets, Figure 16

Comparison of denoising performance of several digital implementations of the curvelet transform. *Top left*: original image. *Top right*: noisy image $\sigma = 20$. *Bottom left*: denoised with DCTG1 using the rectopolar DRT. *Bottom middle*: denoised with DCTG1 using the FSS DRT. Bottom right: denoised with DCTG2 using the wrapping implementation

and right. The second row gives an example of inpainting on `Claudia` image using the DCTG2 with 50% missing pixels.

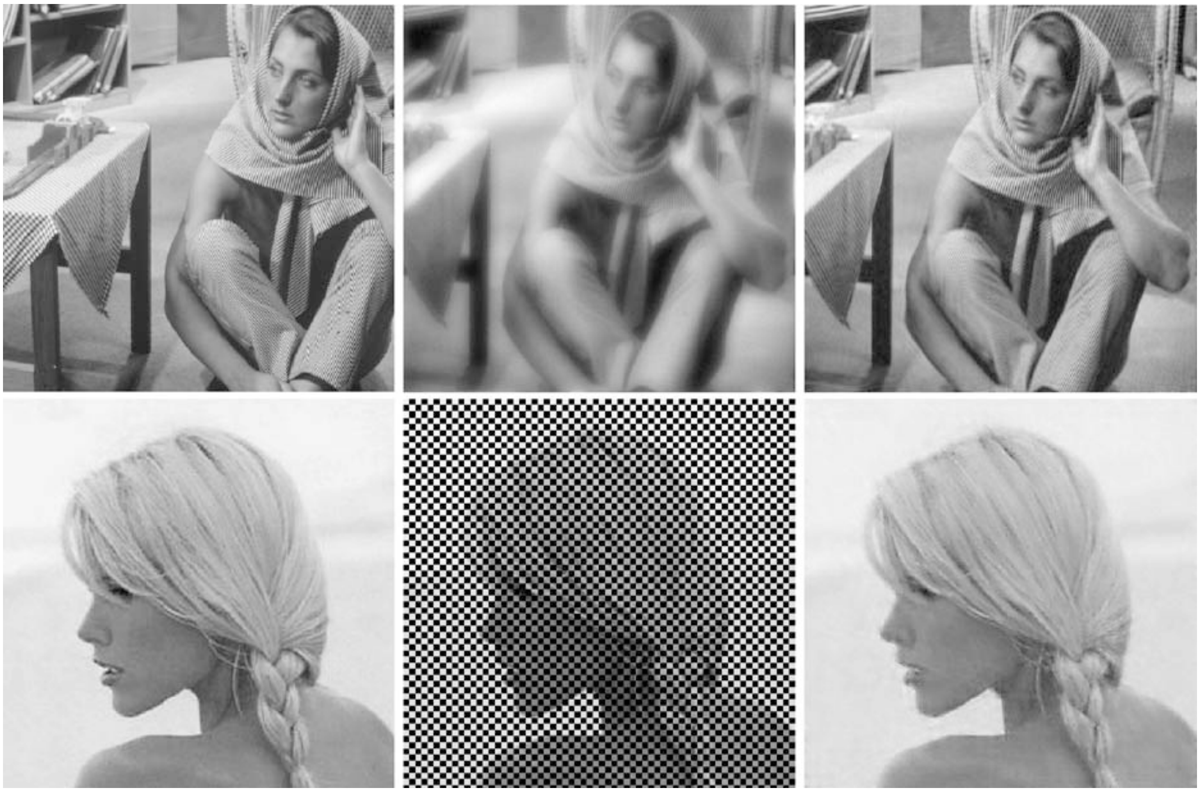
Contrast Enhancement

The curvelet transform has been successfully applied to image contrast enhancement by Starck et al. [73]. As the curvelet transform capture efficiently edges in an image, it is a good candidate for multiscale edge enhancement. The idea is to modify the curvelet coefficients of the input image in order to enhance its edges. The curvelet coefficients are typically modified according to the function displayed in the left plot of Fig. 18. Basically, this plot says that the input coefficients are kept intact (or even shrunk) if they have either low (e. g. below the noise level) or high (strongest edges) values. Intermediate curvelet coefficient values which correspond to the faintest edges are amplified. An example of curvelet-based image enhancement on `Saturn` image is given in Fig. 18.

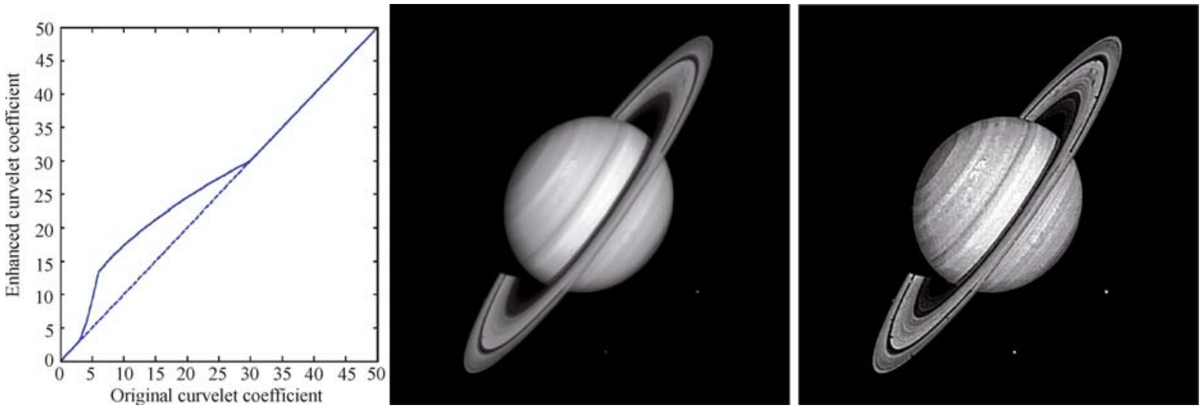
Morphological Component Separation

The idea to morphologically decompose a signal/image into its building blocks is an important problem in signal and image processing. Successful separation of a signal content has a key role in the ability to effectively analyze it, enhance it, compress it, synthesize it, and more. Various approaches have been proposed to tackle this problem.

The Morphological Component Analysis method (MCA) [70,71] is a method which allows us to decompose a single signal into two or more layers, each layer containing only one kind of feature in the input signal or image. The separation can be achieved when each kind of feature is sparsely represented by a given transformation in the dictionary of transforms. Furthermore, when a transform sparsely represents a part in the signal/image, it yields non-sparse representation on the other content type. For instance, lines and Gaussians in a image can be separated using the ridgelet transform and the wavelet transform. Locally oscillating textures can be separated from the piecewise smooth content using the local discrete cosine trans-



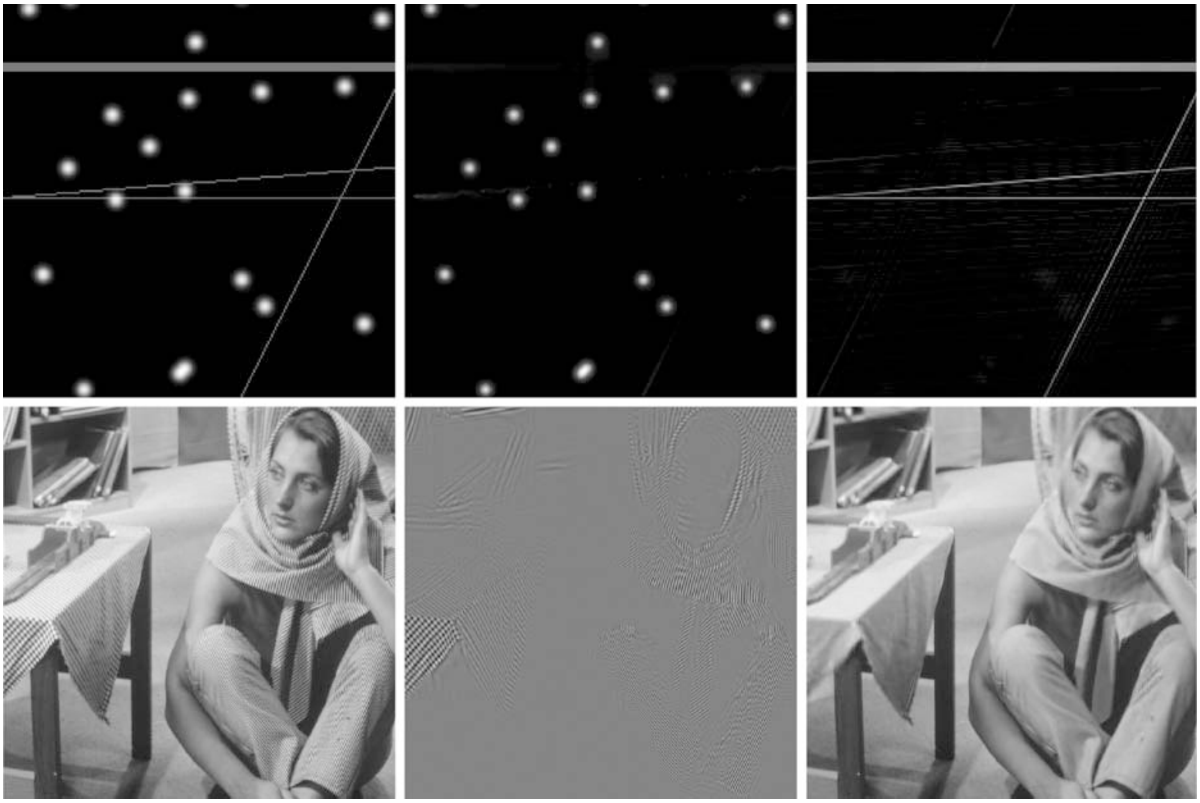
Curvelets and Ridgelets, Figure 17
Illustration of the use of curvelets (DCTG2 transform) when solving two typical linear inverse problems: deconvolution (*first row*), and inpainting (*second row*). First row: deconvolution of `Barbara` image, original (*left*), blurred and noisy (*middle*), restored (*right*). Second row: inpainting of `Claudia` image, original (*left*), masked image (*middle*), inpainted (*right*)



Curvelets and Ridgelets, Figure 18
Curvelet contrast enhancement. Left: enhanced vs original curvelet coefficient. Middle: original `Saturn` image. Right: result of curvelet-based contrast enhancement

form and the curvelet transform [70]. A full description of MCA is given in [70].

The first row of Fig. 19 illustrates a separation result when the input image contains only lines and isotropic Gaussians. Two transforms were amalgamated in the dictionary; namely the à trous WT2D and the DRT. The left, middle and right images in the first row of Fig. 19 represent respectively, the original image, the reconstructed



Curvelets and Ridgelets, Figure 19
First row, **from *left* to *right*: original image containing lines and Gaussians, separated Gaussian component (*wavelets*), separated line component (*ridgelets*). *Second row*, from *left* to *right*: original `Barbara` image, reconstructed local discrete cosine transform part (*texture*), and piecewise smooth part (*curvelets*)**

component from the à trous wavelet coefficients, and the reconstructed layer from the ridgelet coefficients. The second row of Fig. 19 shows respectively the `Barbara` image, the reconstructed local cosine (textured) component and the reconstructed curvelet component. In the `Barbara` example, the dictionary contained the local discrete cosine and the DCTG2 transforms.

Future Directions

In this paper, we gave an overview of two important geometrical multiscale transforms; namely ridgelets and curvelets. We illustrate their potential applicability on a wide range of image processing problems. Although these transforms are not adaptive, they are strikingly effective both theoretically and practically on piecewise images away from smooth contours.

However, in image processing, the geometry of the image and its regularity is generally not known in advance. Therefore, to reach higher sparsity levels, it is necessary to find representations that can adapt themselves to the geometrical content of the image. For instance, geometric transforms such as wedgelets [30] or bandlets [49,50,62] allow to define an adapted multiscale geometry. These transforms perform a non-linear search for an optimal representation. They offer geometrical adaptivity together with stable algorithms. Recently, Mallat [54] proposed a more biologically inspired procedure named the grouplet transform, which defines a multiscale association field by grouping together pairs of wavelet coefficients.

In imaging science and technology, there is a remarkable proliferation of new data types. Beside the traditional data arrays defined on uniformly sampled cartesian grids with scalar-valued samples, many novel imaging modalities involve data arrays that are either (or both):

- acquired on specific “exotic” grids such as in astronomy, medicine and physics. Examples include data defined on spherical manifolds such as in astronomical imaging, catadioptric optical imaging where a sensor overlooks a paraboloidal mirror, etc.

- or with samples taking values in a manifold. Examples include vector fields such as those of polarization data that may rise in astronomy, rigid motions (a special Euclidean group), definite-positive matrices that are encountered in earth science or medical imaging, etc.

The challenge faced with this data is to find multiscale representations which are sufficiently flexible to apply to many data types and yet defined on the proper grid and respect the manifold structure. Extension of wavelets, curvelets and ridgelets for scalar-valued data on the sphere has been proposed recently by [66]. Construction of wavelets for scalar-valued data defined on graphs and some manifolds was proposed by [22]. The authors in [63][see references therein] describe multiscale representations for data observed on equispaced grids and taking values in manifolds such as: the sphere, the special orthogonal group, the positive definite matrices, and the Grassmannian manifolds. Nonetheless many challenging questions are still open in this field: extend the idea of multiscale geometrical representations such as curvelets or ridgelets to manifold-valued data, find multiscale geometrical representations which are sufficiently general for a wide class of grids, etc. We believe that these directions are one of the hottest topics in this field.

Most of the transforms discussed in this paper can handle efficiently smooth or piecewise smooth functions. But sparsely representing textures remains an important open question, mainly because there is no consensus on how to define a texture. Although Julesz [45] stated simple axioms about the probabilistic characterization of textures. It has been known for some time now that some transforms can sometimes enjoy reasonably sparse expansions of certain textures; e. g. oscillatory textures in bases such as local discrete cosines [70], brushlets [56], Gabor [53], complex wavelets [46]. Gabor and wavelets are widely used in the image processing community for texture analysis. But little is known on the decay of Gabor and wavelet coefficients of "texture". If one is interested in synthesis as well as analysis, the Gabor representation may be useless (at least in its strict definition). Restricting themselves to locally oscillating patters, Demanet and Ying have recently proposed a wavelet-packet construction named WaveAtoms [77]. They showed that WaveAtoms provide optimally sparse representation of warped oscillatory textures.

Another line of active research in sparse multiscale transforms was initiated by the seminal work of Olshausen and Field [59]. Following their footprints, one can push one step forward the idea of adaptive sparse representation and requires that the dictionary is not fixed but rather optimized to sparsify a set of exemplar signals/images, i. e. patches. Such a learning problem corresponds to finding a sparse matrix factorization and several algorithms have been proposed for this task in the literature; see [1] for a good overview. Explicit structural constraints such as translation invariance can also be enforced on the learned dictionary [5,58]. These learning-based sparse representations have shown a great improvement over fixed (and even adapted) transforms for a variety of image processing tasks such as denoising and compression [8,36,52], linear inverse problems (image decomposition and inpainting) [61], texture synthesis [60].

Bibliography

1. Aharon M, Elad M, Bruckstein AM (2006) The K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process 54(11):4311–4322
2. Arivazhagan S, Ganesan L, Kumar TS (2006) Texture classification using curvelet statistical and co-occurrence feature. In: Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong (ICPR 2006), vol 2. pp 938–941
3. Averbuch A, Coifman RR, Donoho DL, Israeli M, Waldén J (2001) Fast Slant Stack: A notion of Radon transform for data in a cartesian grid which is rapidly computible, algebraically exact, geometrically faithful and invertible. SIAM J Sci Comput
4. BeamLab 200 (2003) http://www-stat.stanford.edu/beamlab/
5. Blumensath T, Davies M (2006) Sparse and shift-invariant representations of music. IEEE Trans Speech Audio Process 14(1):50–57
6. Bobin J, Moudden Y, Starck JL, Elad M (2006) Morphological diversity and source separation. IEEE Trans Signal Process 13(7):409–412
7. Bobin J, Starck JL, Fadili J, Moudden Y (2007) Sparsity, morphological diversity and blind source separation. IEEE Transon Image Process 16(11):2662–2674
8. Bryt O, Elad M (2008) Compression of facial images using the k-svd algorithm. J Vis Commun Image Represent 19(4):270–283
9. Candès EJ (1998) Ridgelets: theory and applications. Ph D thesis, Stanford University
10. Candès EJ (1999) Harmonic analysis of neural networks. Appl Comput Harmon Anal 6:197–218
11. Candès EJ (1999) Ridgelets and the representation of mutilated sobolev functions. SIAM J Math Anal 33:197–218
12. Candès EJ (1999) Ridgelets: Estimating with ridge functions. Ann Statist 31:1561–1599
13. Candès EJ, Demanet L (2002) Curvelets and Fourier integral operators. C R Acad Sci, Paris, Serie I(336):395–398
14. Candès EJ, Demanet L (2005) The curvelet representation of wave propagators is optimally sparse. Com Pure Appl Math 58(11):1472–1528
15. Candès EJ, Demanet L, Donoho DL, Ying L (2006) Fast discrete curvelet transforms. SIAM Multiscale Model Simul 5(3):861–899

16. Candès EJ, Donoho DL (1999) Ridgelets : the key to high dimensional intermittency? Philos Trans Royal Soc of Lond A 357:2495–2509
17. Candès EJ, Donoho DL (1999) Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In: Cohen A, Rabut C, Schumaker LL (eds) Curve and Surface Fitting: Saint-Malo. Vanderbilt University Press, Nashville
18. Candès EJ, Donoho DL (2000) Curvelets and curvilinear integrals. Approx J Theory 113:59–90
19. Candès EJ, Donoho DL (2000) Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. Ann Stat 30:784–842
20. Candès EJ, Donoho DL (2002) New tight frames of curvelets and optimal representations of objects with piecewise-C2 singularities. Comm Pure Appl Math 57:219–266
21. Candès E,Ying L, Demanet L (2005) 3d discrete curvelet transform. In: Wavelets XI conf., San Diego. Proc. SPIE, vol 5914, 591413; doi:10.1117/12.616205
22. Coifman RR, Maggioni M (2006) Diffusion wavelets. Appl Comput Harmon Anal 21:53–94
23. Daubechies I (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia
24. Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Com Pure Appl Math 57:1413–1541
25. Demanet L, Ying L (2007) Curvelets and wave atoms for mirror-extended images. In: Wavelets XII conf., San Diego. Proc. SPIE, vol 6701, 67010J; doi:10.1117/12.733257
26. Do MN, Vetterli M (2003) The finite ridgelet transform for image representation. IEEE Trans Image Process 12(1):16–28
27. Do MN, Vetterli M (2003) Contourlets. In: Stoeckler J, Welland GV (eds) Beyond Wavelets. Academic Press, San Diego
28. Donoho DL (1997) Fast ridgelet transforms in dimension 2. Stanford University, Stanford
29. Donoho DL (1998) Digital ridgelet transform via rectopolar coordinate transform. Technical Report, Stanford University
30. Donoho DL (1999) Wedgelets: nearly-minimax estimation of edges. Ann Stat. 27:859–897
31. Donoho DL (2000) Orthonormal ridgelets and linear singularities. SIAM J Math Anal 31(5):1062–1099
32. Donoho DL, Duncan MR (2000) Digital curvelet transform: strategy, implementation and experiments. In: Szu HH, Vetterli M, Campbell W, Buss JR (eds) Proc. Aerosense, Wavelet Applications VII, vol 4056. SPIE, pp 12–29
33. Donoho DL, Flesia AG (2002) Digital ridgelet transform based on true ridge functions. In: Schmeidler J, Welland GV (eds) Beyond Wavelets. Academic Press, San Diego
34. Douma H, de Hoop MV (2007) Leading-order seismic imaging using curvelets. Geophys 72(6)
35. Durand S (2007) M-band filtering and nonredundant directional wavelets. Appl Comput Harmon Anal 22(1):124–139
36. Elad M, Aharon M (2006) Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans Image Process 15(12):3736–3745
37. Elad M, Starck JL, Donoho DL, Querre P (2006) Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). J Appl Comput Harmon Anal 19:340–358
38. Fadili MJ, Starck JL (2006) Sparse representation-based image deconvolution by iterative thresholding. In: Murtagh F, Starck JL (eds) Astronomical Data Analysis IV. Marseille
39. Fadili MJ, Starck JL, Murtagh F (2006) Inpainting and zooming using sparse representations. Comput J
40. Fernandes F, Wakin M, Baraniuk R (2004) Non-redundant, linear-phase, semi-orthogonal, directional complex wavelets. In: IEEE Conf. on Acoustics, Speech and Signal Processing, Montreal, Canada, vol 2, pp 953 956
41. Figueiredo M, Nowak R (2003) An EM algorithm for wavelet-based image restoration. IEEE Trans Image Process 12(8):906–916
42. Hennenfent G, Herrmann FJ (2006) Seismic denoising with nonuniformly sampled curvelets. IEEE Comput Sci Eng 8(3):16–25
43. Herrmann FJ, Moghaddam PP, Stolk CC (2008) Sparsity- and continuity-promoting seismic image recovery with curvelet frames. Appl Comput Harmon Anal 24(2):150–173
44. Jin J, Starck JL, Donoho DL, Aghanim N, Forni O (2005) Cosmological non-gaussian signatures detection: Comparison of statistical tests. Eurasip J 15:2470–2485
45. Julesz B (1962) Visual pattern discrimination. Trans RE Inform Theory 8(2):84–92
46. Kingsbury NG (1998) The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement. In: European Signal Processing Conference, Rhodes, Greece, pp 319–322
47. Labate D, Lim W-Q, Kutyniok G, Weiss G (2005) Sparse multidimensional representation using shearlets. In: Wavelets XI, vol 5914. SPIE, San Diego, pp 254–262
48. Lambert P, Pires S, Ballot J, García RA, Starck JL, Turck-Chièze S (2006) Curvelet analysis of asteroseismic data. i. method description and application to simulated sun-like stars. Astron Astrophys 454:1021–1027
49. Le Pennec E, Mallat S (2000) Image processing with geometrical wavelets. In: International Conference on Image Processing, Thessaloniki, Greece
50. Le Pennec E, Mallat S (2005) Bandelet Image Approximation and Compression. Multiscale SIAM Mod Simul 4(3):992–1039
51. Lu Y, Do MN (2003) CRIPS-contourlets: A critical sampled directional multiresolution image representation. In: Wavelet X, San Diego. Proc. SPIE, vol 5207. pp 655–665
52. Mairal J, Elad M, Sapiro G (2008) Sparse representation for color image restoration. IEEE Trans Image Process 17(1):53–69
53. Mallat S (1998) A Wavelet Tour of Signal Processing. Academic Press, London
54. Mallat S (2008) Geometrical grouplets. Appl Comput Harmon Anal
55. Matus F, Flusser J (1993) Image representations via a finite Radon transform. IEEE Trans Pattern Anal Mach Intell 15(10):996–1006
56. Meyer FG, Coifman RR (1997) Brushlets: a tool for directional image analysis and image compression. Appl Comput Harmon Anal 4:147–187
57. Meyer Y (1993) Wavelets: Algorithms and Applications. SIAM, Philadelphia
58. Olshausen BA (2000) Sparse coding of time-varying natural images. In: Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Barcelona, pp 603–608
59. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. Nature 381(6583):607–609

60. Peyré G (2007) Non-negative sparse modeling of textures. In: SSVM. Lecture Notes in Computer Science. Springer, pp 628–639
61. Peyré G, Fadili MJ, Starck JL (2007) Learning adapted dictionaries for geometry and texture separation. In: Wavelet XII, San Diego. Proc. SPIE, vol 6701, 67011T; doi:10.1117/12.731244
62. Peyré G, Mallat S (2007) A review of bandlet methods for geometrical image representation. Num Algorithms 44(3):205–234
63. Rahman I, Drori I, Stodden VC, Donoho DL, Schröder P (2005) Multiscale representations fpr manifold-valued data. Multiscale Mod Simul 4(4):1201–1232
64. Saevarsson B, Sveinsson J, Benediktsson J (2003) Speckle reduction of SAR images using adaptive curvelet domain. In: Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium, IGARSS '03, vol 6., pp 4083–4085
65. Simoncelli EP, Freeman WT, Adelson EH, Heeger DJ (1992) Shiftable multi-scale transforms [or what's wrong with orthonormal wavelets]. IEEE Trans Inf Theory 38(2):587–607
66. Starck JL, Abrial P, Moudden Y, Nguyen M (2006) Wavelets, ridgelets and curvelets on the sphere. Astron Astrophys 446:1191–1204
67. Starck JL, Bijaoui A, Lopez B, Perrier C (1994) Image reconstruction by the wavelet transform applied to aperture synthesis. Astron Astrophys 283:349–360
68. Starck JL, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. IEEE Trans Image Process 11(6):131–141
69. Starck JL, Candès EJ, Donoho DL (2003) Astronomical image representation by the curvelet tansform. Astron Astrophys 398:785–800
70. Starck JL, Elad M, Donoho DL (2004) Redundant multiscale transforms and their application for morphological component analysis. In: Hawkes P (ed) Advances in Imaging and Electron Physics, vol 132. Academic Press, San Diego, London, pp 288–348
71. Starck JL, Elad M, Donoho DL (2005) Image decomposition via the combination of sparse representation and a variational approach. IEEE Trans Image Process 14(10):1570–1582
72. Starck JL, Murtagh F, Bijaoui A (1998) Image Processing and Data Analysis: The Multiscale Approach. Cambridge University Press, Cambridge
73. Starck JL, Murtagh F, Candès E, Donoho DL (2003) Gray and color image contrast enhancement by the curvelet transform. IEEE Trans Image Process 12(6):706–717
74. Starck JL, Nguyen MK, Murtagh F (2003) Wavelets and curvelets for image deconvolution: a combined approach. Signal Process 83(10):2279–2283
75. The Curvelab Toolbox (2005) http://www.curvelet.org
76. Velisavljevic V, Beferull-Lozano B, Vetterli M, Dragotti PL (2006) Directionlets: Anisotropic multi-directional representation with separable filtering. IEEE Trans Image Process 15(7):1916–1933
77. Ying L, Demanet L (2007) Wave atoms and sparsity of oscillatory patterns. Appl Comput Harmon Anal 23(3):368–387
78. Zhang Z, Huang W, Zhang J, Yu H, Lu Y (2006) Digital image watermark algorithm in the curvelet domain. In: Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP'06), Pasadena. pp 105–108

Cytoskeleton and Cell Motility

THOMAS RISLER[1,2]

[1] Institut Curie, Section de Recherche, Laboratoire Physicochimie Curie (CNRS-UMR 168), Paris, France

[2] Université Pierre et Marie Curie Paris VI, Paris, France

Article Outline

Glossary

Cell Structural and functional elementary unit of all life forms. The cell is the smallest unit that can be characterized as living.

Eukaryotic cell Cell that possesses a nucleus, a small membrane-bounded compartment that contains the genetic material of the cell. Cells that lack a nucleus are called *prokaryotic cells* or *prokaryotes*.

Domains of life *archaea*, *bacteria* and *eukarya* – or in English *eukaryotes*, and made of eukaryotic cells – which constitute the three fundamental branches in which all life forms are classified. Archaea and bacteria are prokaryotes. All multicellular organisms are eukaryotes, but eukaryotes can also be single-cell organisms. Eukaryotes are usually classified into four kingdoms: animals, plants, fungi and protists.

Motility Spontaneous, self-generated movement of a biological system.

Cytoskeleton System of protein filaments crisscrossing the inner part of the cell and which, with the help of the many proteins that interact with it, enables the cell to insure its structural integrity and morphology, exert forces and produce motion.

Amoeboid motility Crawling locomotion of a eukaryotic cell by means of protrusion of its leading edge.

Molecular motor Motor of molecular size. In this context, protein or macromolecular complex that converts a specific source of energy into mechanical work.

Filament Here, extended unidimensional structure made of an assembly of repeated protein units that hold together via physical interactions (without covalent bonds). A filament will be either a single *polymer* (or here *biopolymer*), a linear assembly of such polymers, or a linear assembly of molecular motors.

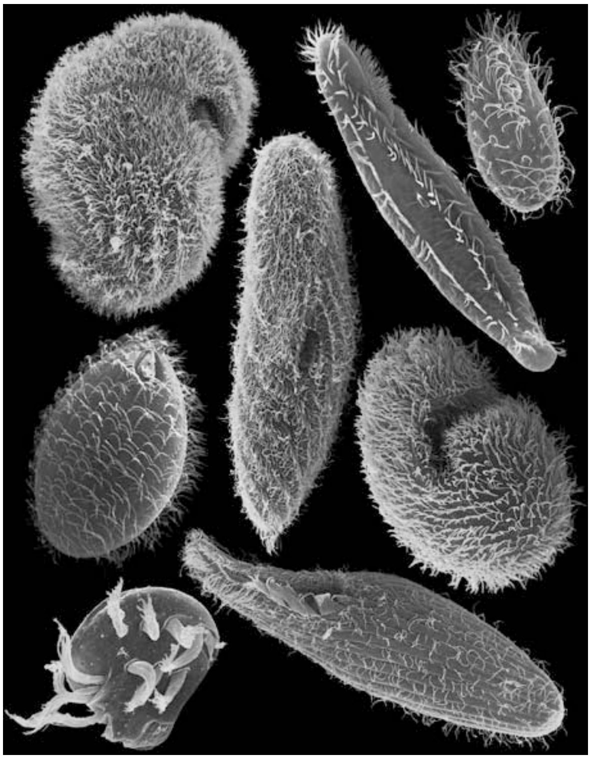
Active gel Cross-linked network of linear or branched polymers interacting by physical means, and that is dynamically driven out of equilibrium by a source of energy.

Definition of the Subject

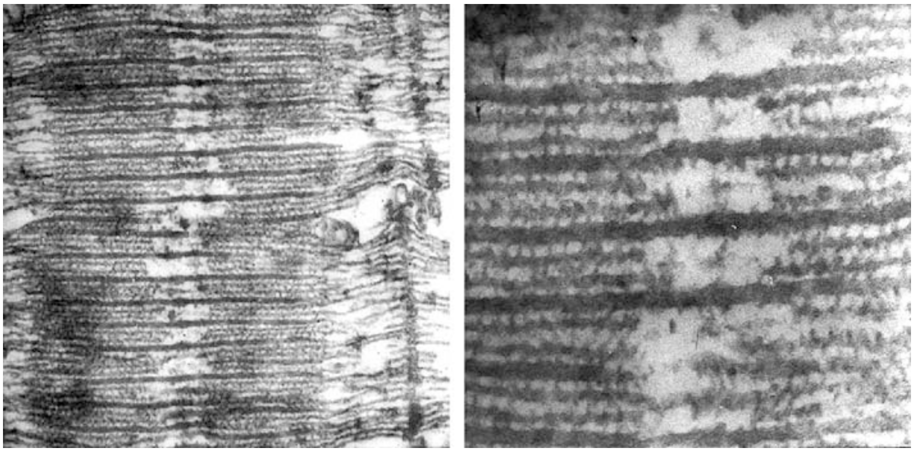
We, as human beings, are made of a collection of cells, which are most commonly considered as the elementary building blocks of all living forms on earth [5]. Whether they belong to each of the three domains of life (*archaea*, *bacteria* or *eukarya*), cells are small membrane-bounded compartments that are capable of homeostasis, metabolism, response to their environment, growth, reproduction, adaptation through evolution and, at the cellular as well as multicellular level, organization. In addition, spontaneous, self-generated movement – also known as *motility* – is one of the properties that we most closely associate with all life forms. Even in the case of apparently inanimate living forms on macroscopic scales, like most plants and fungi, constitutive cells are constantly remodeling their internal structure for the entire organism to perform its metabolism, growth and reproduction [29]. In animals like human beings, cell motility is at the basis of most – if not all – essential processes participating in their lifetime, from their development, maintenance, to eventual death. It is indeed crucially involved for example in embryonic development (where individual as well as collective motions of cells underly morphogenesis), wound healing and recovery from injuries (where cellular migration is essential for tissue repair and regeneration), as well as immune response and most of disease progressions. There, on a biomedical point of view, cellular motility is involved in processes as diverse as neutrophils (white blood cells) and macrophages (cells that ingest bacteria) progressions, axonal regrowth after injuries, multiple sclerosis and cancer metastases. In addition, motility defects of the animal cells themselves can lead to a variety of inherited health problems, including male infertility, deafness and chronic inflammatory diseases.

Introduction

Cell movement was observed and reported for the first time as early as 1674, when Anthony van Leeuwenhoek brought a glass bead that served him as a primitive microscope close to a drop of water taken from a pool. His astonishment was immediate, as he later reported: "… the motion of these animalcules in the water was so swift and various, upwards, downwards and round about, that it was wonderful to see …" [29]. The organisms he saw were probably ciliated *protozoa* – unicellular non-photosynthetic eukaryotic organisms – a fraction of a millimeter in length, swimming by the agitated but coordinated motion of sometimes thousands of hairlike cilia on their surface (see Fig. 1). Despite this very early observation, only relatively recent advances of the past few decades in microscopy, molecular biology and biochemistry have enabled the discovery of the basic underlying molecular mechanisms by which cells are able to feel their environment, exert forces and move in a directed way in search for nutrients or any other task they need to perform. The cytoskeleton, defined as the system of protein filaments that enable the cell to insure its structural integrity and morphology, exert forces and produce motion, was first observed by H. E. Huxley and J. Hanson in



Cytoskeleton and Cell Motility, Figure 1
Electron micrographs of different species of ciliated *protozoa*. Almost all members of the protozoan group are non-pathogenic free-living organisms. Source: Foissner, W. and Zankl, A. (unpublished)



Cytoskeleton and Cell Motility, Figure 2

Skeletal-muscle thick and thin filaments as seen by H. E. Huxley in 1957 [112] (Reproductions from the original 1953 paper [93] were poor). *Left panel*: Thin longitudinal sections of rabbit psoas muscle fibers, showing a single layer of a filament lattice, with individual thick and thin filaments as well as crossbridges between them. *Right panel*: Higher-magnification view of a thin longitudinal section. The relative dimensions were distorted due to axial compression during sectioning: crossbridges' axial spacing is ≃40 nm and thick filaments' diameter is ≃12 nm. Source: reprinted from [114] with permission from Blackwell Publishing Ltd. (based on an original micrograph of 1957, see also [112])

1953, when they discovered the double array of filaments in cross-striated muscles using electron-microscopy techniques [93,110,111]. In parallel with A. F. Huxley and R. Niedergerke, but independently, they published the next year the "sliding-filament model", which explains muscle contraction via the relative sliding of two different types of filaments, originally called "thick" and "thin" filaments [109,115] (see Fig. 2). This, with the help of further genetic, biochemical and crystallographic studies, dated the beginning of a scientific understanding of the subcellular mechanisms that underly cell motility.

In addition to the characterization of the biochemical composition and organization of these subcellular structures, tremendous advances in the past two decades on both physical micro-manipulation and fluorescence-microscopy techniques have enabled the characterization of the processes involved with minute details. On the one hand, thanks to the help of micro-pipettes, atomic-force microscopes and optical tweezers, one can characterize the forces that cells are able to exert as well as their responses to applied stimuli. On the other hand, fluorescence-microscopy techniques provide information on the microscopic dynamics of single molecules in vivo. Finally, combined with biochemistry and gene-expression control, as well as the micro-fabrication of bio-mimetic artificial systems in in vitro assays, these techniques have enabled the study of simplified systems, where some specific aspects of the processes involved can be characterized separately.

In addition to these biochemical and behavioral characterizations, understanding the generic principles that underly cell motility needed an integrated approach to explain how this complex molecular machinery can self-organize and lead to a coherent, purposeful movement at the cellular level. Nothing better than a eukaryotic cell can indeed be categorized as a complex system, in that its behavior integrates the coordinated interplay of more than ten thousand different protein types, numbering together millions and representing 60% of its dry mass [5]. The cytoskeletal machinery is made of hundreds of different molecular players. For example, in 2003, about 160 proteins were known to bind to actin, one of the major biopolymers participating in cell structure and dynamical behavior [53]. Knowledge about this biomolecular machinery is constantly evolving, and its undergoing complexity can be appreciated by consulting up-to-date information on available databases[1]. Therefore, understanding this complexity and describing how it is integrated at the cellular level was made by biophysical studies, both on experimental and theoretical grounds, which helped to identify the generic principles behind cell motility. At the molecular level first, the conversion of chemical energy stored in covalent bounds into mechanical work relies on out-of-equilibrium thermodynamic principles and asymmetrical properties – or polarity – of the structures involved, and happens in a highly-fluctuating environment of brownian particles [13,233]. On larger length scales, the appearance of coordinated motion in large collections of proteins relies on collective phenomena, self-organi-

[1] See, for example, http://www.cytoskeletons.com/database.php.

zation and dynamical symmetry breakings [129]. On yet larger length scales, swimming of microorganisms has attracted the attention of physicists for years [228,263], and morphogenesis and pattern formations in cellular tissues rely on self-organization phenomena, as was envisioned for the first time by Turing in 1952 [141,269]. Therefore, in addition to biophysical experimental techniques, variety of theoretical physics' disciplines spanning the theory of stochastic processes, statistical physics, out-of-equilibrium thermodynamics, hydrodynamics, nonlinear dynamics and pattern formation have contributed and still contribute to our understanding of cell motility.

The present article will mainly focus on the eukaryotic cytoskeleton and cell-motility mechanisms. Bacterial motility as well as the composition of the prokaryotic cytoskeleton will be only briefly mentioned. The article is organized as follows. In Sect. "The Diversity of Cell Motility", we will first present an overview of the diversity of cellular motility mechanisms, which might at first glance be categorized into two different types of behaviors, namely "swimming" and "crawling". Intracellular transport, mitosis – or cell division – as well as other extensions of cell motility that rely on the same essential machinery will be briefly sketched. In Sect. "The Cell Cytoskeleton", we will introduce the molecular machinery that underlies cell motility – the cytoskeleton – as well as its interactions with the external environment of the cell and its main regulatory pathways. Sections "Other Cytoskeleton-Associated Proteins" to "The Prokaryotic Cytoskeleton" are more detailed in their biochemical presentations; readers primarily interested in the theoretical modeling of cell motility might want to skip these sections in a first reading. We will then describe the motility mechanisms that rely essentially on polymerization-depolymerization dynamics of cytoskeleton filaments in Sect. "Filament-Driven Motility", and the ones that rely essentially on the activity of motor proteins in Sect. "Motor-Driven Motility". Finally, Sect. "Putting It Together: Active Polymer Solutions" will be devoted to the description of the integrated approaches that have been developed recently to try to understand the cooperative phenomena that underly self-organization of the cell cytoskeleton as a whole.

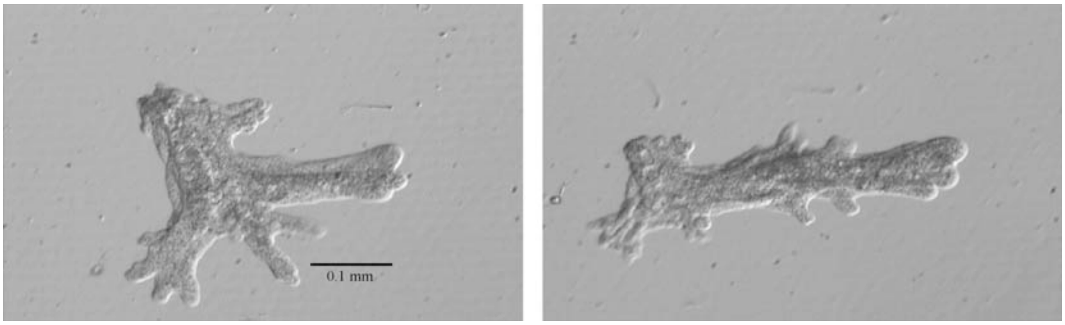
The Diversity of Cell Motility

Swimming

At the cellular level, viscous hydrodynamic forces are several orders of magnitude higher than inertial forces. Therefore, simple reciprocal motions cannot produce forward motion, and cellular swimming patterns need to be asymmetric in space and time for the cell to advance. This hydrodynamic problem faced by cells attempting to swim have been eloquently summarized by Purcell as "life at low Reynolds number" [228]. To solve this problem, bacteria use the rotation of a short helical or corckscrew-shaped *flagellum*, which is a relatively rigid structure made of a collection of hundreds of identical protein subunits called *flagellins* [20]. The swimming of a single bacterium can be impressively rapid, as bacteria such as *Escherichia coli* - the common intestinal bacterium – swim at speeds of 20 to 30 micrometers per second, for the cell itself is only about two-micrometer long and half a micrometer in diameter. The bacterium possesses multiple flagella that gather together during swimming, and can fly apart as the bacterium switches direction. Other bacteria such as *Vibrio cholerae* - the causative agent of cholera – use a single flagellum located at one of their pole, but the propulsion mechanism relies always on the presence of a rotary molecular motor located in the cell membrane, and which is sensitive to modifications of the chemical environment of the cell. Under normal conditions, the bacterium changes direction in an intermittent chaotic way by reversing the rotational direction of its motors, a phenomenon known as *tumbling*. When placed in a concentration gradient of nutrients however, the cell can adapt its tumbling frequency to swim towards nutrient-rich regions, a phenomenon known as *chemotaxis* [181].

Even though it shares the same name, the eukaryotic flagellum shares little structures and propulsion mechanisms with its bacterial counterpart. It is indeed at least ten times larger than a bacterial flagellum in both diameter and length, and instead of being a rigid passive structure animated by a remote motor, it bears its motor activity along its length. Propulsion occurs by the propagation of a bending wave along the flagellum as a result of the relative sliding of a group of about 10 long parallel filaments, which are engulfed in the cell's plasma membrane and are animated by hundreds of motor proteins in a coordinated manner. Eukaryotic cells also use another type of protrusions to swim, the *cilia*, which are much like flagella in their internal structure, but which are shorter and work usually in numbers, covering sometimes the whole cell surface as in the case of paramecia or other ciliated protozoa (see Fig. 1). Their beating pattern is then coordinated at the cellular level, most often in a wave-type of manner known as the *metachronal wave*. Beating cilia are also found in animals, as for example in humans where ciliated cells play major roles in several organs like the brain, the retina, the respiratory tract, the Fallopian tube and the kidney [117].

Other strategies of swimming include the elegant movement used by *Eutreptiella* - called *metaboly* – which



Cytoskeleton and Cell Motility, Figure 3
Two pictures of *amoeba proteus* displaying different shapes of its pseudopodia. Note the dramatic change in cell shape during locomotion. Source: courtesy of Sutherland Maciver

consists in gradually changing the contour of the cell surface to locally increase the drag exerted by the viscous fluid around and move the cell forward [67]. Other organisms like most motile species of *Chrysophytes* – a group of marine photosynthetic protozoa – possess a flagellum attached at their front instead of their back. The flagellum is covered with stiff hairs projecting from its side that allow the cell to move forward as a planar wave propagates from the base to the tip of the flagellum [29]. Finally, one should mention yet another type of motility – namely *walking* – in which cells use also cilia and flagella animated in a coordinated manner to enable the cell to literally "walk" over surfaces. Walking motility relies on the same essential biochemical structures as the ones employed in swimming with collections of cilia.

Crawling

Cell crawling is the common mechanism employed by most eukaryotic animal cells as they move through animal tissues, constituted of other cells or filaments of the extracellular matrix [5,29]. In contrast to swimming cells, crawling cells in general do not employ conspicuous motile organelles that are external to the cell, and which can be studied in isolation. In general however, they either move by means of wormlike cycles of extensions and contractions of their cell body or of some specific protrusions, or slide without visible means of protrusion, a process also referred to as *gliding*. Most of crawling mechanisms rely on the protrusion of specific dynamic extensions at the leading edge of the cell, but gliding seems to rely sometimes on different mechanisms [67,96]. Even though gliding mechanisms are widespread in bacteria, algae and parasitic protozoa, we still do not know for sure the molecular machinery as well as the essential mechanisms that underly these different phenomena [29,96].

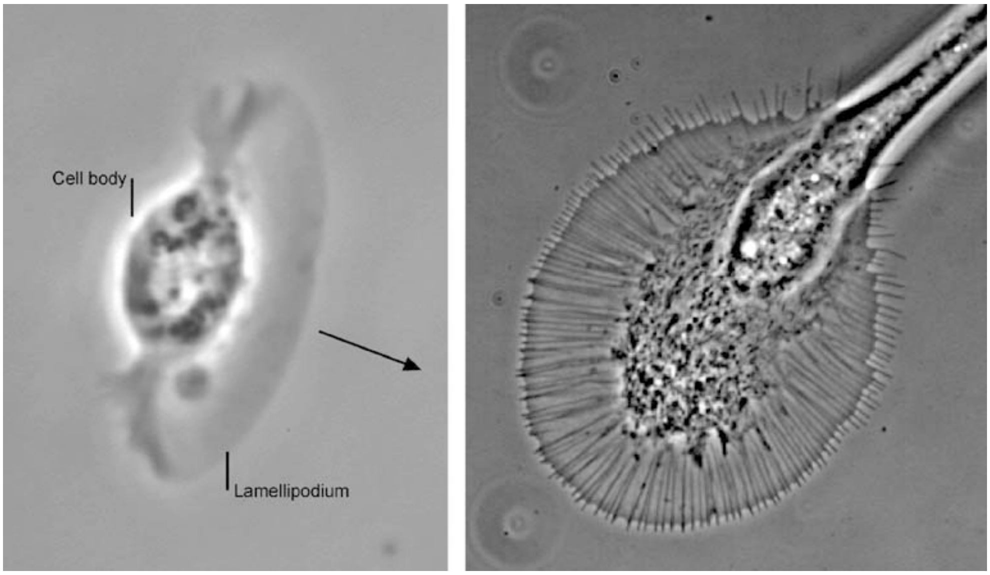
The best characterized crawling mechanism is the so-called *amoeboid motility*, referring to the locomotion of all eukaryotic cells that move by means of protrusion of their leading edge. Originally, the term was referring uniquely to the crawling mechanism of *Amoeba proteus*, a particular species of *amoebae*, whose protrusions are stubby three-dimensional projections called *pseudopodia*[2] (see Fig. 3)[3]. But other types of protrusions exist, that are classified with respect to their shape and dimensional organization. Two-dimensional protrusions are the flat veil-shaped projections called *lamellipodia*, as they occur in fibroblasts' or fish epidermal keratocytes' motility for wound healing[4]. One-dimensional projections are the long thin projections called either *filopodia* or *microspikes*, and which occur for example in neuronal-growth-cone progressions[5]. Filopodia usually protrude as small extensions of a lamellipodium, and are used by the cell to extend its lamellipodium in a given direction [136,257] (see Fig. 4). To

[2] Note that some zoologists also use the term "pseudopodia" or "pseudopods" rather generally to refer to a variety of cell-surface protrusions. These include the different types of protrusions described here as playing a role in amoeboid motility, but also the long extended processes that some cell types use only as feeding apparatus, like *axopodia*.

[3] A short video of a locomoting *Amoeba proteus* can be seen on the following website: http://www.bms.ed.ac.uk/research/others/smaciver/A.prot.Loc.mov.

[4] Fibroblasts are the cells that synthesize and maintain the extracellular matrix in most animal connective tissues. They provide a structural framework (stroma) for many tissues, and play a crucial role in wound healing. Keratocytes are epithelial cells that have been characterized in the epidermis of fish and frogs, and that have been named so because of their abundant keratin filaments. They are specialized in wound healing, and are one of the most spectacular example of fast and persistent locomotion in cells, with velocities up to 30 μm/min [5,158].

[5] Growth cones are structures that are found at the tip of axons and dendrites, by means of which neuron cells extend.



Cytoskeleton and Cell Motility, Figure 4
Two examples of lamellipodia. *Left panel*: A living fish keratocyte extends its leading lamellipodium during crawling. This is a phase-contrast micrograph, a single frame from a video sequence (the whole movie can be seen on the following webpage: http://cmgm.stanford.edu/theriot/movies.htm.). The lamellipodium and the cell body are labeled. The *large arrow* indicates the direction of motion. Source: reprinted from [36] with permission from Nature Publishing Group. Right panel: Snail neuronal growth cone by means of which the nerve fiber elongates at its tip. Clearly visible are the radially-aligned bundle structures that project into filopodia at the leading edge of the lamellipodium. Source: courtesy of Feng-quan Zhou; reprinted from [291] with permission from Rockefeller University Press

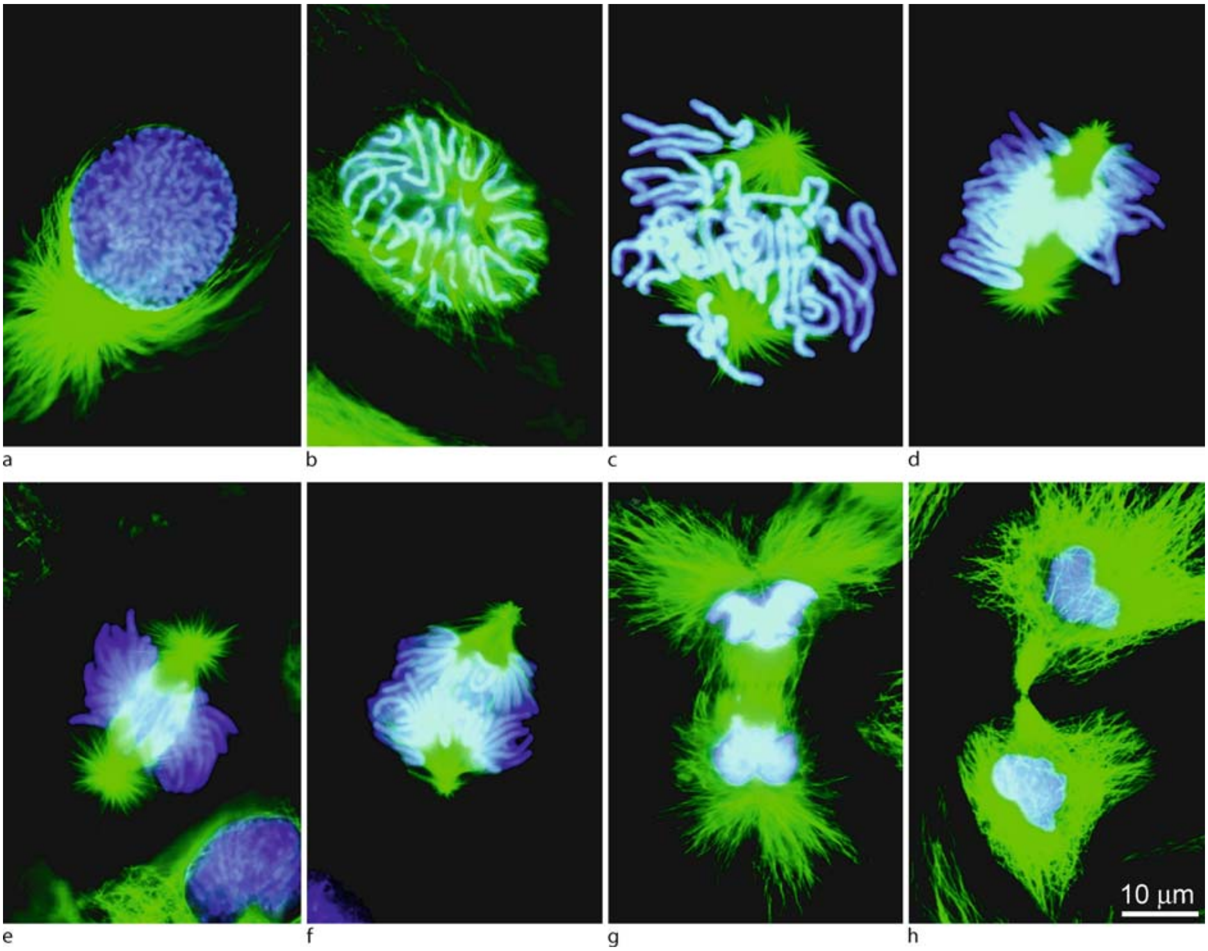
these must be added the spherical membrane protrusions called *blebs*, which occur as a result of cortical contractility, and which have been proposed recently to participate in the initiation of lamellipodium formation and the elaboration of cell polarity for directed motion [211], as well as in the amoeboid motility itself for example in *Dictyostelium*, a model species of *amoebae* [289]. Finally, one should mention a motility mechanism that can be classified as *rolling*, in which some organisms such as *helizoa* use coordinated shortening and lengthening of long radiating needlike extensions called *axopodia* to roll over surfaces. Axopodia happen also to be sticky extensions that are most often used for catching preys in numerous protozoa.

Depending on authors, the process of amoeboid motility can be decomposed into three to five steps that occur simultaneously. First the cell makes a protrusion, where the membrane is pushed forward by means of the polymerization of cytoskeletal filaments. Then the protrusion adheres to the substrate via the formation of anchoring points, and subsequent contraction of the cell cytoskeleton drags the cell body forward. Finally at the rear end, the cell de-adheres and retracts [156,192]. Of these five steps, the two last ones – namely de-adhesion and retraction – involve similar structures and mechanisms as the formation of the anchoring points and cell body drag, which led originally Abercrombie to describe his observation as a three-step cycle [1]. The speed of amoeboid motility can range from less than a micrometer per hour to more than one micrometer per second, depending on cell type and stimulation[6].

Extensions of Cell Motility

In addition to moving the whole cell body, the machinery that is responsible for cellular movement can be employed for quite different tasks, which are as essential to the cell survival and reproduction as its motility per se. As we have earlier stated, even in the case of macroscopically non-aminated live forms, the constitutive cells need constantly to displace their internal organelles for their metabolism to be maintained [241]. When looked under the light microscope, mitochondria, vesicles, lysosomes and ingested particles display a rapid and sporadic movement that is interspaced with relatively long periods of quiescence. Velocities are typically of the order of micrometers per second, as the fastest known organelle transport is performed in the

[6] Cell-motility videos can be seen at http://cellix.imba.oeaw.ac.at.



Cytoskeleton and Cell Motility, Figure 5
a to h Fluorescence micrographs of mitosis in fixed newt lung cells stained with antibodies to reveal the microtubules (MT, *green*) see Sect. "Biopolymers", and with a dye (Hoechst 33342) to reveal the chromosomes (*blue*). The spindle forms as the separating astral MT arrays, associated with each centrosome (a to c), interact with the chromosomes. Once the chromosomes are segregated into daughter nuclei (f and g), new MT-based structures known as stem-bodies form between the new nuclei (g). These play a role in cytokinesis (h), the actual cleavage of the two daughter cells. Source: reprinted from [237] with permission from The American Association for the Advancement of Science

green algae *Chara*, whose chloroplasts are transported at velocities that can achieve 60 micrometers per second [39]. Of all cell types, the need for organelle transport is best illustrated by the mammalian motor neurons whose longest extensions – the axons – even though typically only a few micrometers in diameter, can reach lengths up to one meter. Characteristic times that would be required for a mitochondrion to naturally diffuse that distance in such a geometry range from 10 to 100 years. Instead, membrane vesicles and organelles are actively transported in both directions at speeds of about one to five micrometers per second, which allows the whole journey to be made in just a few days [29]. Finally, the probably most-spectacular event of intracellular transport occurs during the essential process of eukaryotic mitosis, by which duplicated chromosomes are segregated from the mother cell and delivered to each of the nascent daughter cells. For this process to occur, major structural reorganizations of the whole-cell cytoskeleton are needed, during which a large and complex cellular structure – the *mitotic spindle* - assembles and drives the chromosomes apart in a coordinated manner [193] (see Fig. 5).

Cell motility can also occur by means of molecular machineries located outside the cell that needs to move. This is the case in particular for mammalian pathogene bacteria such as *Listeria monocytogenes* and *Shigella* and *Rick-*

ettsia species, but also for some viruses like *vaccinia virus*. These organisms propel themselves within and across the cells they invade by utilizing the cytoskeleton of their hosts [41,83,87,266]. Among these organisms, *Listeria* in particular has become a model organism for studying actin-based motility, a simplified version of the whole amoeboid motility, and which can be seen as a representation of just the first step of this complex process in the original Abercrombie description [36,194,222]. Other particular systems use different specific structures from purely cell-cytoskeleton-based motility. Among these, vertebrate skeletal muscles contrast with standard cellular motility, in that they are structured in enormous multinucleated cells that evolved specifically to generate extremely rapid, repetitive and forceful movements. The cytoplasm of these giant cells is crammed full of a highly-organized, almost-crytsalline array of cytoskeletal filaments, whose only function is to produce contractile forces [29,113]. Another essential event of cell division, namely *cytokinesis* – the actual cleavage of the two daughter cells – also involves the contraction of relatively-sliding cytokeletal filaments, this time under the form of a dividing ring [82].

Finally, other types of cells use mechanisms that do not rely on their cytoskeleton for their motility. Of the most spectaculars is the motility based on the stored mechanochemical energy in some supra-molecular springs, which can then contract at velocities as high as eight centimeters per second [179]. Yet another mechanism relies on some stored purely-elastic energy that allows some insect-eating plants to catch their preys. This is the case for example of the Venus flytrap *Dionaea muscipula*, whose leafs can close in about 100 ms, one of the fastest movements in all plant kingdom. To achieve such a performance, the plant relies on a snap-buckling instability, whose onset is actively controlled by the plant after the arrival of a fly has triggered some biochemical response via the disturbance of mechano-sensitive hairs located inside the trap [69].

The Cell Cytoskeleton

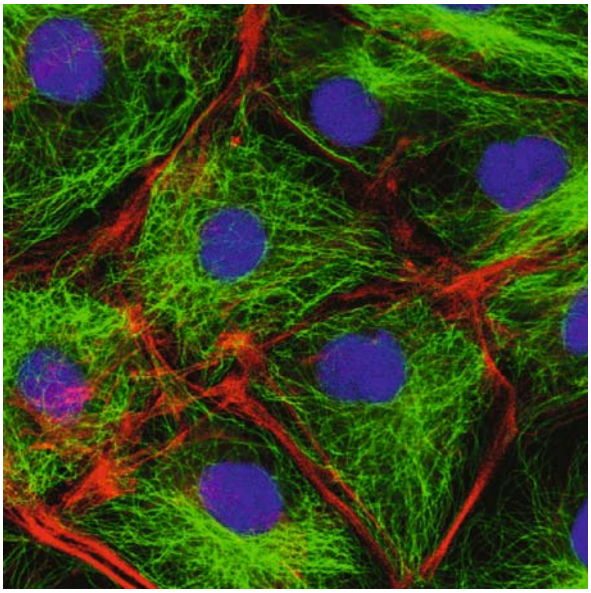
The eukaryotic *cytoskeleton* is defined as the system of protein filaments that enable the cell to insure its structural integrity and rigidity, regulate its shape and morphology, exert forces and produce motion. As a framework that insures structural integrity, the cytoskeleton is mainly constituted of a cohesive meshwork of protein filaments that extend throughout the cytoplasm of the cell. But being the essential structure that produces movement at the cellular level, and thereby needing to be highly adaptable to extracellular stimuli or rapid environmental changes, the cytoskeleton has evolved into a highly-dynamic structure. In fact, cytoskeletal filaments constantly grow and shrink, associate and dissociate via multiple linkages, organize on large scales into a dynamic network, and serve as an intricated set of tracks to motor proteins that transport cargos from one part of the cell to the other, or slide filaments with respect to one another to produce contractile forces. This section is devoted to the biochemical description of this very-complex structure. In addition, its interaction with the cell's external world and its regulatory pathways will be briefly presented, as well the prokaryotic cytoskeleton which, even though biochemically different, appears more and more to resemble its eukaryotic counterpart on a functional point of view.

Biopolymers

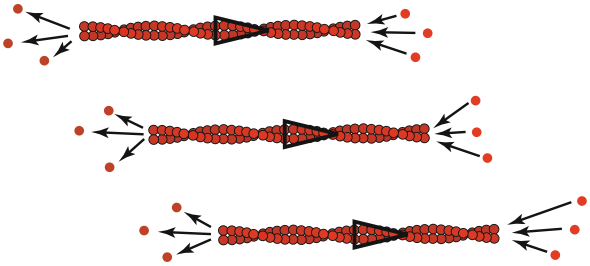
How can a eukaryotic cell, with a diameter of 10 microns or more, be spatially organized by cytoskeletal proteins that are typically 2000 times smaller in linear dimensions? The answer lies in *polymerization*, this ability of the elementary protein subunits (called *monomers*) to assemble via physical interactions into extended linear structures that are made of a large number of them, typically thousands (called thereby *polymers*, or here more precisely *biopolymers*). There are three types of biopolymers in a given eukaryotic cell, namely *actin filaments*, *microtubules* and *intermediate filaments* (see Fig. 6). Although they are classified according to their respective thickness, more interesting for cellular structures and functions are their rigidity, which at thermodynamic equilibrium is characterized by their persistence length L_p [7] [81].

Actin filaments – or F-actin – have a persistence length that is usually accepted to be of the order of 15 to 17 μm [81,210], even though it has been reported that actin rigidity should depend on the way it is decorated, ranging from 9 ± 0.5 μm for bare F-actin to 20 ± 1 μm for tropomyosin-bound actin filaments in skeletal-muscle structures [119]. Actin filaments are two-stranded helical polymers, 5 to 9 nm in diameter, and are built from dimer pairs of globular-actin monomers – or G-actin – that are polar in nature [5,29]. The two halves of an actin monomer are separated by a cleft that can bind adenosine triphos-

[7] The persistence length L_p is defined as follows: consider a thin flexible rod of fixed length L, submitted to thermal forces. Its shape is completely specified by the tangent angle $\theta(s)$ in three dimensions along the arc length of the rod [154]. The persistence length L_p is defined as the characteristic arc length above which thermal fluctuations of the angle $\theta(s)$ become uncorrelated. Specifically, $\langle \cos[\Delta\theta(s)] \rangle = \exp(-s/L_p)$, where $\Delta\theta(s)$ is the three-dimensional angle change over the arc length s. L_p is related to the rod's material Young modulus E and its geometrical moment of inertia I by $L_p = EI/k_B T$, where $k_B T$ represents thermal energy [81].



Cytoskeleton and Cell Motility, Figure 6
Animal cells as seen in the fluorescence microscope after fixation and labeling with specific probes. Actin filaments are stained in *red*, microtubules in *green*, and the nuclei in *blue*. Source: courtesy of Mark Shipman, James Blyth and Louise Cramer, MRC-Laboratory Molecular Cell Biology and Cell Biology Unit, UCL, London UK (unpublished)



Cytoskeleton and Cell Motility, Figure 7
Schematic representation of a treadmilling actin filament. The arrows indicate the polarity of the filaments. Monomers are added to the plus end and removed from the minus end at the same rate, such that while the filament's length remains constant, its center of mass is advancing. Top to bottom shows three subsequent times. Source: courtesy of Karsten Kruse

phate (ATP) or its hydrolyzed form adenosine diphosphate (ADP)[8]. This is responsible for the existence of two distinct ends to the whole filament, namely a fast growing end – called "plus end" or "barbed end" – where mostly ATP-bound monomers are located, and a slow growing end – known as "minus end" or "pointed end" – that is rich in ADP-bound monomers. The minus end has a critical actin-monomer concentration that is roughly six times as high as that of the plus end. At steady state, and with the help of monomeric diffusion, this drives the phenomenon of *treadmilling*, a dynamic evolution of the actin filament where actin monomers are added to the plus end, and removed from the minus end at the same rate. During this process, the total length of the treadmilling filament is kept constant, while its center of mass is displaced at a constant velocity, even though each individual polymerized monomer do not move on average[9] (see Fig. 7). For pure actin at physiological concentrations, this process is rather slow and occurs at velocities of the order of a few micrometers per hour. But as we shall see in the following, specialized actin-binding proteins allow the cell to increase this speed substantially, which makes actin forces exerted via polymerization-depolymerization mechanisms one of the key players in cellular motility.

Actin is the most abundant protein in a eukaryotic cell (several grams per liter), and has been highly conserved throughout evolution. It organizes into a variety of structures, namely linear bundles, two-dimensional networks or three-dimensional gels, and is mainly concentrated in a layer located just beneath the plasma membrane and called the *actin cortex*. Of primary importance for cell motility are the two-dimensional highly cross-linked networks that actin forms in lamellipodia, and the linear bundles that are found in filopodia and which protrude from the lamellipodia in a directed way. There, as we shall see in Sect. "Filament-Driven Motility", actin polymerization plays a crucial role in driving cell motility. Finally, one should mention the cortical rings that contract during the process of cytokinesis to cleave the two daughter cells, as well as the formation of *stress fibers*, which are force-producing structures that are attached to anchoring points, and which enable the cell to exert traction forces on the substrate on which the cell is crawling[10] (see Sect. "Cell Anchoring and Regulatory Pathways").

Microtubules are the stiffest of all polymers, with persistence lengths ranging from 100 µm up to 6 mm [213]. They are hollow cylinders with an outer diameter of ≃25 nm and are made of tubulin subunits arranged in 13 adjacent protofilaments. Tubulin is a heterodimer formed of α- and β-subunits, which can bind either

[8] The hydrolysis reaction of ATP (ATP $\rightleftharpoons$ ADP + P_i, where P_i designates inorganic phosphate) breaks a high-energy chemical bond – here a phosphoanhydride bond – to drive many chemical reactions in the cell.

[9] Animated movies of this process can be seen at http://www.uni-leipzig.de/~pwm/kas/actin/actin.html or http://cellix.imba.oeaw.ac.at/actin-polymerisation-drives-protrusion.

[10] Illustrations of these structures can be found at http://cellix.imba.oeaw.ac.at. See also [136,257].

guanosine triphosphate (GTP) or guanosine diphosphate (GDP)[11]. Microtubules share some important properties with actin filaments, in that they are polar, treadmill, and can exert forces [51]. They typically organize radially from a single microtubule-organizing center called the *centrosome*, and connect to the actin cortex with their plus ends towards the cell edge. In addition to giving the cell its structural rigidity and shape, they actively participate in regulating the actin cortex dynamics, focal-adhesion assembly and disassembly, and in some cell types participate in determining the cell polarity and its subsequent migrating direction (see Sect. "Cell Anchoring and Regulatory Pathways").

Intermediate filaments are the most flexible polymers of the cell cytoskeleton, with persistence lengths of the order of 0.3 to 1.0 μm. They range in diameter from 7 to 12 nm, in-between that of actin and microtubules. There are different classes of intermediate filaments such as *vimentin*, *desmin*, *keratin*, *lamin* and *neurofilaments*, and they constitute together a large and heterogeneous family, of which different cells possess different members. Unlike actin filaments and microtubules, they are not polar, do not treadmill, and are therefore thought to contribute essentially to the structural and elastic properties of the cell, but little to its dynamics and motility. One particular example of intermediate-filament structure is the *nuclear lamina*, located just beneath the inner nuclear membrane, and that is responsible for its structural integrity.

Molecular Motors

Molecular motors constitute the subset of proteins and macromolecular complexes that convert a given source of energy into mechanical work [5,103]. The energy they need is generally stored into either of two forms by the cell: high-energy chemical bonds, such as the phosphoanhydride bonds found in ATP and GTP, and asymmetric ion gradients across membranes. Known molecular motors can be classified into roughly five categories, namely (1) rotary motors, (2) linear-stepper motors, (3) assembly-disassembly motors, (4) extrusion nozzles, and (5) prestressed springs. A nice table of the major different cell-movements' categories with the different cellular structures and molecular motors they rely on, can be found in [67].

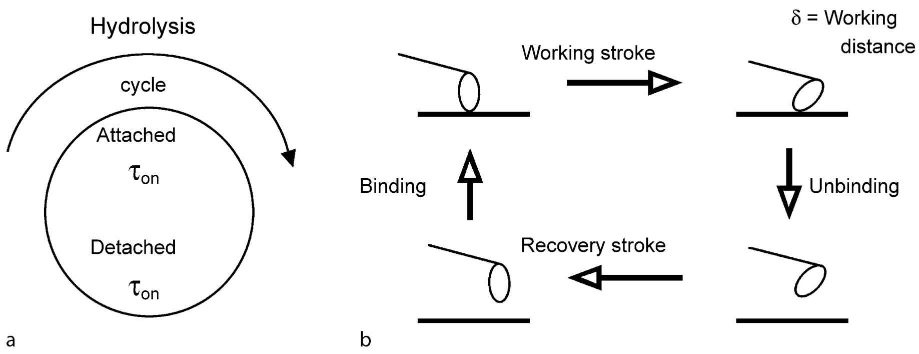
All known biological rotary motors use ion-gradient-based sources of energy, and most of them use electrochemical forces based on hydrogen-ion (or proton) gradients, also known as *proton-motive forces*. This is the case for example for the propulsion motor of bacteria that is responsible for their flagella to rotate [20,181], as well as for the surprising rotary motor F0F1-ATPase that is responsible for ATP synthase in mitochondria and bacteria [290]. This rotary machine usually converts the electrochemical energy stored in proton-concentration gradients, first into mechanical motion, and then back into chemical energy under the form of ATP. But the motor is also reversible, in that it can harness the chemical energy of ATP to produce or maintain the transmembrane electrochemical gradient of proton concentration. This reversibility is best seen in bacteria, when they switch from aerobic to anaerobic conditions [5].

Most of the motors used in amoeboid motility are linear-stepper motors [103,246,258]. We shall therefore focus on this class of motor proteins in the remaining of the present article. They walk on the linear tracks formed by the polymerized cytoskeletal filaments, and can be classified into two different categories, namely *processive* and *non-processive* molecular motors, sometimes designated as "porters" and "rowers" [159]. The processivity is linked to the *duty ratio*, the proportion of time that the molecule spends attached to the filament as compared to the whole motor cycle, namely one ATP-hydrolysis cycle [102] (see Fig. 8). Typically, porters are individual walkers that carry cargos across the cell, and therefore most often participate in intra-cellular traffic. Rowers however work in numbers, and are usually involved in generating contractile forces, like it is the case in skeletal-muscle fibers, stress fibers or contractile rings that form during cytokinesis [82,109,113,115]. Structurally, all these motor proteins can be divided into a motor domain, called the *head*, and a *tail* or *base*. The head is the site of conformational change of the protein during ATP-hydrolysis, and with which the motor attaches to the filament. The tail connects the motor to its cargo or to other motors. Processive motors are (homo-)dimers, such that as one head is attached to the filament, the other can move to a new binding site. In that case, the two tails of the associated monomers wind up together to hold to each other. Non-processive motors can also be found in dimeric forms, one of the two heads being then just unused.

Motor Families

Eukaryotic cytoskeletal motor proteins are divided into three superfamilies, namely *myosins*, *kinesins* and *dyneins*. The motor proteins known longest belong to the myosin superfamily [21], because of their high concentration in skeletal muscles. All myosin motors walk on actin filaments through a general four-step process: binding,

[11] Similarly to ATP, GTP is a stored source of energy for the cell that is consumed via a hydrolysis reaction, here GTP $\rightleftharpoons$ GDP + P_i.



Cytoskeleton and Cell Motility, Figure 8

a ATP-hydrolysis cycle, with the respective durations τ_{on} and τ_{off} of the attached and detached states of the motor. These durations define the duty ratio r as $r = \tau_{on}/(\tau_{on} + \tau_{off})$. b During the attached phase, the head of the motor makes a working stroke of working distance δ. The motor then unbinds from the filament, and makes a recovery stroke during the detached phase. By recovering its initial conformation while detached, the motor avoids stepping backwards and so progresses by a distance equal to the working stroke during each cycle. Source: reprinted from [102] with permission from Nature Publishing Group

power-stroke, unbinding, and recovery-stroke[12] [102] (see Fig. 8). Today, they are classified into 18 different classes, with possibly dozens of different members in each class, even in a single organism. The skeletal-muscle myosins belong to the Myosin II family; they have long tails that form dimeric α-helices and associate into the so-called "thick filaments" originally observed by H. E. Huxley and J. Hanson, while the "thin filaments" are F-actin polymers [93,110,111] (see Fig. 2 and 9). Most myosin molecules are plus-ended directed (to the exception of Myosin VI), and non-processive (to the exception of Myosin V, which is involve in vesicular transport). Their very diverse mechanical features, in terms of step sizes, duty ratios and stepping speeds, are very fine-tuned to their functions[13] [102].

Kinesin proteins share very similar structural features with myosins in their head domain and are therefore thought to have branched from a common ancestor with myosins, but diverge in their tail structures [273]. They walk on microtubules instead of actin filaments, are processive, and are involved mainly in intracellular transport like the transport of organelles along nerve axons[14]. The kinesin superfamily has been divided into 14 families, and a number of "orphans" that are so far ungrouped [189]. Most kinesin motors are plus-ended directed, like the conventional kinesin I that founded the family [274]. Members of the Kinesin-13 family are unconventional, in that they can processively induce microtubule depolymerization, a process that is essential to chromosome segregation during mitosis[15] [107] (see next Section).

Dynein proteins are less well-characterized. It is also unknown whether they share a common ancestor with myosins and kinesins, or whether they are the result of convergent evolution. Two major groups of dyneins exist: axonemal dyneins, which drive the bending of eukaryotic cilia and flagella by inducing the relative sliding of microtubules [226], and cytoplasmic dyneins, which are involved in organelle and vesicular transport, as well as cell division [133]. Most dyneins are minus-ended directed, and interestingly, some dyneins can be non-processive at high, but processive at low ATP concentrations.

Other Cytoskeleton-Associated Proteins

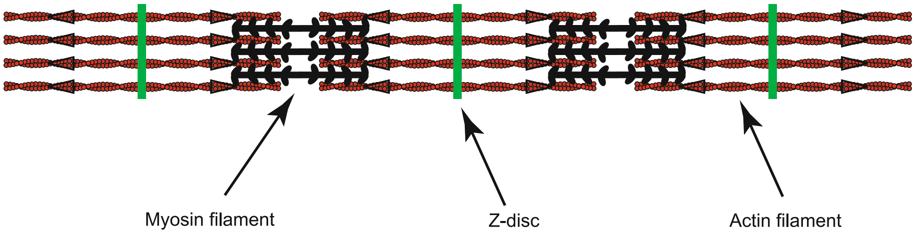
The coordination of the numerous different processes that happen during amoeboid motility rely on a tight regulation of the activity of the cell cytoskeleton, as well as its anchoring to the substrate. In particular, as we shall see below, the protrusion of the leading edge of the cell – the first step of amoeboid motility – relies on the formation of a highly-cross-linked and dynamic network of actin filaments. Its formation and dynamical regulation are carried out with the help of numerous accessory proteins [224,257]. Following Pollard's presentation [223,225] (see Fig. 10), we can focus on the main proteins that are involved in the formation, structure and dynamics of

[12] Animated movies of myosin skeletal fibers' detailed motion can be seen at http://www.scripps.edu/cb/milligan/research/movies/myosin.mov, or http://valelab.ucsf.edu.

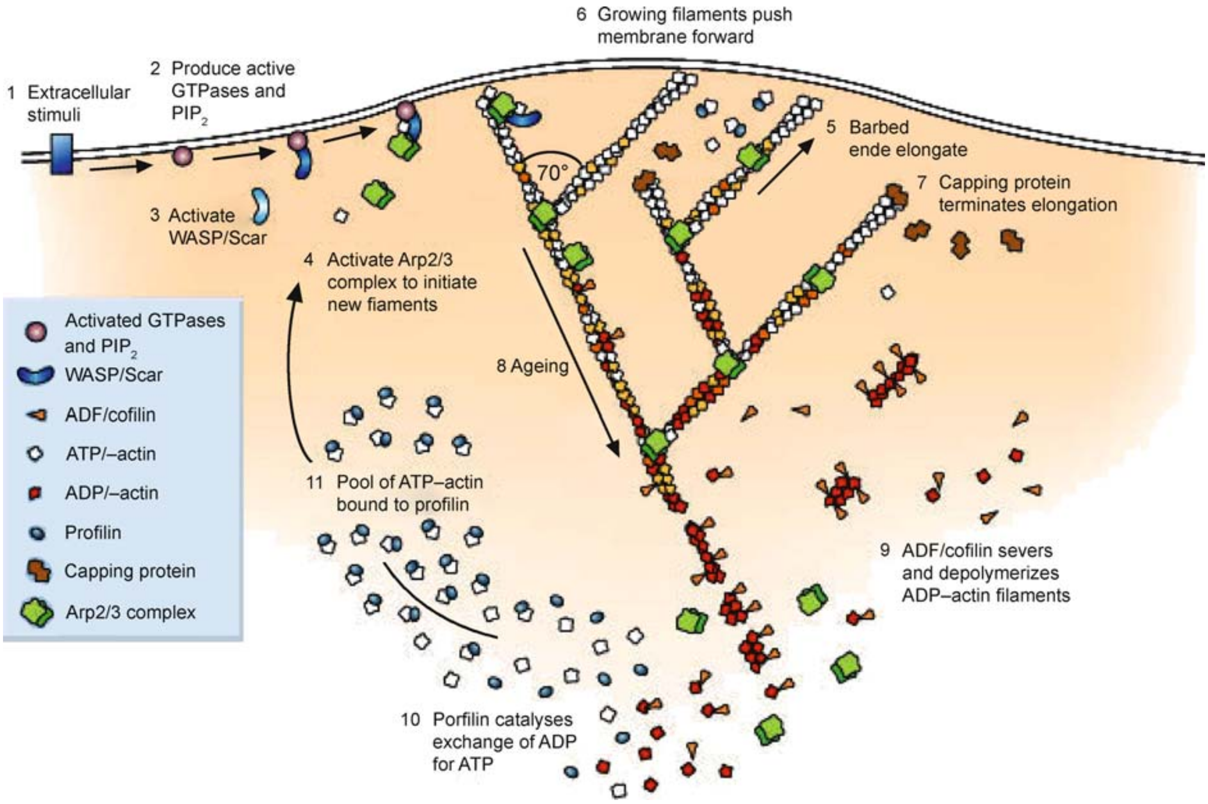
[13] Up-to-date information about myosin motors can be found at http://www.proweb.org.

[14] Animated movie of kinesin's detailed motion can be seen at http://www.scripps.edu/cb/milligan/research/movies/kinesin.mov or http://valelab.ucsf.edu.

[15] Up-to-date information about kinesin motors can be found at http://www.proweb.org.



Cytoskeleton and Cell Motility, Figure 9
Schematic representation of muscle myofibrils, the basic contractile fibers of skeletal muscles. Actin and myosin filaments are periodically arranged in a polarity-alternated fashion. Between two "Z discs" is found the elementary structure that is periodically repeated, the *sarcomere*, and where relative sliding of actin and myosin filaments leads to contraction. Source: courtesy of Karsten Kruse


1 Extracellular stimuli
2 Produce active GTPases and PIP$_2$
3 Activate WASP/Scar
4 Activate Arp2/3 complex to initiate new filaments
70°
6 Growing filaments push membrane forward
5 Barbed ende elongate
7 Capping protein terminates elongation
8 Ageing
9 ADF/cofilin severs and depolymerizes ADP–actin filaments
10 Porfilin catalyses exchange of ADP for ATP
11 Pool of ATP–actin bound to profilin
Activated GTPases and PIP$_2$
WASP/Scar
ADF/cofilin
ATP/–actin
ADP/–actin
Profilin
Capping protein
Arp2/3 complex

Cytoskeleton and Cell Motility, Figure 10
Dynamical organization of the actin network at the leading edge of a protruding lamellipodium. (1) External cues activate signalling pathways that lead to GTPases and PIP$_2$ activation (2). These then activate proteins of the WASP family (3), which in turn activate Arp2/3 complexes that initiate new filaments as branching from existing ones (4). Each new filament grows rapidly (5), fed by a high concentration of profilin-bound actin stored in the cytoplasm, and this pushes the plasma membrane forward (6). Capping proteins bind to the growing ends, terminating elongation (7). Actin-depolymerizing factor (ADF)/cofilin sever and depolymerize the ADP filaments, mainly in the "older regions" of the filaments (8, 9). Profilin re-enters the cycle at this point, promoting dissociation of ADP and binding of ATP to dissociated subunits (10). ATPactin binds to profilin, refilling the pool of subunits available for assembly (11). Source: reprinted from [223] with permission from Nature Publishing Group (image based on an original figure of [224])

the actin network. Nucleation of the network starts after biochemical signals have been integrated via G-protein-linked membrane receptors, namely small GTPases and PIP$_2$ pathways [174] (see Sect. "Cell Anchoring and Regulatory Pathways"). Then members of the Wiscott Aldrich syndrome protein (WASP) family that are anchored to the cell's plasma membrane (like Scar [175]), activate Arp2/3 (for actin-related proteins 2 and 3) complexes that are responsible for the nucleation and maintenance of branching points in the network[16]. Then, in order to promote growth of the actin gel, recycling of G-actin monomers as well as the creation of new F-actin plus ends are stimulated by mainly two types of proteins: (1) Actin-binding proteins – such as profilin – that bind to actin monomers, catalyze the exchange of ADP for ATP, and inhibits ATP hydrolysis, a process that is antagonized by monomer-sequestering proteins – like thymosineβ-4 – that stabilize ADP-bound G-actin. (2) Actin-depolymerizing factors (ADF) – such as cofilin (or ADF/cofilin) – that sever and depolymerize ADP-actin filaments, thereby increasing the pool of available G-actin monomers. The structure of the network is further controlled by capping proteins that can bind to F-actin plus ends to terminate their growth, and thus limit the increase of free-growing plus ends. Finally, cross-linked structures are formed with the help of actin cross-linkers like filamin, and actin-bundling proteins like fascin, fimbrin and α-actinin. α-actinin and filamin are most present in lamellipodium structures, as fimbrin and fascin and most observed in filopodia [257]. Finally, the same as well as other actin-binding proteins (like espin, fascin, fimbrin and villin) exist in other structures where actin-bundles are formed, like bristles, microvilli and stereocilia [234]. In skeletal muscles, tropomyosin strengthens the actin filaments and prevents myosin motors from binding to actin when muscles need to be at rest[17]. For a relatively recent review on the actin-binding proteins, see [53].

Microtubule-associated proteins (MAPs) have been classified into two types, and participate to microtubules' stability and organization. MAPs of Type I are large filamentous proteins that comprise a microtubule-binding domain and a projection domain, thereby controlling the spacing of microtubules. MAPs of Type II have similar structures and cross-link microtubules to membranes, intermediate filaments or other microtubules. In addition, both types of MAPs promote microtubule assembly and stability, and compete with motor proteins for binding sites, such that they participate in microtubule-transport regulation. Other MAPs that do not belong to these classes are denoted XMAPs, as they have been originally identified in the Xenopus-frog eggs. Among these are the plus-end-binding proteins (or +TIPs) that bind to the microtubule growing ends and participate in their stability, and the highly-conserved stathmin or oncoprotein 18 which, instead, destabilizes microtubules [10,249]. The best understood microtubule end-binding proteins are the MCAKs (for mitotic centromere-associated kinesins), also known as Kin I kinesins, which are unusual kinesins in that, instead of moving along the surface of microtubules like other kinesin proteins do, they bind to microtubules' ends and trigger depolymerization in a processive way [47]. In particular, they depolymerize microtubules during mitosis to drive chromosome segregation [180]. For a review, see [104].

Cell Anchoring and Regulatory Pathways

The two first steps of cell crawling in the Abercrombie classification consist in the protrusion of the leading edge and its adhesion to the substrate [1,156]. Although they were thought to be largely independent processes, evidences are accumulating that adhesion and protrusion are highly interrelated [9,24,90,236]. Protrusion results primarily from actin polymerization at the leading edge of the migrating cell (see Sect. "Filament-Driven Motility"), and is regulated by the small GTPases Rho, Rac and Cdc42 [25,235]. As Rho is known to activate actomyosin contractility, Rac and Cdc42 induce actin polymerization and the formation of actin-filled protrusions such as lamellipodia and filopodia [91]. Through these pathways, the cell can respond to the chemical composition of its environment, an example of chemotaxis: as a function of the gradients of chemoattractants or chemorepellants in its environment, the cell regulates its sites of fastest actin polymerization in order to move towards or away from the source[18] [215]. Of primary importance for the accurate spatial regulation of these processes is the role of microtubules, which in some cell types play a crucial role in determining cell polarity and directional migration [136,204,236,255,256]. Microtubules have been proposed to activate Rac and Rho, the latest via the release of the GDP-GTP exchange factor GEF-H1 during microtubule depolymerization [143,283].

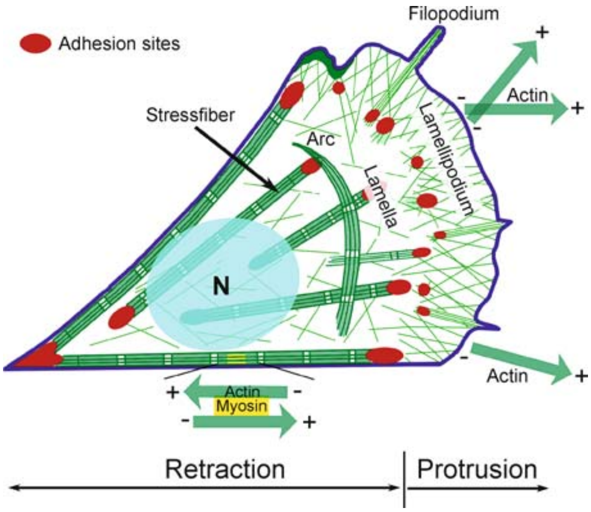
Adhesion occurs via the formation of adhesion sites, which rely primarily on molecules such as *integrins*, also involved in regulating the cell behavior via different signal-transduction pathways [24,78,190]. An important aspect of that process is that it allows the cell to "feel" the mechanical properties of its environment. This has been shown to be important for the migration of fibroblasts, in that they seem to migrate preferentially towards regions of

[16] In addition to Arp2 and Arp3, which are members of the Actin related proteins (Arp) family in that they have sequences and structures that are similar to actin, the Arp2/3 complex contains five other smaller proteins.

[17] Up-to-date informations can be found at http://www.bms.ed.ac.uk/research/others/smaciver/Cyto-Topics/actinpage.htm.

[18] See the movies associated with ref [215], as well as the one of a neutrophil cell that chases a bacterium at http://www.biochemweb.org/fenteany/research/cell_migration/movement_movies.html.

stiff substrates, a process referred to as *durotaxis* [169]. In addition, the mechanical properties of the cell's environment have been proposed to be relevant for tissue-growth directionality, as well as cell differentiation [56,243]. Adhesion sites can be roughly divided into two broad categories, namely *focal complexes*, which locate beneath microspikes or filopodia, and *focal adhesions*, which locate at the termini of stress fibers and serve in long-term anchorage [136] (See Fig. 11). Interestingly, these adhesion sites are also regulated by small GTPases of the Rho, Rac and Cdc42 families: focal complexes are signaled via Rac1 and Cdc42, and can either turnover on a minute time-scale or differentiate into long-lived focal adhesions via the intervention of RhoA [136]. A schematic representation of the integrated role of the small GTPases in regulating cell migration can be seen in Fig. 12. Such regulations are crucial for cell migration to occur optimally. Indeed, whereas adhesion sites are necessary at the leading edge of the cell to provide anchoring points on which the cell can exert traction forces, these need to be released at the rear for the cell to move forward. This results in a biphasic response of the cell-migration speed as a function of adhesive-ligand concentrations, in that too-low or too-high ligand concentrations prevent either the traction forces to be exerted, or the rear to be released [131,156]. How these regulatory pathways lead to a spatio-temporal feedback mechanism between actomyosin regulation and the focal-adhesion system is still under investigation [24,90].



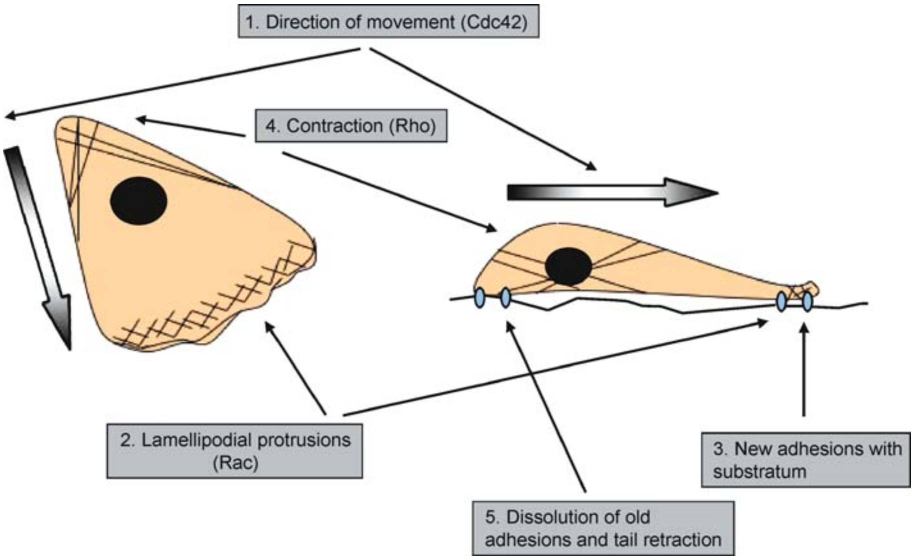
Cytoskeleton and Cell Motility, Figure 11
Schematic representation of the actin cytoskeleton in a polarised fibroblast. The different organisational forms of actin filaments and their relations to adhesion sites to the substrate are depicted: diagonal actin filament meshwork in the lamellipodium, with associated radial bundles that sometimes protrude into filopodia; contractile bundles of actin (stress fibers) in the cell body and at the cell edge; and a loose actin network throughout the cell. Arc-shaped bundles are sometimes observed that move inwards under the dorsal cell surface (arc). The diagram shows an idealized cell: in reality, actin arrays are interconnected in various combinations and geometries. Adhesion sites are indicated in red. The flat region behind the lamellipodium and in front of the nucleus (N) is termed the lamella. At the cell front, in lamellipodia and filopodia, actin filaments are all polarized in one direction, with their fast-growing ends directed forward for producing pushing forces and inducing protrusion; in the cell body, actin filaments form bipolar assemblies with myosin proteins (stress fibers) for retraction. Source: courtesy of Vic Small; modified from [136] with permission from Elsevier Limited

The Prokaryotic Cytoskeleton

This section is completely independent of the rest of the article. Readers not interested in the biochemical composition of the prokaryotic cytoskeleton might want to skip this section.

As cytoskeletal proteins' structures are highly conserved throughout the three domains of life (archaea, bacteria and eukarya), prokaryotic cytoskeletal proteins differ strongly in their sequences from their eukaryotic counterparts. For this reason, and the fact that prokaryotes have a relatively simple organization as compared to eukaryotic cells, it was long thought that they were lacking a cytoskeleton. It is only in the 1990s that prokaryotic homologs of tubulin, actin and intermediate filaments started to be discovered. The first bacterial cytoskeletal proteins to be brought to knowledge was the protein FtsZ, whose relation to tubulin was discovered independently by three groups in 1992 [43,200,232]. Later, it was found that FtsZ could assemble into protofilaments that can be either straight or curved as a function on the state of the nucleotides, similarly to microtubules [58,104], and that its structure at the level of protein folding was nearly identical to that of tubulin [171,206]. The second prokaryotic cytoskeletal proteins to be discovered were MreB and ParM also in 1992, and were shown to be distant relative of the actin superfamily by sophisticated sequence-alignment techniques [26]. Later, it is only in 2001 that MreB was proven to be capable of self-assembly into cytoskeletal filaments that resemble much closely F-actin structures [126,276]. Finally, an homolog of intermediate filaments has been found recently in the bacterium *Caulobacter crescentus*, but only in this particular species so far [16]. Since it is responsible for giving the bacterium its crescent shape, it was given the name of *crescentin*.

Despite their sequencial differences with their eukaryotic counterparts, prokaryotic cytoskeletal proteins share with them strong homologies in their structural as well as


1. Direction of movement (Cdc42)
4. Contraction (Rho)
2. Lamellipodial protrusions (Rac)
3. New adhesions with substratum
5. Dissolution of old adhesions and tail retraction

Cytoskeleton and Cell Motility, Figure 12

Schematic representation of the integrated roles of Rho, Rac and Cdc42 proteins in regulating cell migration. By inducing actin-filament assembly, filopodia and focal-complexes formations, Cdc42 regulates the direction of migration (1). Rac induces actin polymerization at the cell periphery (2) and promotes lamellipodia protrusion. It also induces the formation of focal complexes at the leading edge (3). Rho plays a role in regulating longer-lived structures, namely activating actomyosin contraction in the stress fibers located in the cell body and at the rear (4), as well as promoting the assembly of focal-adhesion complexes. Source: courtesy of Alan Hall; reprinted from [229] with permission from Elsevier Limited

functional properties [8,188]. They are classified into four groups [253]: (i) Actin homologs are constituted by MreB and MreB homologs, ParM, and MamK. MreBs play an important role in a number of cellular functions, such as regulation of cell shape, chromosome segregation, establishment of cell polarity and organization of membranous organelles. ParM proteins are involved essentially in plasmid partitioning, and MamK is involved in the subcellular organization of membrane-bounded organelles. Similarly to actin, MreB and ParM protein families present polymerization-depolymerization dynamics that are driven by ATP hydrolysis. Less is known about MamK. (ii) Tubulin homologs contain FtsZ and the BtubA/B proteins, which constitute two other families of GTPases as compared to tubulin. As FtsZ is crucially involved in cytokinesis via its ability to form contractile rings and spiral structures, the role of BtubA/B proteins, which are much less widespread in the bacterial kingdom, has less been characterized so far. (iii) The intermediate filaments' homolog crescentin has only been found in *Caulobacter crescentus* and, as for its eukaryotic counterparts, is mainly involved in cell shape and structural integrity. (iv) Finally, the large MinD/ParA superfamily is made of prokaryotic cytoskeletal proteins that have no counterparts in eukaryotes. They however have the ability to organize into polymeric filaments, and present ATPase activity. Proteins of the MinD group are involved in placement of the bacterial and plasmid division sites, whereas proteins of the ParA subgroup are primarily involved in DNA partitioning.

Interestingly, cytoskeletal proteins seem to have been strongly conserved throughout evolution in each of the three separate domains of life, but differ quite substantially across domains [57]. Bacterial FtsZs proteins are 40–50% identical in sequence across species, and share even the same amount of similarities with their archaeal counterparts. Bacterial MreBs are generally 40% conserved. Among eukaryotes, the conservation is even stronger: it reaches 75–85% for tubulin and 88% for actin, one of the most conserved protein in the eukaryotic domain. In the case of archaea, MreB and actin homologs have not yet been identified for sure [57].

Filament-Driven Motility

Many kind of movements in eukaryotic cells are driven by polymerization-depolymerization mechanisms of cytoskeleton filaments, for which motor proteins per se are not required. Instead, the chemical energy stored in high-energy hydrogen bounds (under the form of ATP or GTP) is converted into movement via treadmilling mechanisms [264]. Two types of filaments have this ability, namely microtubules and actin filaments.

Microtubule Growth and Catastrophes

Microtubules are the stiffest of all cytoskeletal filaments, which confers them the ability to organize and stabilize both the cell structure and its transport network for internal communication and distribution. Depending on the cell need, they constantly reorganize or exert forces on the cell membrane or other organelles they transport. This is the case for example during the organization of the mitotic spindle, the structure formed prior to chromosomal segregation during mitosis [248]. There, chromosomes gather in a plane halfway from two microtubule-organizing centers – the centrosomes – that are located at each pole of the future dividing cell, and from which the microtubules tear the chromosome pairs apart [6,118] (see Fig. 6, especially panels C to F). For this mechanism to happen, microtubules constantly exert pulling and pushing forces both on the chromosomes and the cell membrane, a mechanism that allows for correct positioning of the site of cell-division [288]. Coupled with kinesin-motor activity, correct positioning of the centrosomes relies crucially on the ability of microtubules to grow and shrink spontaneously, a dynamics that provides feedback to centrosome positioning and leads to oscillations orthogonal to the cell spindle axis [85,86]. Such a mechanism is also responsible for the correct positioning of the nucleus in the fission yeast *Schizosaccharomyces pompe* [268].

Understanding microtubules' polymerization-depolymerization dynamics started with the first observation that they display phases of relatively slow growth, alternated with phases of rapid shrinkage [191]. Changes from one type of behavior to the other are referred to as *catastrophes* for the conversion from growing to shrinking, and *rescues* for the opposite transition. Observations of this behavior were further made in culture cells [244] and cellular extracts [19], which confirmed the existence of such dynamic instabilities in vivo. During mitotic-spindle formation, it has later been shown that the specialized structures that connect the microtubules to the chromosomes, the *kinetochores*, can "capture" and stabilize growing microtubules, preventing them from undergoing catastrophes [95]. For a review, see [134].

Further characterization of microtubules' biomechanical properties came from experimental studies of the forces produced by their polymerization-based growth. Analyzing force-induced microtubule buckling [50], microtubule forces were characterized as being potentially as high as those produced by motor proteins – typically a few pico-Newtons [123] – and to be able to deform membranes [74] or center asters in mirofabricated chambers [62,101], a mechanism that imitates nucleus positioning in fission yeast. For reviews, see [51,104].

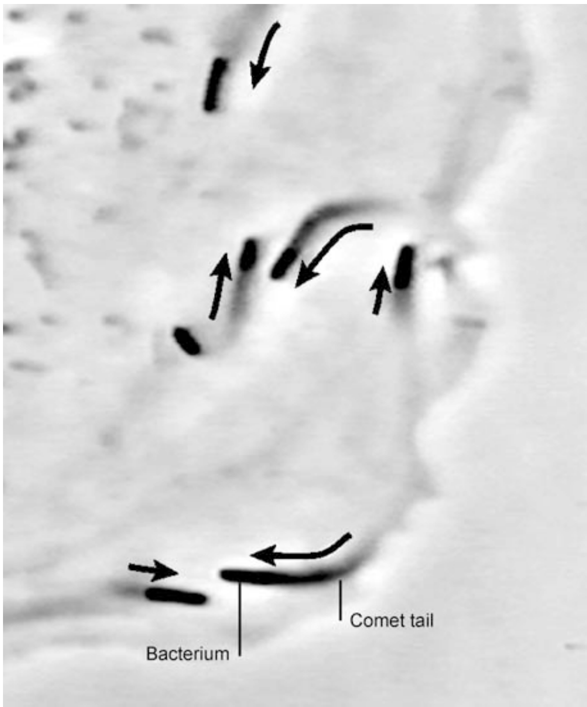
Actin Gels

As earlier stated, the first step of amoeboid motility in the original Abercrombie classification occurs via protrusion of the leading edge of the cell. This mechanism relies mainly on the polymerization dynamics of actin filaments [194,214,225]. Actin polymerization is known to play a primary role at the plasma membrane, where it is nucleated by proteins of the WASP family via Arp2/3 complexes (see Sect. "Other Cytoskeleton-Associated Proteins"). It has also been proposed to be responsible for driving endocytosis and the movement of endosomes, both in cultured cells and yeast [132,186].

Our understanding of eukaryotic actin-based motility has grandly benefitted from the motility mechanism of the bacterium *Listeria monocytogenes*. This pathogene moves at velocities of the order of several micro-meters per minute by nucleating the formation of an actin "comet-tail" that, while polymerizing thanks the host's cytoskeletal machinery, pushes the pathogene forward [265] (see Fig. 13). This particular motility mechanism, studied in in vitro assays, has allowed for the identification of the minimal set of proteins needed to actin-based motility, as well as the role of several of the main actin-related proteins [170,284]. It has also been used as a probe for the cell cytoskeleton network structures and visco-elastic properties in a position-dependent manner [153], and has shed light into the basic elementary principles of actin-based motility [71,87,194,214].

Except for very recent reports [68], nearly no force measurement has yet been done on single actin filaments. Due to their smaller bending rigidity, the corresponding stall force is expected to be orders of magnitude smaller than that of microtubules because of buckling phenomena. Instead, large forces can only be obtained when highly cross-linked actin filaments work as a whole and form a relatively rigid network, as it is the case in filopodia protrusion. Forces generated during actin-based propulsion have been measured on polymerizing actin gels, in particular using in vitro assays based on artificial biomimetic systems. Forces in the range of a few nano-Newtons have been found for gel comets originating from 2-µm-size polystyrene beads [182].

Other bio-mechanical characterizations of actin network's properties concern the study of its gel-like viscoelastic properties. In particular, transitions between a solid-like elastic material and a solution-like viscous material have been observed [73]. These could rely in part on



Cytoskeleton and Cell Motility, Figure 13
***Listeria* bacteria moving in a cell via actin-based motility. This is a phase-contrast micrograph, a single frame from a video sequence (the whole movie can be seen on the following webpage: http://cmgm.stanford.edu/theriot/movies.htm). The kidney epithelial cell was infected about five hours before the acquisition of this video sequence. All of the bacteria in this cell are clonal descendants of a single individual. A bacterium and its associated comet tail are labeled. Bacteria are moving in the direction of the arrows. Source: reprinted from [36] with permission from Nature Publishing Group**

the biochemical-dependent mechanical properties of the actin filaments themselves [119], on the generic properties of such semiflexible-filament networks [75], or on the activity of motor proteins that help disentangling the network and thereby lead to its fluidization [106]. It has also been observed that cross-linked actin networks display an increase of their elastic modulus as a function of the stress applied, a nonlinear behavior known as *stress-stiffening* [63,259]. This might explain partly the interestingly rich properties of cellular rheology [99,135], which have been partly reproduced in in vitro measurements [77]. Among these, dynamical scaling of the stress stiffening (see e. g. [76]) has been proposed to be the signature of underlying self-similar mechanical properties of the cell cytoskeleton [18,61]. Finally, the intermediate filaments as well as the biochemical environment or preparation of the actin network have been proposed to play an important role in modulating its rheological properties [75,122]. This might contribute to the observed local changes in the elasticity of the cell as it moves, a crucial aspect for driving its motility [9,73].

Modeling Polymerization Forces

With general thermodynamic considerations, growth velocities of polymerizing filaments can be understood as follows: if k_{on} and k_{off} are the association and dissociation constants for monomers at the polymer tip, and δ is the distance a filament grows under addition of a single monomer, a typical growth velocity of the polymerizing filament is given by: $v = \delta[k_{\text{on}} - k_{\text{off}}]$. When experiencing a force f opposing polymerization, like the cellular membrane resistance at the leading edge of the advancing cell, filament-growth velocity becomes:

$$v(f) = \delta\left[k_{\text{on}} \exp\left(-q\frac{fa_1}{k_B T}\right) - k_{\text{off}} \exp\left((1-q)\frac{fa_1}{k_B T}\right)\right]. \quad (1)$$

In this expression, $k_B T$ represents thermal energy, fa_1 is the most probable work needed to add a monomer in the presence of the force f, and q is a parameter describing how much the force f influences the on-rate as compared to the off-rate. Under these assumptions, the maximal force a given filament can produce via polymerization, or *stall force*, is expressed as $f_s = (k_B T/a_1) \cdot \ln(k_{\text{on}}/k_{\text{off}})$. Even though good overall agreement with experimental data was obtained for individual microtubules while choosing $13a_1 = a$ and $q = 1$ (with a being the size of a tubulin monomer) [50], the so-derived stall force was too large as compared with experimental measurements. This led to revising the dynamics of the microtubule-polymerizing end, proposing $a_1 \simeq a$ and $q \simeq 0.22$ as better fitting parameters, pointing to a rich dynamics of microtubule polymerization [140].

To understand the origin of polymerization forces, the standard microscopic model relies on the ratchet mechanism, a rectified Brownian motion originally introduced in this context by Peskin et al. [219] to explain filopodia protrusion, *Listeria* propulsion as well as protein translocation. Filopodia protrusion in particular is thought to rely essentially on actin-polymerization forces: when reaching the cell membrane, growing F-actin filaments feel a force opposing their growth, and therefore exert a force on the membrane. Because of thermal fluctuations and membrane's as well as actin-filaments' finite bending rigidities, some space is constantly opened between the growing filament and the membrane. From time to time, an additional monomer can thereby be added to the growing filament, which pushes the membrane forward[19]. In the simplest case of a single stiff protofilament, the distribution P of distances x between the filament and the membrane is given by the following Fokker–Planck equation:

$$\begin{aligned}
\partial_t P(x) &= D\partial_x^2 P(x) + f\frac{D}{k_\mathrm{B}T}\partial_x P(x) \\
&\quad + k_\mathrm{on}P(x+\delta) - k_\mathrm{off}P(x) \quad \text{for } x < \delta \\
\partial_t P(x) &= D\partial_x^2 P(x) + f\frac{D}{k_\mathrm{B}T}\partial_x P(x) \\
&\quad + k_\mathrm{on}[P(x+\delta) - P(x)] \\
&\quad - k_\mathrm{off}[P(x) - P(x-\delta)] \qquad \text{for } x > \delta\,,
\end{aligned} \tag{2}$$

where notations are similar to the ones used in Eq. (1), and to which must be added the effective diffusion coefficient D for the distances x between the filament and the membrane. The time-dependence of P is implicit. Using vanishing-current conditions at the leading edge $x = 0$, the stall force can be obtained and is given by an expression analog to that of microtubules models, with $a_1 = \delta$ being the size of a G-actin monomer. Including bending fluctuations of the growing filament, this led to the Elastic Brownian Ratchet Model [195], a generalization of which is the Tethered Elastic Brownian Ratchet Model [197] that considers that some filaments are attached to the membrane via protein complexes (as it has been observed in the *Listeria*-propulsion mechanism for example) and therefore do not exert polymerization forces. When typical parameter values are plugged into these models, single actin-filament force generation is estimated to be of the order of 5–7 pN [198]. Taken into account that at the leading edge several hundreds of actin filaments per micron work together to drive the cell forward, the resulting force is of the order of nanonewtons per micron [197], a force large enough to tackle the membrane load and resistance. However, it has since then been claimed that motor proteins, called *end-tracking motors*, should be required to explain observed forces in the case of *Listeria* propulsion for example [49]. This work has been reviewed in [194].

Lateral interactions between filaments in an actin network have been investigated via models that take into account the branching structure of the network [37,38]. In particular, Autocatalytic Models assume that new branches are generated from existing ones, which leads to a growth velocity that is independent of the load [38]. To investigate the consequences of these models, two approaches have been followed, namely stochastic simulations of the growing actin network, tracking each filament position and orientation [37], and deterministic rate equations that include growth, capping and branching rates, and which led to a comparison between ratchet and autocalytic models [38]. Experimental tests of the two models have been performed in in-vitro systems, using *Listeria* propulsion as well biomimetic systems [271] (see next paragraph). While some *Listeria* studies favored the Tethered Elastic Brownian Ratchet Model [184], some studies using biomimetic systems favored the Autocatalytic Model [285], and several others neither of them. A possible explanation for these apparently contradictory results may be that different experimental studies led to analyzing different regimes of the force-velocity curve.

A Model System for Studying Actin-Based Motility: The Bacterium *Listeria monocytogenes*

As earlier stated, our understanding of eukaryotic actin-based motility has grandly benefited from the motility mechanism of the bacterium *Listeria monocytogenes* (see Sect. "Actin Gels"). While velocities of *Listeria* bacteria in a homogeneous environment are typically constant, some mutants progress in a saltatory manner [155]. This observation has been later reproduced in in vitro motility assays using latex beads coated with the bacterium transmembrane protein ActA (that further recruits Arp2/3) [35], or directly with VCA proteins, a sub-domain of WASP that is responsible for actin-branching and polymerization nucleation [22]. Such biomimetic systems have allowed for the direct measurement of the characteristic polymerization force that is produced by an actin gel [182], and for the overall study of actin-based motility mechanisms in simple and well-controlled conditions [194,222,271].

Theoretical understanding of such actin-based propulsion mechanisms has come from two different angles, namely molecular and mesoscopic, continuum models. We have already reviewed the molecular models that rely on brownian-ratchet mechanisms. They, in particular, have led to force-velocity curves that are consistent with some observations of *Listeria* motion [184,197]. Continuum models describe the actin network as a compressible elastic gel with an elastic modulus of about 5000 Pa [79,182,207] (see also Sect. "Macroscopic Phenomenological Approaches: The Active Gels"). When growing over a curved surface like the bacterium *Listeria* or a coated bead, the gel deforms as it grows by monomer additions on the particle surface, which in turn generates a stress that pushes the particle forward [79,207]. Monomer transport to the inner surface of the growing gel is purely diffusive, with a diffusion constant that has been estimated to be of the order of 2 μm^2/s for actin monomers

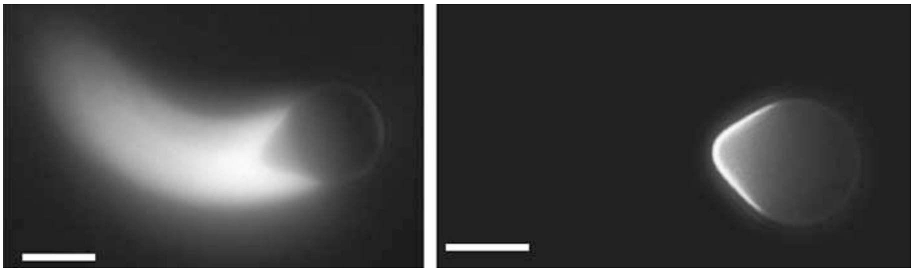
[19] For an animated illustration, see http://www.jhu.edu/cmml/movies/anim/eBRatchet2.swf.

in an ActA-produced gel [221]. When originally initiated on a spherical object like a rigid bead, the growth of the gel layer starts isotropic, but ruptures into a comet-type growth because of mechanical instability. The instability relies on a positive feedback that involves creation of a tensile stress as the gel grows because of geometrical effects, and enhancement of the depolymerization rate or rupture of the gel in regions of enhanced tensile stress [252]. The instability occurs less rapidly with increasing bead size, which explains why movement is more often observed with small beads. This mechanism of tensile-stress accumulation and rupture can also explain the saltatory motion observed with *Listeria* mutants and coated beads in some conditions [22]: rapid phases of motion are due to the sudden rupture of the gel that pushes the particle forward, as slow phases of motion correspond to progressive build-up of lateral tensile stress. Depending on the size of the bead as well as the concentration of proteins at its surface, this dynamic instability can be present or not, which explains the observation of both continuous and saltatory regimes with coated beads as well as *Listeria* bacteria [22,23].

To further explore the properties of the actin gel and the *Listeria*-propulsion mechanism, experiments with soft objects like liposomes [80,270], endosomes [262] and oil droplets [28] have been performed. They show that the actin gel squeezes the object, compressing its sides and pulling its rear, an effect that gives it a pear-like shape (see Fig. 14). Analysis of the contour of the deformed objects provides informations on the distribution of the normal stress on the surface of the object. This could in principle allow for the derivation of the total force exerted on the load in the case for example of oil droplets, where interfacial tension is measurable and normal stress can be deduced from Laplace's law [28]. But in fact, assuming a constant surface tension, the integration of the normal stress over the surface of the droplet gives a zero net value of the force that is independent of the droplet shape, the latest being regulated by the variation of the polymerization velocity with normal stress. Instead, the distribution of actin-polymerization promoters on the surface of the droplet follows the gel elastic deformations, which in turn creates pressure variations inside the droplet, and thereby surface-tension gradients along its surface (see Fig. 14); and these are at the origin of the final non-zero net pushing force [28]. Finally, direct observation of the actin comet during its growth on coated beads has shown that the actin gel constantly undergoes deformations that depend on the protein composition of the motility medium they are placed in [212]. As a function of bead size and the concentration of cross-linkers or regulatory proteins, the bead velocity can be limited either by diffusion of the monomers to the coated surface, by polymerization velocity at the surface of the bead, or by the elastic stress built up in the gel. These findings, supported by experimental results, buttress the idea that actin-based movement is governed by the mechanical properties of the actin network, themselves tightly regulated by the proteins that are involved in actin dynamics and assembly (see Sect. "Other Cytoskeleton-Associated Proteins").

Another Example of Filament-Driven Amoeboid Motility: The Nematode Sperm Cell

Even though, as described above, protrusion of the leading edge in amoeboid motility is most commonly actin-driven, other cells, the nematode sperm cells, use another cytoskeletal protein to drive their motility: the Major Sperm Protein (MSP) [240]. Nematodes constitute one of the most common phyla of all animal kingdom, with over 80.000 different described species, and their sperm is



Cytoskeleton and Cell Motility, Figure 14
Actin-based propulsion with liquid drops. Oil drops are covered with VCA, placed in cell extracts that are supplemented with actin, and observed by fluorescence microscopy. *Left panel*: Note the bright actin comet and the pear-like shape of the droplet due to squeezing forces exerted by the actin gel. Scale bar is 4 µm. *Right panel*: VCA is labeled with fluorescin isothiocyanate (FITC). Note the inhomogeneous distribution of the actin-polymerization promoter on the surface of the droplet. Scale bar is 3 µm. Source: reprinted from [28] with permission from The American Physical Society

thought to be the only eukaryote cell type that do not possess the globular protein G-actin. These cells offer an ideal to study cell crawling since, dislike actin, MSP is a simpler, more specialized protein that do not possess as many regulatory or associated proteins, and in particular is not known to bind any molecular motor (at the exclusion of end-tracking proteins). MSPs being also apolar [31], the nematode motility constitutes one of the simplest of all cytoskeleton-driven motility mechanisms known to operate in vivo [240,286].

Similarly to what has been done with biomimetic systems to study actin-based propulsion, motility assays using vesicles derived from the leading-edge of nematode sperm cells of *Ascaris* species, have shed light into the mechanisms at play [120]. In the presence of ATP, growth of MSP fibers are capable of pushing the vesicle forward, their polymerization being driven by specialized proteins located within the vesicle membrane, a mechanism that ressembles very much the *Listeria*-propulsion mechanism. But, contrary to ATP-driven actin treadmilling, MSPs assemble into apolar filaments and lack a nucleotide binding site for ATP hydrolysis. To power membrane protrusion, it has recently been proposed that motor end-tracking proteins processively polymerize MSP filaments, while keeping the elongating filaments' ends in contact with membrane-associated proteins [48]. For cell progression however, a second force is required, namely a traction force that pulls the cell body forward once the advancing lamellipodium has been anchored to the substrate on which the cell is crawling. In actin-based amoeboid motility, this process is motor-driven, but nematode sperm cells use instead the sensitivity of their MSP to pH, whose decrease provokes reorganization, depolymerization and in fine contraction of the network [121,187,286]. Polarity in the cell is maintained by an influx of protons close to the cell body, which creates a pH gradient in the lamellipodium and powers this process [137].

To quantitatively understand the mechanism underlying this motility, both microscopic and phenomenological models have been proposed. In the proposed microscopic models, mechanisms underlying the traction-force generation by solely cytoskeletal disassembly can be qualitatively understood as follows [27,286]: because of pH gradient, MSP filaments tend to bundle at the front, and split apart and disassemble at the rear [187]. A bundle with N filaments being much stiffer than N isolated individuals (with an effective persistence length of $N^2 L_p$ as each individual has a persistence length L_p), it pushes the cell membrane at the leading edge where filaments are bundled, while splitting filaments exert contractile forces at the rear. Indeed, because of entropic effects, filaments tend to retract once split apart. Finally, even further decrease in pH creates weakening of the attachments and dissociation of the filaments for monomeric MSPs to be recycled at the front [286] (see also [196]). In the proposed phenomenological approach however [124], the sensitivity to pH is described as influencing the equilibrium swelling properties of the gel only. As the gel treadmills towards the rear end where acidic conditions are found, it tends to contract by an isotropic multiplicative factor Λ that is position-dependent. General elasticity theory of continuous media allows to express the strain tensor as:

$$u_{\alpha\beta} = \tfrac{1}{2}(1 - \Lambda^2)\delta_{\alpha\beta} + \tfrac{1}{2}(\partial_\alpha u_\beta + \partial_\beta u_\alpha)\,, \qquad (3)$$

where u_α are the components of the displacement vector (with $\alpha = x, z$). Assuming linear elasticity theory[20], the stress tensor is then obtained as $\sigma_{\alpha\beta} = \lambda u_{\gamma\gamma}\delta_{\alpha\beta} + 2\mu u_{\alpha\beta}$, where λ and μ are the Lamé coefficients, which further leads to a position-dependent tensile stress as was introduced phenomenologically in [27]. While traveling through the lamellipodium, tangential stress builds up, which leads to rupture of the adhesion points once a critical force has been passed, and eventually drags the cell body forward. Therefore, within this framework, only one parameter is directly controlled by the pH – namely Λ – and the pH in particular does not need to influence directly adhesion strength.

Motor-Driven Motility

Generic Considerations

Despite the major role played by polymerization forces in cellular motility, and in particular as we have previously seen in amoeboid motility, a vast amount of diverse motile processes in eukaryotic cells is driven by motor proteins (see Sect. "Molecular Motors"). Theoretical studies of molecular motors started with the cross-bridge model published independently by A. F. Huxley and H. E. Huxley to explain the relative sliding of myosin filaments with respect to actin filaments in cross-striated muscle fibers [108,112]. This approach was later formalized by Hill [97], who introduced the notion of different "states" of a motor protein, each of these corresponding to a thermodynamic-equilibrium state. Interpretation of these different states was given in terms of different conformations of the motor protein and its interaction with the filament, or in terms of the state of the hydrolysis reaction of ATP, or both [54,159]. Justification for considering different thermodynamic-equilibrium states relied on

[20] For an introduction to the elasticity of continuous media, see, e. g., [154].

the observation that for the transient response of muscles, the fastest response was known to be in the range of milliseconds, as thermal equilibrium on molecular characteristic length scales of 10 nm occurs after at most a few hundreds of nanoseconds. In this class of models, progression of the motor along the filament relies on asymmetric transition rates of the particle between the different states, for which asymmetry of the filament and energy consumption by the motor is required. Typically, after one cycle of conformational states, the motor protein has progressed by one or several allowed binding sites on the periodic lattice represented by the cytoskeletal filament. In between, up to five or six different states could be involved [97,172]. Experimental confirmation came later with the direct observation of walking steps displayed by advancing molecular motors [258]. Such observations were first obtained studying kinesin motors in in-vitro motility assays, and later with myosin motors displacing a filament that was attached to two glass beads placed in laser-trap potentials [32]. For a review, see [246].

Another class of models relies on the generalization of Feynman's famous "thermal ratchet", in which the presence of different heat baths (namely thermal baths at different temperatures) can rectify the brownian motion of a given particle and lead to its directed motion [64]. For motor proteins, as we have already discussed, temperature inhomogoneities in the system cannot hold long enough to ground the mechanism. Instead, various different isothermal rectifying models have been discussed to describe the underlying mechanisms of different biophysical processes [13,177]. Among these, one can mention the translocation of proteins and force-generation by linear molecular motors (which includes cytoskeletal motors, but also motors acting on DNA or RNA, like DNA-polymerases, RNA-polymerases and helicases), the ion transport in ion pumps, and the rotary-motor processes such as the one found in the F0F1-ATPase or the bacterial flagellar motor. Such isothermal rectifying processes and their underlying physical principles have been extensively reviewed in [129,233]. They all rely on a Langevin type of description of an overdamped particle of position x, moving in a spatially-periodic potential $W(x)$ that reflects the motor-filament interaction, and subjected to a viscous friction with coefficient ξ and a fluctuating force $f(t)$ that reflects the stochasticity of thermal fluctuations:

$$\xi \frac{\mathrm{d}x}{\mathrm{d}t} = -\partial_x W(x) + f(t) \,. \tag{4}$$

To rectify brownian motion, three different approaches have been mainly followed, namely (i) random forces $f(t)$ whose fluctuations do not satisfy the fluctuation-dissipation (FD) relation, (ii) fluctuating potentials $W(x,t)$ that are time-dependent, and (iii) particle fluctuating between states, where different states indexed by $i = 1, \ldots, N$ reflects different conformations of the protein and interactions with the filament.

In the following, no attempt will be made to extensively present the literature on molecular motors. We shall instead only briefly sketch the generic considerations of the main proposed models, and focus more closely on a particular example of them, the two-state model, which has allowed for an understanding of the appearance of spontaneous oscillations in systems of coupled motors. This generic mechanism has been proposed to underly axonemal beating, the generic mechanism that powers eukaryotic flagellar and ciliary-based motilities.

Phenomenological Description Close to Thermodynamic Equilibrium

Sufficiently close to thermal equilibrium, out-of-equilibrium perturbations can be described using a generic linear-response theory that introduces generalized forces which drive generalized currents [45]. In the context of molecular motors, the generalized forces that drive the system out of equilibrium are the mechanical force f_{ext} acting on the motor (including drag), and the chemical-potential difference $\Delta\mu$ of the chemical reaction ATP $\rightleftharpoons$ ADP + P_i that drives motor power [129,217]. Linear-response theory then gives:

$$v = \lambda_{11} f_{\text{ext}} + \lambda_{12} \Delta\mu \tag{5}$$

$$r = \lambda_{21} f_{\text{ext}} + \lambda_{22} \Delta\mu \,, \tag{6}$$

where the coefficients λ_{ij} are phenomenological response coefficients. Here λ_{11} and λ_{22} can be viewed respectively as a standard and generalized mobilities, and λ_{12} and λ_{21} as mechano-chemical couplings. Onsager relations impose that $\lambda_{12} = \lambda_{21}$, and the Second Law of Thermodynamics insures that the dissipation rate is positive: $T\dot{S} = f_{\text{ext}} v + r\Delta\mu \geq 0$. Whenever both of the two terms that appear in this inequality are positive, the system is passive, but it works as a motor when $f_{\text{ext}} v < 0$, and as generator of chemical energy when $r\Delta\mu < 0$. The latter function is not known for linear motors, but is the common mode of operation of F0F1-ATPase, the protein complex that synthetizes ATP from electro-chemical energy that is stored in proton gradients [5,290] (see Sect. "Molecular Motors"). The reversibility of this rotary engine can be related to the predicted reversibility that comes out of linear-response theory: in the absence of an external force, reversing the chemical potential difference $\Delta\mu$ should reverse the sign of the velocity v without a need for a change in the mechanism.

Hopping and Transport Models

Within the first class of models that we have mentioned earlier, namely hopping models between different discrete equilibrium states of the motor-filament system, generic transition rates and periodicity in theses transition rates, related to periodicity of the filament, are generally assumed. Within this framework, one can calculate the mean velocity v and the diffusion coefficient D of the molecular motor from analyzing the generic associated Master Equation [46]. For non-zero mean velocity to occur, at least one of the transitions between states must break detailed balance, a feature that can be associated with chemical energy consumption. In the simplest case of only two possible states of the motor protein, one can derive simple compact expressions for v and D [66]. Their dependence on the external force further leads to the derivation of the force-velocity curve, as well as a simple expression for the stall force, namely the force at which the motor protein ceases to progress on average.

To describe protein trafficking on a filament where many motors are simultaneously engaged, like it is commonly the case for example in organelle transport by kinesin proteins along microtubules, one can reduce the number of states that a motor can occupy to one per filament binding site. Motors are then represented by particles that move on a one-dimensional lattice with homogeneous transition rates, to which attachment and detachment rates from and toward the bulk can be added. This description belongs to a class of driven lattice-gas models that are used to study various transport phenomena, like ionic transport in solids or traffic flow with bulk on-off ramps [247]. In the simplest case of the absence of particle attachment and detachment, the model reduces to the Asymmetric Simple Exclusion Process (ASEP) [250], originally introduced to describe the translation of messenger RNA by ribosomes [173]. Including attachment-detachment rates, the next simplest case describes the space surrounding the filament as a reservoir of uniformly-distributed particles [147,216]. A second possibility is to include the dynamics of unbound particles explicitly, for example on a cubic lattice [161]. Boundary terms can also play an important role, and different possible choices have been considered depending on the biological situation [138].

Consequences of these models are illustrated by various important phenomena. Among these, one can find the followings: anomalous transport due to repeated attachments and detachments [4,161,205], domain walls that separate regions of high and low motor densities in the filament [161,216], phase separation in systems with two motor species [60], and phase transitions when cooperative binding-unbinding is introduced [139]. For a recent review on these collective traffic phenomena, see [40] and references therein.

The Two-State Model

One model that proved to be particularly useful for describing the rectification of brownian motion via coupling to chemical hydrolysis reactions, is the so-called “two-state model”. In this description, the molecular motor switches stochastically between two different interaction states with the filament, that are described by two different asymmetric and l-periodic potentials W_1 and W_2 representing polarity and periodicity of the filament [14,15,128,129,178, 217,218,227] (see Fig. 15a). The dynamics of this system can be conveniently represented in terms of two coupled Fokker–Planck equations that describe the evolution of the probability density $P_i(x,t)$ of the motor to be in state $i = 1,2$ at position x at time t. Explicitly, we have:

$$\begin{aligned} \partial_t P_1 + \partial_x J_1 &= -\omega_1 P_1 + \omega_2 P_2 \\ \partial_t P_2 + \partial_x J_2 &= \omega_1 P_1 - \omega_2 P_2 \,. \end{aligned} \tag{7}$$

The currents $J_i (i = 1,2)$ result from diffusion, interaction with the potentials W_i, and the external force f_{ext}:

$$J_i = \mu_i [-k_B T \partial_x P_i - P_i \partial_x W_i + P_i f_{\text{ext}}] \,. \tag{8}$$

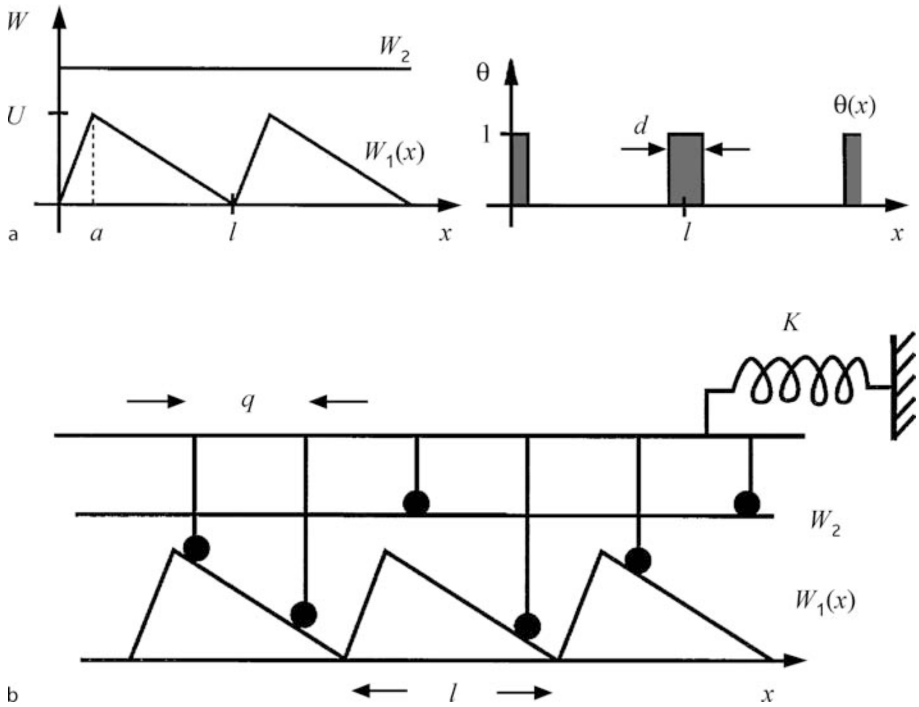
The transition rates $\omega_i(x) (i = 1,2)$ between the two states are driven out of equilibrium by ATP consumption, whose strength can be represented by a single parameter Ω using the following form:

$$\omega_1(x) = \omega_2(x) \exp\left[\frac{W_1(x) - W_2(x)}{k_B T}\right] + \Omega\Theta(x) \,, \tag{9}$$

where $\Theta(x)$ is a l-periodic function of integral one over one period. For $\Omega = 0$, detailed balance is satisfied. Within this formalism, it has been shown that both spatial symmetry and detailed balance need to be broken for directed motion to occur, which has been quantified in terms of an effective potential W_{eff} [227].

Coupled Motors and Spontaneous Oscillations

Directly interesting for eukaryotic cellular motility driven by cilia and flagella, is the motion of motors with respect



Cytoskeleton and Cell Motility, Figure 15

Schematic representation of the two-state model as used in [128] for the calculation of the collective behavior of rigidly-coupled non-processive motors. a, *left panel*: W_1 represents the attached state of the motor to the filament and is therefore asymmetric and *l*-periodic; W_2 represents the detached state where no interaction with the filament occurs. a, *right panel*: Illustration of the *l*-periodic function Θ (*x*) as it appears in Eq. (9) and which represents ATP consumption in the system, here at preferred locations of the motor on the filament, or at preferred configurations of the motor protein for ATP hydrolysis. b Schematic representation of a collection of rigidly-coupled motors with periodicity *q* interacting with the filament, and coupled to its framework by a spring of stiffness *K*. Under some conditions, this model leads to spontaneous oscillations. Both structures and behaviors are reminiscent of skeletal-muscle myofibrils' oscillations or axonemal beating (see Sect. "Axonemal Beating"). Source: reprinted from [128] with permission from The American Physical Society

to their associated cytoskeletal filament when a collection of them is rigidly coupled. Such structures are typical of skeletal-muscle structures (where Myosin II motors associate into the so-called "thick filaments"), or of the axonemal structure that drives oscillatory motions in cilia and flagella. Such systems have been studied using the two-state thermal-ratchet model [127,128], and a crossbridge model [30,97,108,112,279,280]. Here we shall discuss essentially the case of an ensemble of motors that are rigidly linked to each other and walk collectively on a cytoskeletal filament whose interaction with the motors is described by the two-state model [129]. In the case of randomly distributed motors, or motors distributed periodically with a period q that is incommensurate with the filament period l, the probability density $P(\xi, t)$ of finding a particle at position $\xi = x \bmod l$ in either state $i = 1$ or 2, approaches the value $1/l$ in the case of an infinitely-large number of motors. In a mean-field approximation, equations of motion for the probability densities read:

$$\begin{aligned}\partial_t P_1 + v\partial_\xi P_1 &= -\omega_1 P_1 + \omega_2 P_2\\ \partial_t P_2 + v\partial_\xi P_2 &= \omega_1 P_1 - \omega_2 P_2\ .\end{aligned} \tag{10}$$

The force-velocity curve can then be computed using the fact that $f_{\text{ext}} = \eta v - f$, where f_{ext} is the external force applied, η is the friction coefficient per motor protein, and f is the force per motor protein exerted by the potentials:

$$f = -\int_0^a \mathrm{d}\xi (P_1 \partial_\xi W_1 + P_2 \partial_\xi W_2)\ . \tag{11}$$

Expressing P_2 as $P_2 = 1/l - P_1$, and P_1 as a series expansion in powers of the velocity v, one finds a generic series expansion for the force-velocity curve f_{ext} as a function of v in the steady state. As a function of the distance to thermal equilibrium Ω, controlled by ATP consumption by the motors, and which appears as a control parameter for the dynamics with a critical value Ω_c, the curve $f_{ext}(v)$ can be strictly monotonic for $\Omega \leq \Omega_c$, or present some multi-valuated regions for $\Omega > \Omega_c$, where two stable velocity regimes exist for a given external force. For symmetric potentials, the system is quiescent with $v = 0$ at zero force for $\Omega \leq \Omega_c$, but present two possible opposite spontaneous velocities for $\Omega > \Omega_c$, a spontaneous symmetry breaking that is characteristic of second-order phase transitions with characteristic mean-field exponents. Such a reversible spontaneous movement has been observed in a motility assay with NK11 proteins, a mutant of the kinesin protein Ncd that has lost its directionality [55]. In addition, when the external force is varied, a hysteresis is found for $\Omega > \Omega_c$, an experimental observation of which has been reported for a myosin II motility assay under near-stalling conditions induced by electric fields [239]. Numerical simulations of both situations have been performed using the two-state model with a finite number of motors and in the presence thermal noise [17].

Axonemal Beating

The previous and related models have been used to describe the spontaneous oscillations that have been observed in skeletal-muscle myofibrils' oscillations or axonemal beating [72,209,287]. In these cases, it has been proposed that the coupling of the motor backbone to a spring prevents spontaneous steady-state velocities to occur, but instead leads to spontaneous oscillations [128] (see Fig. 15b). In the case of axonemal beating, of most relevance for eukaryotic cell swimming, the elastic force results from bending of the microtubules and leads to self-organization of the dynein motors. This collective behavior has been proposed to explain the bending waves of cilia and flagella [30,176] and analyzed in the framework of the two-state model [33,34]. Close to the oscillatory instability, wave-patterns can be computed, whose frequencies and shapes depend on the filament length and boundary conditions, and which are in good agreement with observed flagellar beating patterns [278].

In the case of cilia however, beating patterns are typically assymetric, like it is at best exemplified in the case of the two cilia of the green alga *Chlamydomonas* [29], an observation that cannot be accounted for by investigating beating patterns at the oscillatory instability only. Using the same underlying model, it has been proposed that this assymetry originates from the presence of transverse external flow that occurs as the organism is swimming [89]. In that case, the cilium tends to beat faster and quite straight in the direction of the flow, whereas it comes back slower and more curved against it, a beating pattern that evokes power and recovery strokes. Hydrodynamics has also been proposed to be responsible for dynamic coupling of adjacent cilia, which results in both spontaneous symmetry breaking and synchronization of their beating pattern [89]. This effect could be at the basis of the observed beating waves that propagate for example on the surface of *paramecia*[21] as they swim, which originate from a constant phase difference in the beating of the adjacent cilia, and which have been called *metachronal waves* [29,88,89]. This could also underly the breaking of symmetry that occurs in mammalian development during gastrulation, and which is responsible for left-right asymmetry. In that case, it has been shown that beating of cilia located in a transiently-formed epithelial chamber known as the *node*, create a directional flow which transports signaling molecules preferentially to one side [29,117,185]. There, beating patterns are unusual in that cilia swirl in vortical fashion rather than beat [208], and hydrodynamic-driven synchronization of these three-dimensional beating patterns has also been studied [281].

Putting It Together: Active Polymer Solutions

The last part of this review is devoted to the presentation of some generic descriptions of the cell cytoskeleton, when considered as a network of long protein filaments that are cross-linked by a variety of smaller proteins. As already discussed, filamentous proteins that are involved in the cell-cytoskeleton dynamics are mostly F-actin and microtubules (made of G-actin and tubulin monomers), with which interact cross-linkers that can be either passive and stationary (such as α-actinin), or active and mobile, consisting then of clusters of molecular motors (mostly myosin and kinesin motors) (see Sect. "The Cell Cytoskeleton"). To model these systems, different complementary approaches have been developed, namely computer simulations [202,203,260], and analytical descriptions that can be roughly divided into three categories, namely microscopic, mesoscopic, and macroscopic or phenomenological hydrodynamic descriptions. The first analytical approaches that have been developed correspond to mesoscopic descriptions. There, starting from a microscopic description of the filaments, the effect of active cross-linkers is described via motor-induced relative velocities of paired filaments, where the form of such velocities is inferred from general symmetry considerations [144,145,148,201,251]. Microscopic approaches start from what is known about the properties of the different molecular players involved and their interactions, and aim

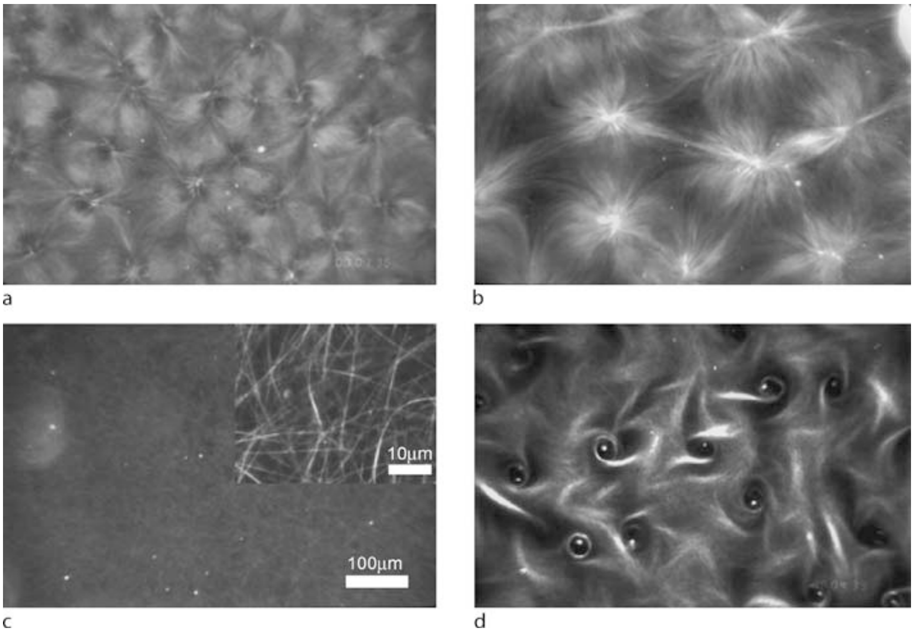
[21] An illustration of a paramecium can be seen in Fig. 1, center.

to build large-scaled coarse-grained theories from statistical physics' principles [2,3,11,12,146,165,166,168]. Finally, macroscopic hydrodynamic approaches have adopted a more phenomenological point of view: they harness the generic symmetry and dynamical properties of the players involved, to derive directly effective continuous theories in terms of a few coarse-grained fields [94,125,130,150, 151,152,157,231,245,254,267,282,295]. Recently, attempts have been made to bridge microscopic to macroscopic models, and compare what results in being similar and different in the two types of approaches [3,164,166].

Interests for describing the cytoskeleton as an ensemble of filamentous polymers actively connected by crosslinkers have come to the scene since self-organizations of motor-filament mixtures were observed experimentally [203,260,272]. Among these, complex patterns that include asters, vortices, spirals and connected poles or networks have been observed in confined quasi-two-dimensional systems in in vitro experiments [203,260] (see Fig. 16). Patterns where shown to be selected in a way that is dependent on motor and ATP concentrations, and numerical simulations based on microscopic models of rigid rods connected by active elements have shown to be capable of reproducing the experimental results [260]. Further experiments were performed on systems that resemble more closely a living cell, and which, while being simplified versions of it, still exhibit some of its behaviors. Along these lines, formations of bipolar spindles that do not contain any microtubule-organizing center were observed using cell extracts [116], and cell fragments that contain only the actin cortex where found to self-propagate on a substrate, with coexistence of locomoting and stationary states [59,277].

Mesoscopic Approaches

Theoretical modeling of the cell cytoskeleton have benefitted from the knowledge accumulated in equilibrium statistical physics of polymer solutions and liquid crystals [44]. However, the cell cytoskeleton is an active medium for which new analysis techniques needed to be developed in order to describe, for example, its ability to actively self-organize, exert forces and create motion. First, theoretical models have aimed to describe pattern formations in sys-



Cytoskeleton and Cell Motility, Figure 16

Different large-scale patterns formed through self-organization of microtubules and kinesin motors as reported in [203]. Initially uniform mixtures of proteins heated to 37°C displayed different patterns after 7 min of self-organization. Patterns are shown at equal magnification; the samples differ in kinesin concentration. a A lattice of asters and vortices obtained at 25 g ml^{-1} kinesin concentration. b An irregular lattice of asters obtained at 37.5 g ml^{-1} kinesin concentration. c Microtubules form bundles at 50 g ml^{-1} kinesin concentration (scale bar, 100 µm); insert: at higher magnification (scale bar, 10 µm). d A lattice of vortices obtained at a kinesin concentration smaller than 15 g ml^{-1}. Source: courtesy of François Nédélec; reprinted from [203] with permission from Nature Publishing Group

tems of actively-driven rigid filaments in one-dimensional geometries [144]. Such configurations are represented in vivo for example by stress fibers that are important for cell-force generation, contractile rings that form during cytokinesis, or the formation of filopodia for forward protrusion during amoeboid motility. There, dynamical equations that govern individual filaments were introduced using "mesoscopic" mean-field models, where the relative sliding of paired polar filaments is described by an effective relative velocity that is induced by many individual events of motor activity. General constraints on these relative velocity fields are imposed by symmetry considerations that rely on the orientational polarity of the filaments (see, e. g., [144]). In such systems, polarity sorting [201], contraction [144,251], as well a propagating waves [148] emerge from the models. Interestingly, it has been shown that relative velocity of filaments of the same orientation is important for contraction to occur [144], a phenomenon that has been suggested to rely on motor-density inhomogeneities along the filament that create inhomogeneous filament interactions along their lengths [147]. More generally, the whole bifurcation diagram of this generic one-dimensional model has been established, and motor-distribution dynamics has been introduced that lead to contractile states with the generation of contractile forces, of most relevance for stress fibers' as well as contractile rings' dynamics [145]. Interestingly, the simplest version of these models, when only one possible polymer orientation is considered, has been mapped to hopping models that describe driven-diffusive systems [199]. In the absence of active cross-linkers, the model reduces to a class of hopping models known as the Zero Range Process (ZRP), for which exact analytical solutions of the steady state as well as one-dimensional phase transitions have been described [92]. In the generic case however, the dynamics defines a new class of driven-diffusive systems, which can still be mapped in some cases to the ZRP analytic solution, even though with a different criterium for condensation to occur [199].

Microscopic Approaches

Microscopic approaches to describing the cell cytoskeleton dynamical behavior model explicitly all the considered different processes and interactions that occur between the different molecular players, and aim to derive effective dynamical equations for the different density fields that enter the description, by coarse-graining the microscopic dynamics. Most studies that have done so model the motor-filament system as an ensemble of rigid rods of fixed lengths, which interact via point-like cross-linkers that can induce relative sliding as well as rotational motions [2,3,146,165,166]. Exceptions to this rule are theoretical descriptions of the mechanical response of active-filament solutions to high frequency stimuli [162,168]. There, anomalous fluctuations occur that are dominated by the bending modes of the filaments in combination with the activity of the cross-linkers. Inspired by polymer physics at thermodynamic equilibrium, excluded-volume interactions as well as entanglements are taken into account in the description. In particular, the system exhibits accelerated relaxation at long times due to directed reptation that relies on active phenomena.

Attempts at deriving the motor-mediated interaction between filaments from microscopic descriptions have been performed in [2,3,11,12,165]. See also the review [163]. In [11,12], a generalization of the Maxwell model of binary collisions in a gas is used to describe the dynamics of polar rods whose inelastic and anisotropic interactions reflect the presence of active crosslinkers. Orientational instabilities lead to bundling as well as the formation of asters and vortices patterns. In [2,3,165], filaments are described as rigid rods of fixed length, and hydrodynamics is obtained by coarse-graining the Smoluchowsky equation for rods in solution, coupled via excluded-volume and motor-mediated interactions. There are two main motor-mediated mechanisms for force exchange among filaments. First, active crosslinkers induce bundling of filaments, building up density inhomogeneities. Second, they induce filament sorting as a function of their polarization state. As a result, phase diagrams are derived that show instabilities of the homogeneous states at high filaments' and crosslinkers' densities. In particular, all homogeneous states are rendered unstable by the same mechanism of filament bundling, a fact reminiscent of the effect described in [144] where the interaction between filaments of the same orientation has been shown to be important for contraction to occur. Interestingly, the broken directional symmetry of the polarized state yields an effective drift velocity that describes filament advection. This convective-type term describes a genuine out-of-equilibrium contribution that is structurally not present in phenomenological descriptions based on systematic linear expansions close to thermodynamic equilibrium (see below). Such a term is reminiscent of the one introduced in earlier studies of self-propelled nematic particles [230,254,267], as well as of the explicit flow of the solvent taken into account in [150,151,282]. Other effects of higher-order nonlinear terms have also been discussed in [292,293], where pattern selection between stripe patterns and periodic asters occurs via nonlinear interactions.

Macroscopic Phenomenological Approaches: The Active Gels

The third category of approaches that have been developed to try to understand the dynamical behavior of the cell cytoskeleton as a whole, are effective phenomenological theories which rely on the hypothesis that large length- and time-scales behaviors of the cytoskeleton are largely independent of the microscopic details that underly its dynamics, but depend instead only on a few macroscopic fields that capture the relevant behavior. Sufficiently close to thermodynamic equilibrium, these relevant fields describe the *hydrodynamic modes* (or *slow modes*) of the dynamics, namely the modes whose relaxation rates go to zero at long wavelengths. As for equilibrium systems close to a critical point, such hydrodynamic modes correspond to the *conserved densities* on the one hand, and the *order parameters* that break continuous symmetries on the other hand [45,100]. To write generic theories for the dynamics of these hydrodynamic modes, standard approaches consist in writing systematic expansions in the different couplings that are allowed by the symmetry properties of the system. In the vicinity of a critical point, which occurs generally when a continuous symmetry is spontaneously broken after a second-order phase transition has been traversed, the concept of renormalization group has given a theoretical framework to identify universality classes: starting from the full nonlinear expansions of the underlying stochastic dynamics, only a few relevant parameters matter for the asymptotic scaling laws that occur at the transition [65,100,294]. Even though originally developed to study equilibrium critical points, the renormalization-group concept has allowed for the characterization of some out-of-equilibrium universality classes (see e. g. [70,238,261] and the review [98]).

Away from such remarkable points however, any term allowed by symmetry in a systematic expansion is a priori relevant. The standard approach for systems close to thermodynamic equilibrium consists in writing generalized thermodynamic forces and fluxes that are related to each other by linear-response theory. Constraints on the generic coupling constants to linear order emerge from the spatio-temporal symmetries of the system, and correspond to the Onsager relations and the Curie principle [45]. Inspired by the dynamics of liquid crystals [44,183], a hydrodynamic theory has been developed that describes the cytoskeleton as a visoelastic polar gel, driven out of equilibrium by a source of chemical energy [150,151,152,282,295] (this work has been reviewed in [130]). Among other applications, such or similar approaches have been applied to the description of pattern formation in motor-microtubule mixtures [157,245], as well as the collective dynamics of self-propelled particles [52,94,254,267].

Originally, this hydrodynamic theory has been presented as a generic theory for active viscoelastic materials made of polar filaments, referred to as *active polar gels* [151]. Within the framework of the previously-described general formalism, here applied to the cytoskeleton, conserved quantities are the different number densities that enter the dynamics, namely the number densities of subunits in the gel, of free monomers, and of respectively bound and unbound motors to the filaments. To these must be added the solvent density and the total mechanical momentum. Source terms in the conservation equations correspond to polymerization and depolymerization of cytoskeleton filaments, attachment and detachment of motor proteins to the filaments, and the potential presence of an external force. Order parameters correspond to orientational order parameters that originate from the polarity of the filaments. Namely, they correspond to momenta of the local polarization vector of individual filaments $\boldsymbol{u}$, and most often only the first momentum $\boldsymbol{p} = \langle \boldsymbol{u} \rangle$ is considered that represents the locally-averaged polarity in the gel[22]. To these must be added a crucial parameter that drives the system out of equilibrium, and which originates from the actively-maintained source of chemical energy in the cell, corresponding to out-of-equilibrium concentrations of ATP versus ADP and P_i. This parameter $\Delta\mu$ is expressed as the difference in chemical potentials of ATP versus ADP plus P_i: $\Delta\mu = \mu_{\mathrm{ATP}} - (\mu_{\mathrm{ADP}} + \mu_{\mathrm{P}_i})$. After identification of the different conjugated generalized fluxes and forces, that are split into dissipative and reactive parts as a function of their properties under time-reversal symmetry, the constitutive equations that specify the dynamics are written in terms of a generalized Maxwell model, which describes the viscoelastic dynamical properties of the gel. Under its simplest form and for nonpolar viscoelastic gels, the Maxwell model writes

$$\left(1 + \tau \frac{D}{Dt}\right) \sigma'_{\alpha\beta} = 2\eta \left(v_{\alpha\beta} - \frac{1}{d} \delta_{\alpha\beta} v_{\gamma\gamma} \right) + \bar{\eta} \delta_{\alpha\beta} v_{\gamma\gamma} \,, \tag{12}$$

in d dimensions. Here $v_{\alpha\beta}$ and $\sigma'_{\alpha\beta}$ are the symmetric parts respectively of the velocity-gradient tensor $\partial_\alpha v_\beta$ and the viscous stress tensor, η and $\bar{\eta}$ are respectively the shear and bulk viscosities, and $\tau = E/\eta$ is the viscoelastic relaxation time that is related to the Young elastic modulus E

[22] The next momentum $q_{\alpha\beta} = \langle u_\alpha u_\beta - d^{-1} \, \boldsymbol{p}^2 \, \delta_{\alpha\beta} \rangle$, where d is the dimension of space, is a symmetric traceless tensor of order two that corresponds to nematic order.

and to the shear viscosity η. This relaxation time describes the crossover between an elastic behavior at short times that resembles that of a solid gel and a viscous behavior at long times that resembles that of a fluid[23]. Finally, D/Dt represents a convective corotational derivative that takes into account invariance with respect to translations and rotations in the system. In the general framework of active polar gels close to thermodynamic equilibrium, generic linear couplings are added to this model following the general procedure described above. Extensive presentations of the complete formalism can be found in refs. [130,151], which also include discussions about its limitations and some of its possible extensions. In particular, in [130], extensions that aim to include the contributions of rotational viscoelasticity, of some nonlinear couplings and of passive as well as active sources of noise are briefly discussed. The main developed extension so far concerns the generalizations of this formalism to multi-component active gels that are now being developed [125,167]. These allow in particular to take into account the possible permeation of the cytosol through the cytoskeletal gel, which affects force balance in the system, and which might be of importance for cell motility.

Despite its recent development, this generic theory of active polar gels has been applied to the description of some systems that are of particular relevance for cell motility or experimental situations observed in vitro. Its first application was the study of topological defects in the polarity field of the gel that lead to the formation of patterns such as asters, vortices and spirals [150]. As a function of two dimensionless parameters representing the relative strength of the coupling to the chemical potential $\Delta\mu$ and to the bend and splay moduli of the polar gel, a phase diagram was derived where vortices and asters give rise to rotating spirals via dynamic instabilities. These relate to the spatial patterns that have been observed in vitro [203,260], as well as to the creation of spontaneous motion, of most relevance for cell motility. In a different geometry, namely a cylinder of finite diameter and length, the formalism has been applied to establish a phase diagram of ring formation that contains phases of one or multiple rings, and which can be quiescent or oscillating [295]. This is relevant for understanding the formation and localization of cortical rings that form prior to cytokinesis and for which double-ring formation has been observed with certain plant cells [84]. To understand the generation of active flows that might be of relevance for cell crawling, a generic phase diagram has been derived for a two-dimensional active polar film that is compressible [282]. Compressibility here might refer to different thicknesses in a three-dimensional incompressible gel that is described in two dimensions after integration of the density fields over its thickness. Within this framework, density fluctuations couple generically to polarity splay, and different topological phases of the gel-polarity organization are found that could correspond to some of the previously-observed patterns in the experimental literature. Finally, the description of spontaneous movements of thin layers of active gels has been applied to the study of cell locomotion on a solid substrate that occurs via the protrusion of the actin-filled lamellipodium at the leading edge of the cell [152]. Reducing the lamellipodium description to a two-dimensional gel protruding in one dimension, and with a spatially-dependent thickness, the steady-state thickness profile as well as the flow and force fields have been computed. One particularly striking aspect of cell crawling that is described by this formalism is the presence of a retrograde flow of the gel as the cell is crawling. This aspect has been quantified in earlier experiments performed on fish epidermal keratocytes [275]. It has been shown that while the cell is crawling, treadmilling of actin filaments happens faster than global motion of the cell, such that the actin cortex is moving rearward with respect to the substrate, in a direction opposite to the movement of the cell [42,131,160]. Similar questions have been addressed using different theoretical frameworks in [7,142,242].

Comparisons of the Different Approaches to Describing Active Polymer Solutions

With these different ways of approaching the description of the dynamics of the cell cytoskeleton as a whole, a natural question is to ask to what extent these different approaches are similar and different, and which aspects of the cytoskeleton or cell behavior can be or not described by each of the theories. For answering these questions, connections between the different approaches have been made, first between mesoscopic and hydrodynamic descriptions [149,164]. In [164], a generalization of the mesoscopic model introduced in [144,148] is developed to obtain a set of continuum equations in unconfined geometries. A phase-diagram is derived that results from the stability analysis of the homogeneous state of actively crosslinked polymers, taking into account excluded-volume interactions and estimates of entanglement in two and three dimensions. It is found that an instability occurs as the bundling rate between filaments of the same orientation is

[23] Note that only one relaxation time is assumed to characterize the system, as some experiments suggest that a power-law distribution of relaxation times is better suited to describe cytoskeleton dynamics, potentially because of some scale-invariant dynamical properties in the system [18,61].

increased, which at low filament density happens first via a density-fluctuation instability, and at high filament density via an orientational-fluctuation instability. In the presence of a finite sorting rate between filaments of different orientations, propagating modes appear that reflect oscillatory behavior. In [149], the continuum theory is related to nonlocal descriptions of filament-motor systems, since filaments can transmit stresses over finite distances. The effective parameters of the continuum theory are recovered from the previously-published mesoscopic description [144], even though with missing coefficients that are thought to correspond to microsocpic multi-particle interactions, not described in [144]. Effects of polymerization-depolymarization dynamics via effective source and sink terms in the local filament densities are also discussed (see also [125,167]) – like it is the case in the effective macroscopic theories – as well as the role of polarity. In particular, it is found that nonpolar arrangements of filaments do not exhibit oscillatory instabilities and propagating modes, which might be of relevance for muscle sarcomeric structures. As seen previously, in these systems, spontaneous oscillations that have been observed correspond more to oscillatory instabilities of rigidly-coupled collective motors than to solitary-wave solutions, as they are found in systems of active polar filaments.

Microscopic theories present the advantage of being able in principle to give rise to a full description of a given system with arbitrary precision and specificity, and to take into account the nonlinear effects that are of direct relevance for the system's behavior. However, they rely on the microscopic knowledge that one has on the system under consideration, and are therefore limited by the available information on the different agents. In addition, they end up with effective descriptions that are model-dependent, in that the different parameters of the so-obtained theory, which describe its physical behavior, depend on the interactions that are taken into account at the microscopic level. Also, an important aspect of active cytoskeleton dynamics that is usually not described in such microscopic approaches is the very important phenomenon of treadmilling that relies on polymerization-depolymerization dynamics of cytoskeletal polymers, and which we have seen to be of crucial importance for some mechanisms of cellular motility such as *Listeria* propulsion or nematode-sperm-cell locomotion (see Sect. "Filament-Driven Motility"). However, despite the absence of these effects, which are taken into account effectively in macroscopic hydrodynamic descriptions, such microscopic approaches allow for the derivation of the forces exchanged between the motors and the filaments from microscopic knowledge, while they appear as effective parameters of unknown explicit origin in effective macroscopic descriptions. Thereby, questions can be addressed that concern the role played by the specific physical properties of motor-filament interactions at the microscopic level in controlling the system behavior on large scales. Indeed, the richness of the observed self-organized structures raises the question of how much is generic, and how much is specific in cytoskeleton behavior. For example, experiments have shown that very different self-organizing structures occur with processive as opposed to non-processive motor proteins: at high motor concentrations, microtubule-kinesin mixtures self-organize in a variety of spatial patterns [203,260], as homogeneous states are more robust with acto-myosin systems [106], an effect that can be thought of as the influence of motor processivity on the dynamical large-scale parameters [164].

Extensions and Future Directions

Cell motility is a complex and integrated process that relies on self-organization of the cytoskeleton, carefully and precisely orchestrated by the cell with the help of numerous different types of molecular players. If one includes the subcellular movements that are responsible for intracellular traffic and material exchange between the inner parts and external parts of the cell, cell-motility mechanisms are found to ground the activity of all life forms on earth. When looked under the microscope, motility mechanisms and structural changes of the diverse cell types appear so vast and various that a comprehensive understanding of their underlying mechanisms seems to be an overwhelming challenge. However, as we have seen from the literature covered in the present article, our understanding of cell motility has tremendously progressed over the past two decades. On the one hand, complexity has even further emerged, since the biochemical characterization of the molecular players involved has revealed that at least hundreds of different protein types participate in the structural and dynamical organization of the cell cytoskeleton. On the other hand, despite the existence of such very complex regulation processes that rely on the integrated interplay of the whole set of different molecular players, the characterization of the cell cytoskeleton has revealed that its main structures and functions are due to just a few types of key proteins, namely three types of biopolymers and three superfamilies of molecular motors. Even more striking is the evolutionary conservation of the main molecular players involved in building the cytoskeleton dynamical filaments, both within the eukaryotic domain of life on the direct sequence point of view, and even across the three domains of life when structural and functional properties are considered. These striking observations indicate that the different underlying mechanisms of cell motility all rely on generic principles that can be understood on a biophysical point of view. In addition, further help from micro-manipulation and fluorescence-microscopy techniques, as well as the development of simplified systems based on gene-expression control and biomimetic artificial systems, has enabled the experimental biophysical investigation of the different specific aspects of the processes at play.

The theoretical analyses reviewed in this article have shown that central concepts that underly the cytoskeleton dynamics are self-organization and dynamic instabilities, here grounded on out-of-equilibrium nonlinear dynamics', thermodynamics' and statistical physics' principles. Such concepts are at the basis of all the theoretical approaches that have been developed to understand the mechanisms of diverse phenomena such as polymerization-depolymerization force and movement generation, molecular motors' individual behaviors and collective phenomena, as well as the generic behaviors of active-polymer solutions which lead to a description of the cytoskeleton dynamics as a whole. On all of these topics, microscopic as well coarse-grained effective macroscopic approaches have been developed. As already discussed, they both have their advantages, powers and limitations, and represent important complementary steps in the ultimate goal of an integrated description of the universal principles that underly cell motility.

As we have seen in this article, our understanding of cell motility and cell cytoskeleton dynamics has grandly benefitted from the interplay between experiments and modeling, each for its own reasons guiding the other in its directions of investigation. To further understand the integrated processes at play in cell motility, such fruitful interactions will certainly be further required and developed. On the theoretical point of view, bridges between understanding simplified systems or some particular aspects of cell motility and the phenomenon of cell motility as a whole at the global cellular level, have already started being investigated, but further developments of these two different ways of approaching the cytoskeleton dynamics as well as understanding the links that ultimately relate them are required. Another important aspect whose understanding represents a challenge is the potentially crucial role of noise that has been so far most of the time absent from the macroscopic effective theoretical approaches. Indeed, noise in nonlinear dynamical systems is known to potentially have important constructive effects, whose main representatives are stochastic resonance, coherence resonance and noise-induced transitions, as well as the extensive gallery of different spatially-extended phenomena such as array-enhanced stochastic and coherence resonance, or noise-enhanced synchronization of nonlinear oscillators (see e. g. [220]). Such phenomena have already been recognized to play an important role in some biological cytoskeleton-based pattern formations (see e. g. [105]), and could play a crucial role in driving other cytoskeletal self-organization phenomena, especially close to dynamical instabilities, where the effect of noise is highest. Finally, having at hand the underlying biochemical and biophysical mechanisms of cell forces and motility, a great challenge is to understand self-organization at yet larger scales, namely in animal tissues, where collections of cells present integrated coherent behaviors that drive diverse key processes such as morphogenesis, wound healing, immune response, tumor development and metastases formations. There, the same scheme involving "microscopic" as well as effective "macroscopic" approaches can certainly play an equally major role, "microscopic" approaches then potentially integrating the whole knowledge acquired at the level of a single cell, and partially reviewed in this article.

Acknowledgments

To write this article, I grandly benefited from Karsten Kruse's habilitation thesis, which constituted a very good starting point as an extensive review of the existing biophysical literature on the cytoskeleton. I acknowledge Cécile Sykes and Jean-François Joanny for discussions, careful reading of the manuscript and constructive criticisms and suggestions. I also acknowledge Andrew Callan-Jones for pointing to me important references.

Bibliography

Primary Literature

1. Abercrombie M (1980) The crawling movement of metazoan cells. Proc R Soc Lond B Biol Sci 108:387–393
2. Ahmadi A, Liverpool TB, Marchetti MC (2005) Nematic and polar order in active filament solutions. Phys Rev E Stat Nonlin Soft Matter Phys 72(6 Pt 1):060901
3. Ahmadi A, Marchetti MC, Liverpool TB (2006) Hydrodynamics of isotropic and liquid crystalline active polymer solutions. Phys Rev E Stat Nonlin Soft Matter Phys 74(6 Pt 1):061913
4. Ajdari A (1995) Transport by active filaments. Europhys Lett 31(2):69–74
5. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular biology of the cell, 4th edn. Garland, New York
6. Albertson DG (1984) Formation of the first cleavage spindle in nematode embryos. Dev Biol 101(1):61–72
7. Alt W,Dembo M (1999) Cytoplasm dynamics and cell motion: two-phase flow models. Math Biosci 156(1–2):207–228

8. Amos LA, van den Ent F, Lowe J (2004) Structural/functional homology between the bacterial and eukaryotic cytoskeletons. Curr Opin Cell Biol 16(1):24–31
9. Ananthakrishnan R, Ehrlicher A (2007) The forces behind cell movement. Int J Biol Sci 3(5):303–317
10. Andersen SS (2000) Spindle assembly and the art of regulating microtubule dynamics by MAPs and Stathmin/OP18. Trends Cell Biol 10(7):261–267
11. Aranson IS, Tsimring LS (2005) Pattern formation of microtubules and motors: Inelastic interaction of polar rods. Phys Rev E Stat Nonlin Soft Matter Phys 71(5 Pt 1):050901
12. Aranson IS, Tsimring LS (2006) Theory of self-assembly of microtubules and motors. Phys Rev E Stat Nonlin Soft Matter Phys 74(3 Pt 1):031915
13. Astumian RD (1997) Thermodynamics and kinetics of a brownian motor. Science 276(5314):917–922
14. Astumian RD, Bier M (1994) Fluctuation driven ratchets: Molecular motors. Phys Rev Lett 72(11):1766–1769
15. Astumian RD, Bier M (1996) Mechanochemical coupling of the motion of molecular motors to ATP hydrolysis. Biophys J 70(2):637–653
16. Ausmees N, Kuhn JR, Jacobs-Wagner C (2003) The bacterial cytoskeleton: an intermediate filament-like function in cell shape. Cell 115(6):705–713
17. Badoual M, Jülicher F, Prost J (2002) Bidirectional cooperative motion of molecular motors. Proc Natl Acad Sci USA 99(10):6696–6701
18. Balland M, Desprat N, Icard D, Fereol S, Asnacios A, Browaeys J, Henon S, Gallet F (2006) Power laws in microrheology experiments on living cells: Comparative analysis and modeling. Phys Rev E Stat Nonlin Soft Matter Phys 74(2 Pt 1):021911
19. Belmont LD, Hyman AA, Sawin KE, Mitchison TJ (1990) Realtime visualization of cell cycle-dependent changes in microtubule dynamics in cytoplasmic extracts. Cell 62(3):579–589
20. Berg HC (2003) The rotary motor of bacterial flagella. Annu Rev Biochem 72:19–54
21. Berg JS, Powell BC, Cheney RE (2001) A millennial myosin census. Mol Biol Cell 12(4):780–794
22. Bernheim-Groswasser A, Wiesner S, Golsteyn RM, Carlier M-F, Sykes C (2002) The dynamics of actin-based motility depend on surface parameters. Nature 417(6886):308–311
23. Bernheim-Groswasser A, Prost J, Sykes C (2005) Mechanism of actin-based motility: a dynamic state diagram. Biophys J 89(2):1411–1419
24. Bershadsky A, Kozlov M, Geiger B (2006) Adhesion-mediated mechanosensitivity: a time to experiment, and a time to theorize. Curr Opin Cell Biol 18(5):472–481
25. Bishop AL, Hall A (2000) Rho GTPases and their effector proteins. Biochem J 348(Pt 2):241–255
26. Bork P, Sander C, Valencia A (1992) An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. Proc Natl Acad Sci USA 89(16):7290–7294
27. Bottino D, Mogilner A, Roberts T, Stewart M, Oster G (2002) How nematode sperm crawl. J Cell Sci 115(Pt 2):367–384
28. Boukellal H, Campas O, Joanny J-F, Prost J, Sykes C (2004) Soft listeria: Actin-based propulsion of liquid drops. Phys Rev E Stat Nonlin Soft Matter Phys 69(6 Pt 1):061906
29. Bray D (2001) Cell movements, 2nd edn. Garland, New York
30. Brokaw CJ (1975) Molecular mechanism for oscillation in flagella and muscle. Proc Natl Acad Sci USA 72(8):3102–3106
31. Bullock TL, Roberts TM, Stewart M (1996) 2.5 A resolution crystal structure of the motile major sperm protein (MSP) of Ascaris suum. J Mol Biol 263(2):284–296
32. Bustamante C, Macosko JC, Wuite GJ (2000) Grabbing the cat by the tail: manipulating molecules one by one. Nat Rev Mol Cell Biol 1(2):130–136
33. Camalet S, Jülicher F (2000) Generic aspects of axonemal beating. New J Phys 2:1–23
34. Camalet S, Jülicher F, Prost J (1999) Self-organized beating and swimming of internally driven filaments. Phys Rev Lett 82(7):1590–1593
35. Cameron LA, Footer MJ, van Oudenaarden A, Theriot JA (1999) Motility of ActA protein-coated microspheres driven by actin polymerization. Proc Natl Acad Sci USA 96(9): 4908–4913
36. Cameron LA, Giardini PA, Soo FS, Theriot JA (2000) Secrets of actin-based motility revealed by a bacterial pathogen. Nat Rev Mol Cell Biol 1(2):110–119
37. Carlsson AE (2001) Growth of branched actin networks against obstacles. Biophys J 81(4):1907–1923
38. Carlsson AE (2003) Growth velocities of branched actin networks. Biophys J 84(5):2907–2918
39. Chaen S, Inoue J, Sugi H (1995) The force-velocity relationship of the ATP-dependent actin-myosin sliding causing cytoplasmic streaming in algal cells, studied using a centrifuge microscope. J Exp Biol 198(Pt 4):1021–1027
40. Chowdhury D, Schadschneider A, Nishinari K (2005) Physics of transport and traffic phenomena in biology from molecular motors and cells to organisms. Phys Life Rev 2(4):318–352
41. Cossart P, Bierne H (2001) The use of host cell machinery in the pathogenesis of Listeria monocytogenes. Curr Opin Immunol 13(1):96–103
42. Coussen F, Choquet D, Sheetz MP, Erickson HP (2002) Trimers of the fibronectin cell adhesion domain localize to actin filament bundles and undergo rearward translocation. J Cell Sci 115(Pt 12):2581–2590
43. de Boer P, Crossley R, Rothfield L (1992) The essential bacterial cell-division protein FtsZ is a GTPase. Nature 359(6392): 254–256
44. de Gennes P-G, Prost J (1993) The physics of liquid crystals, 2nd edn. Clarendon Press, Oxford
45. de Groot SR, Mazur P (1984) Non-equilibrium thermodynamics. Dover Publications Inc, New York
46. Derrida B (1983) Velocity and diffusion constant of a periodic one-dimensional hopping model. J Stat Phys 31(3):433
47. Desai A, Verma S, Mitchison TJ, Walczak CE (1999) Kin I kinesins are microtubule-destabilizing enzymes. Cell 96(1): 69–78
48. Dickinson RB, Purich DL (2007) Nematode sperm motility: nonpolar filament polymerization mediated by end-tracking motors. Biophys J 92(2):622–631
49. Dickinson RB, Caro L, Purich DL (2004) Force generation by cytoskeletal filament end-tracking proteins. Biophys J 87(4):2838–2854
50. Dogterom M, Yurke B (1997) Measurement of the force-velocity relation for growing microtubules. Science 278(5339): 856–860
51. Dogterom M, Kerssemakers JWJ, Romet-Lemonne G, Janson ME (2005) Force generation by dynamic microtubules. Curr Opin Cell Biol 17(1):67–74
52. Dombrowski C, Cisneros L, Chatkaew S, Goldstein RE, Kessler

JO (2004) Self-concentration and large-scale coherence in bacterial dynamics. Phys Rev Lett 93(9):098103

53. dos Remedios CG, Chhabra D, Kekic M, Dedova IV, Tsubakihara M, Berry DA, Nosworthy NJ (2003) Actin binding proteins: Regulation of cytoskeletal microfilaments. Physiol Rev 83(2):433–473
54. Duke T, Leibler S (1996) Motor protein mechanics: a stochastic model with minimal mechanochemical coupling. Biophys J 71(3):1235–1247
55. Endow SA Higuchi H (2000) A mutant of the motor protein kinesin that moves in both directions on microtubules. Nature 406(6798):913–916
56. Engler AJ, Sen S, Sweeney HL, Discher DE (2006) Matrix elasticity directs stem cell lineage specification. Cell 126(4): 677–689
57. Erickson HP (2007) Evolution of the cytoskeleton. Bioessays 29(7):668–677
58. Erickson HP, Taylor DW, Taylor KA, Bramhill D (1996) Bacterial cell division protein FtsZ assembles into protofilament sheets and minirings, structural homologs of tubulin polymers. Proc Natl Acad Sci USA 93(1):519–523
59. Euteneuer U, Schliwa M (1984) Persistent, directional motility of cells and cytoplasmic fragments in the absence of microtubules. Nature 310(5972):58–61
60. Evans MR, Foster DP, Godreche C, Mukamel D (1995) Spontaneous symmetry breaking in a one dimensional driven diffusive system. Phys Rev Lett 74(2):208–211
61. Fabry B, Maksym GN, Butler JP, Glogauer M, Navajas D, Fredberg JJ (2001) Scaling the microrheology of living cells. Phys Rev Lett 87(14):148102
62. Faivre-Moskalenko C, Dogterom M (2002) Dynamics of microtubule asters in microfabricated chambers: The role of catastrophes. Proc Natl Acad Sci USA 99(26):16788–16793
63. Fernandez P, Pullarkat PA, Ott A (2006) A master relation defines the nonlinear viscoelasticity of single fibroblasts. Biophys J 90(10):3796–3805
64. Feynman RP, Leighton RB, Sands M (1963) The Feynman lectures on physics, vol 1. Addison-Wesley, Reading
65. Fisher ME (1998) Renormalization group theory: Its basis and formulation in statistical physics. Rev Mod Phys 70(2): 653–681
66. Fisher ME, Kolomeisky AB (1999) The force exerted by a molecular motor. Proc Natl Acad Sci USA 96(12):6597–6602
67. Fletcher DA, Theriot JA (2004) An introduction to cell motility for the physical scientist. Phys Biol 1(1–2):T1–10
68. Footer MJ, Kerssemakers JWJ, Theriot JA, Dogterom M (2007) Direct measurement of force generation by actin filament polymerization using an optical trap. PNAS 104(7):2181–2186
69. Forterre Y, Skotheim JM, Dumais J, Mahadevan L (2005) How the venus flytrap snaps. Nature 433(7024):421–425
70. Frey E, Täuber UC, Cauber U (1994) 2-loop renormalization-group analysis of the Burgers-Kardar-Parisi-Zhang equation. Phys Rev E 50(2):1024–1045
71. Frischknecht F, Way M (2001) Surfing pathogens and the lessons learned for actin polymerization. Trends Cell Biol 11(1):30–38
72. Fujita H, Ishiwata S (1998) Spontaneous oscillatory contraction without regulatory proteins in actin filament-reconstituted fibers. Biophys J 75(3):1439–1445
73. Fukui Y (2002) Mechanistics of amoeboid locomotion: signal to forces. Cell Biol Int 26(11):933–944
74. Fygenson DK, Marko JF, Libchaber A (1997) Mechanics of microtubule-based membrane extension. Phys Rev Lett 79(22):4497–4500
75. Gardel ML, Shin JH, MacKintosh FC, Mahadevan L, Matsudaira PA, Weitz DA (2004) Elastic behavior of cross-linked and bundled actin networks. Science 304(5675):1301–1305
76. Gardel ML, Shin JH, MacKintosh FC, Mahadevan L, Matsudaira PA, Weitz DA (2004) Scaling of F-actin network rheology to probe single filament elasticity and dynamics. Phys Rev Lett 93(18):188102
77. Gardel ML, Nakamura F, Hartwig JH, Crocker JC, Stossel TP, Weitz DA (2006) Prestressed F-actin networks cross-linked by hinged filamins replicate mechanical properties of cells. Proc Natl Acad Sci USA 103(6):1762–1767
78. Geiger B, Bershadsky A, Pankov R, Yamada KM (2001) Transmembrane crosstalk between the extracellular matrix–cytoskeleton crosstalk. Nat Rev Mol Cell Biol 2(11):793–805
79. Gerbal F, Chaikin P, Rabin Y, Prost J (2000) An elastic analysis of Listeria monocytogenes propulsion. Biophys J 79(5): 2259–2275
80. Giardini PA, Fletcher DA, Theriot JA (2003) Compression forces generated by actin comet tails on lipid vesicles. Proc Natl Acad Sci USA 100(11):6493–6498
81. Gittes F, Mickey B, Nettleton J, Howard J (1993) Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. J Cell Biol 120(4):923–934
82. Glotzer M (2001) Animal cell cytokinesis. Annu Rev Cell Dev Biol 17:351–386
83. Goldberg MB (2001) Actin-based motility of intracellular microbial pathogens. Microbiol Mol Biol Rev 65(4):595–626
84. Granger C, Cyr R (2001) Use of abnormal preprophase bands to decipher division plane determination. J Cell Sci 114(Pt 3): 599–607
85. Grill SW, Gonczy P, Stelzer EH, Hyman AA (2001) Polarity controls forces governing asymmetric spindle positioning in the caenorhabditis elegans embryo. Nature 409(6820):630–633
86. Grill SW, Kruse K, Jülicher F (2005) Theory of mitotic spindle oscillations. Phys Rev Lett 94(10):108104
87. Gruenheid S, Finlay BB (2003) Microbial pathogenesis and cytoskeletal function. Nature 422(6933):775–781
88. Gueron S, Levit-Gurevich K, Liron N, Blum JJ (1997)Cilia internal mechanism and metachronal coordination as the result of hydrodynamical coupling. Proc Natl Acad Sci USA 94(12):6001–6006
89. Guirao B, Joanny J-F (2007) Spontaneous creation of macroscopic flow and metachronal waves in an array of cilia. Biophys J 92(6):1900–1917
90. Gupton SL, Waterman-Storer CM (2006) Spatiotemporal feedback between actomyosin and focal-adhesion systems optimizes rapid cell migration. Cell 125(7):1361–1374
91. Hall A (2005) Rho GTPases and the control of cell behaviour. Biochem Soc Trans 33(Pt 5):891–895
92. Hanney T, Evans MR (2004) Condensation transitions in a two-species zero-range process. Phys Rev E Stat Nonlin Soft Matter Phys 69(1 Pt 2):016107
93. Hanson J, Huxley HE (1953) Structural basis of the cross-striations in muscle. Nature 172(4377):530–532
94. Hatwalne Y, Ramaswamy S, Rao M, Simha RA (2004) Rheology of active-particle suspensions. Phys Rev Lett 92(11):118101
95. Hayden JH, Bowser SS, Rieder CL (1990) Kinetochores capture astral microtubules during chromosome attachment to

the mitotic spindle: Direct visualization in live newt lung cells. J Cell Biol 111(3):1039–1045

96. Heintzelman MB (2003) Gliding motility: The molecules behind the motion. Curr Biol 13(2):R57–9
97. Hill TL (1974) Theoretical formalism for the sliding filament model of contraction of striated muscle, part I. Prog Biophys Mol Biol 28:267–340
98. Hinrichsen H (2006) Non-equilibrium phase transitions. Physica A 369(1):1–28
99. Hoffman BD, Massiera G, Van Citters KM, Crocker JC (2006) The consensus mechanics of cultured mammalian cells. Proc Natl Acad Sci USA 103(27):10259–10264
100. Hohenberg PC, Halperin BI (1977) Theory of dynamic critical phenomena. Rev Mod Phys 49(3):435–479
101. Holy TE, Dogterom M, Yurke B, Leibler S (1997) Assembly and positioning of microtubule asters in microfabricated chambers. Proc Natl Acad Sci USA 94(12):6228–6231
102. Howard J (1997) Molecular motors: Structural adaptations to cellular functions. Nature 389(6651):561–567
103. Howard J (2001) Mechanics of Motor Proteins and the Cytoskeleton. Sinauer Associates Inc, Sunderland
104. Howard J, Hyman AA (2003) Dynamics and mechanics of the microtubule plus end. Nature 422(6933):753–758
105. Howard M, Rutenberg AD (2003) Pattern formation inside bacteria: Fluctuations due to the low copy number of proteins. Phys Rev Lett 90(12):128102
106. Humphrey D, Duggan C, Saha D, Smith D, Käs J (2002) Active fluidization of polymer networks through molecular motors. Nature 416(6879):413–416
107. Hunter AW, Caplow M, Coy DL, Hancock WO, Diez S, Wordeman L, Howard J (2003) The kinesin-related protein MCAK is a microtubule depolymerase that forms an ATP-hydrolyzing complex at microtubule ends. Mol Cell 11(2):445–457
108. Huxley AF (1957) Muscle structure and theories of contraction. Prog Biophys Biophys Chem 7:255–318
109. Huxley AF, Niedergerke R (1954) Structural changes in muscle during contraction; interference microscopy of living muscle fibres. Nature 173(4412):971–973
110. Huxley HE (1953) Electron microscope studies of the organisation of the filaments in striated muscle. Biochim Biophys Acta 12(3):387–394
111. Huxley HE (1953) X-ray analysis and the problem of muscle. Proc R Soc Lond B Biol Sci 141(902):59–62
112. Huxley HE (1957) The double array of filaments in cross-striated muscle. J Biophys Biochem Cytol 3(5):631–648
113. Huxley HE (1996) A personal view of muscle and motility mechanisms. Annu Rev Physiol 58:1–19
114. Huxley HE (2004) Fifty years of muscle and the sliding filament hypothesis. Eur J Biochem 271(8):1403–1415
115. Huxley HE, Hanson J (1954) Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. Nature 173(4412):973–976
116. Hyman AA, Karsenti E (1996) Morphogenetic properties of microtubules and mitotic spindle assembly. Cell 84(3):401–410
117. Ibanez-Tallon I, Heintz N, Omran H (2003) To beat or not to beat: roles of cilia in development and disease. Hum Mol Genet 12(Spec no 1):R27–35
118. Inoue S, Salmon ED (1995) Force generation by microtubule assembly/disassembly in mitosis and related movements. Mol Biol Cell 6(12):1619–1640
119. Isambert H, Venier P, Maggs AC, Fattoum A, Kassab R, Pantaloni D, Carlier MF (1995) Flexibility of actin filaments derived from thermal fluctuations. Effect of bound nucleotide, phalloidin, and muscle regulatory proteins. J Biol Chem 270(19):11437–11444 (1995)
120. Italiano JE Jr, Roberts TM, Stewart M, Fontana CA (1996) Reconstitution in vitro of the motile apparatus from the amoeboid sperm of ascaris shows that filament assembly and bundling move membranes. Cell 84(1):105–114
121. Italiano JE Jr, Stewart M, Roberts TM (1999) Localized depolymerization of the major sperm protein cytoskeleton correlates with the forward movement of the cell body in the amoeboid movement of nematode sperm. J Cell Biol 146(5):1087–1096
122. Janmey PA, Euteneuer U, Traub P, Schliwa M (1991) Viscoelastic properties of vimentin compared with other filamentous biopolymer networks. J Cell Biol 113(1):155–160
123. Janson ME, Dogterom M (2004) Scaling of microtubule force-velocity curves obtained at different tubulin concentrations. Phys Rev Lett 92(24):248101
124. Joanny J-F, Jülicher F, Prost J (2003) Motion of an adhesive gel in a swelling gradient: A mechanism for cell locomotion. Phys Rev Lett 90(16):168102
125. Joanny J-F, Jülicher F, Kruse K, Prost J (2007) Hydrodynamic theory for multi-component active polar gels. New J Phys 9:422
126. Jones LJ, Carballido-Lopez R, Errington J (2001) Control of cell shape in bacteria: Helical, actin-like filaments in bacillus subtilis. Cell 104(6):913–922
127. Jülicher F, Prost J (1995) Cooperative molecular motors. Phys Rev Lett 75(13):2618–2621
128. Jülicher F, Prost J (1997) Spontaneous oscillations of collective molecular motors. Phys Rev Lett 78(23):4510–4513
129. Jülicher F, Ajdari A, Prost J (1997) Modeling molecular motors. Rev Mod Phys 69(4):1269–1281
130. Jülicher F, Kruse K, Prost J, Joanny JF (2007) Active behavior of the cytoskeleton. Phys Rep 449(1–3):3–28
131. Jurado C, Haserick JR, Lee J (2005) Slipping or gripping? fluorescent speckle microscopy in fish keratocytes reveals two different mechanisms for generating a retrograde flow of actin. Mol Biol Cell 16(2):507–518
132. Kaksonen M, Sun Y, Drubin DG (2003) A pathway for association of receptors, adaptors, and actin during endocytic internalization. Cell 115(4):475–487
133. Karki S, Holzbaur EL (1999) Cytoplasmic dynein and dynactin in cell division and intracellular transport. Curr Opin Cell Biol 11(1):45–53
134. Karsenti E, Vernos I (2001) The mitotic spindle: A self-made machine. Science 294(5542):543–547
135. Kasza KE, Rowat AC, Liu J, Angelini TE, Brangwynne CP, Koenderink GH, Weitz DA (2007) The cell as a material. Curr Opin Cell Biol 19(1):101–107
136. Kaverina I, Krylyshkina O, Small JV (2002) Regulation of substrate adhesion dynamics during cell motility. Int J Biochem Cell Biol 34(7):746–761
137. King KL, Essig J, Roberts TM, Moerland TS (1994) Regulation of the ascaris major sperm protein (MSP) cytoskeleton by intracellular pH. Cell Motil Cytoskeleton 27(3):193–205
138. Klumpp S, Lipowsky R (2004) Asymmetric simple exclusion processes with diffusive bottlenecks. Phys Rev E Stat Nonlin Soft Matter Phys 70(6 Pt 2):066104

139. Klumpp S, Lipowsky R (2004) Phase transitions in systems with two species of molecular motors. Europhys Lett 66(1):90–96
140. Kolomeisky AB, Fisher ME (2001) Force-velocity relation for growing microtubules. Biophys J 80(1):149–154
141. Kondo S (2002) The reaction-diffusion system: A mechanism for autonomous pattern formation in the animal skin. Genes Cells 7(6):535–541
142. Kozlov MM, Mogilner A (2007) Model of polarization and bistability of cell fragments. Biophys J 93(11):3811–3819
143. Krendel M, Zenke FT, Bokoch GM (2002) Nucleotide exchange factor GEF-H1 mediates cross-talk between microtubules and the actin cytoskeleton. Nat Cell Biol 4(4):294–301
144. Kruse K, Jülicher F (2000) Actively contracting bundles of polar filaments. Phys Rev Lett 85(8):1778–1781
145. Kruse K, Jülicher F (2003) Self-organization and mechanical properties of active filament bundles. Phys Rev E Stat Nonlin Soft Matter Phys 67(5 Pt 1):051913
146. Kruse K, Jülicher F (2006) Dynamics and mechanics of motor-filament systems. Eur Phys J E Soft Matter 20(4):459–465
147. Kruse K, Sekimoto K (2002) Growth of fingerlike protrusions driven by molecular motors. Phys Rev E Stat Nonlin Soft Matter Phys 66(3 Pt 1):031904
148. Kruse K, Camalet S, Jülicher F (2001) Self-propagating patterns in active filament bundles. Phys Rev Lett 87(13):138101
149. Kruse K, Zumdieck A, Jülicher F (2003) Continuum theory of contractile fibres. Europhys Lett 64(5):716–722
150. Kruse K, Joanny JF, Jülicher F, Prost J, Sekimoto K (2004) Asters, vortices, and rotating spirals in active gels of polar filaments. Phys Rev Lett 92(7):078101
151. Kruse K, Joanny JF, Jülicher F, Prost J, Sekimoto K (2005) Generic theory of active polar gels: A paradigm for cytoskeletal dynamics. Eur Phys J E Soft Matter 16(1):5–16
152. Kruse K, Joanny JF, Jülicher F, Prost J (2006) Contractility and retrograde flow in lamellipodium motion. Phys Biol 3(2): 130–137
153. Lacayo CI, Theriot JA (2004) Listeria monocytogenes actin-based motility varies depending on subcellular location: A kinematic probe for cytoarchitecture. Mol Biol Cell 15(5):2164–2175
154. Landau LD, Lifschitz EM (1995) The theory of elasticity. Butterworth-Heinemann, Boston
155. Lasa I, Gouin E, Goethals M, Vancompernolle K, David V, Vandekerckhove J, Cossart P (1997) Identification of two regions in the N-terminal domain of ActA involved in the actin comet tail formation by Listeria monocytogenes. EMBOJ 16(7): 1531–1540
156. Lauffenburger DA, Horwitz AF (1996) Cell migration: A physically integrated molecular process. Cell 84(3):359–369
157. Lee HY, Kardar M (2001) Macroscopic equations for pattern formation in mixtures of microtubules and molecular motors. Phys Rev E Stat Nonlin Soft Matter Phys 64(5 Pt 2):056113
158. Lee J, Ishihara A, Jacobson K (1993) The fish epidermal keratocyte as a model system for the study of cell locomotion. Symp Soc Exp Biol 47:73–89
159. Leibler S, Huse DA (1993) Porters versus rowers: A unified stochastic model of motor proteins. J Cell Biol 121(6): 1357–1368
160. Lin CH, Espreafico EM, Mooseker MS, Forscher P (1996) Myosin drives retrograde F-actin flow in neuronal growth cones. Neuron 16(4):769–782
161. Lipowsky R, Klumpp S, Nieuwenhuizen TM (2001) Random walks of cytoskeletal motors in open and closed compartments. Phys Rev Lett 87(10):108101
162. Liverpool TB (2003) Anomalous fluctuations of active polar filaments. Phys Rev E Stat Nonlin Soft Matter Phys 67(3 Pt 1):031909
163. Liverpool TB (2006) Active gels: Where polymer physics meets cytoskeletal dynamics. Philos Trans A Math Phys Eng Sci 364(1849):3335–3355
164. Liverpool TB, Marchetti MC (2003) Instabilities of isotropic solutions of active polar filaments. Phys Rev Lett 90(13):138102
165. Liverpool TB, Marchetti MC (2005) Bridging the microscopic and the hydrodynamic in active filament solutions. Europhys Lett 69(5):846–852
166. Liverpool TB, Marchetti MC (2006) Rheology of active filament solutions. Phys Rev Lett 97(26):268101
167. Liverpool TB, Marchetti MC (2008) Hydrodynamics and rheology of active polar filaments. In: Lenz P (ed) Cell motility. Springer, New York
168. Liverpool TB, Maggs AC, Ajdari A (2001) Viscoelasticity of solutions of motile polymers. Phys Rev Lett 86(18):4171–4174
169. Lo CM, Wang HB, Dembo M, Wang YL (2000) Cell movement is guided by the rigidity of the substrate. Biophys J 79(1): 144–152
170. Loisel TP, Boujemaa R, Pantaloni D, Carlier MF (1999) Reconstitution of actin-based motility of Listeria and Shigella using pure proteins. Nature 401(6753):613–616
171. Lowe J, Amos LA (1998) Crystal structure of the bacterial cell-division protein FtsZ. Nature 391(6663):203–206
172. Lymn RW, Taylor EW (1971) Mechanism of adenosine triphosphate hydrolysis by actomyosin. Biochemistry 10(25): 4617–4624
173. MacDonald CT, Gibbs JH, Pipkin AC (1968) Kinetics of biopolymerization on nucleic acid templates. Biopolymers 6(1):1–5
174. Machesky LM, Insall RH (1999) Signaling to actin dynamics. J Cell Biol 146(2):267–272
175. Machesky LM, Mullins RD, Higgs HN, Kaiser DA, Blanchoin L, May RC, Hall ME, Pollard TD (1999) Scar, a WASp-related protein, activates nucleation of actin filaments by the Arp2/3 complex. Proc Natl Acad Sci USA 96(7):3739–3744
176. Machin KE (1958) Wave propagation along flagella. J Exp Biol 35:796
177. Magnasco MO (1993) Forced thermal ratchets. Phys Rev Lett 71(10):1477–1481
178. Magnasco MO (1994) Molecular combustion motors. Phys Rev Lett 72(16):2656–2659
179. Mahadevan L, Matsudaira P (2000) Motility powered by supramolecular springs and ratchets. Science 288(5463): 95–100
180. Maney T, Wagenbach M, Wordeman L (2001) Molecular dissection of the microtubule depolymerizing activity of mitotic centromere-associated kinesin. J Biol Chem 276(37): 34753–34758
181. Manson MD, Armitage JP, Hoch JA, Macnab RM (1998) Bacterial locomotion and signal transduction. J Bacteriol 180(5):1009–1022
182. Marcy Y, Prost J, Carlier M-F, Sykes C (2004) Forces generated during actin-based propulsion: A direct measurement by micromanipulation. Proc Natl Acad Sci USA 101(16):5992–5997

183. Martin PC, Parodi O, Pershan P (1972) Unified hydrodynamic theory for crystals, liquid crystals and normal fluids. Phys Rev A 6:2401
184. McGrath JL, Eungdamrong NJ, Fisher CI, Peng F, Mahadevan L, Mitchison TJ, Kuo SC (2003) The force-velocity relationship for the actin-based motility of Listeria monocytogenes. Curr Biol 13(4):329–332
185. McGrath JL, Somlo S, Makova S, Tian X, Brueckner M (2003) Two populations of node monocilia initiate left-right asymmetry in the mouse. Cell 114(1):61–73
186. Merrifield CJ, Moss SE, Ballestrem C, Imhof BA, Giese G, Wunderlich I, Almers W (1999) Endocytic vesicles move at the tips of actin tails in cultured mast cells. Nat Cell Biol 1(1):72–74
187. Miao L, Vanderlinde O, Stewart M, Roberts TM (2003) Retraction in amoeboid cell motility powered by cytoskeletal dynamics. Science 302(5649):1405–1407
188. Michie KA, Lowe J (2006) Dynamic filaments of the bacterial cytoskeleton. Annu Rev Biochem 75:467–492
189. Miki H, Setou M, Kaneshiro K, Hirokawa N (2001) All kinesin superfamily protein, KIF, genes in mouse and human. Proc Natl Acad Sci USA 98(13):7004–7011
190. Miranti CK, Brugge JS (2002) Sensing the environment: A historical perspective on integrin signal transduction. Nat Cell Biol 4(4):E83–90
191. Mitchison TJ, Kirschner M (1984) Dynamic instability of microtubule growth. Nature 312(5991):237–242
192. Mitchison TJ, Cramer LP (1996) Actin-based cell motility and cell locomotion. Cell 84(3):371–379
193. Mitchison TJ, Salmon ED (2001) Mitosis: A history of division. Nat Cell Biol 3(1):E17–21
194. Mogilner A (2006) On the edge: modeling protrusion. Curr Opin Cell Biol 18(1):32–39
195. Mogilner A, Oster G (1996) Cell motility driven by actin polymerization. Biophys J 71(6):3030–3045
196. Mogilner A, Oster G (2003) Cell biology. Shrinking gels pull cells. Science 302(5649):1340–1341
197. Mogilner A, Oster G (2003) Force generation by actin polymerization II: The elastic ratchet and tethered filaments. Biophys J 84(3):1591–1605
198. Mogilner A, Oster G (2003) Polymer motors: Pushing out the front and pulling up the back. Curr Biol 13(18):R721–33
199. Mohanty PK, Kruse K (2007) Driven diffusive systems of active filament bundles. J Stat Phys 128(1–2):95–110
200. Mukherjee A, Dai K, Lutkenhaus J (1993) Escherichia coli cell division protein FtsZ is a guanine nucleotide binding protein. Proc Natl Acad Sci USA 90(3):1053–1057
201. Nakazawa H, Sekimoto K (1996) Polarity sorting in a bundle of actin filaments by two-headed myosins. J Phys Soc Jpn 65(8):2404–2407
202. Nedelec F, Surrey T, Maggs AC (2001) Dynamic concentration of motors in microtubule arrays. Phys Rev Lett 86(14):3192–3195
203. Nedelec FJ, Surrey T, Maggs AC, Leibler S (1997) Self-organization of microtubules and motors. Nature 389(6648):305–308
204. Nelson WJ (2003) Adaptation of core mechanisms to generate cell polarity. Nature 422(6933):766–774
205. Nieuwenhuizen TM, Klumpp S, Lipowsky R (2004) Random walks of molecular motors arising from diffusional encounters with immobilized filaments. Phys Rev E 69(6):061911
206. Nogales E, Wolf SG, Downing KH (1998) Structure of the alpha beta tubulin dimer by electron crystallography. Nature 391(6663):199–203
207. Noireaux V, Golsteyn RM, Friederich E, Prost J, Antony C, Louvard D, Sykes C (2000) Growing an actin gel on spherical surfaces. Biophys J 78(3):1643–1654
208. Okada Y, Takeda S, Tanaka Y, Izpisua Belmonte J-C, Hirokawa N (2005) Mechanism of nodal flow: A conserved symmetry breaking event in left-right axis determination. Cell 121(4):633–644
209. Okamura N, Ishiwata S (1988) Spontaneous oscillatory contraction of sarcomeres in skeletal myofibrils. J Muscle Res Cell Motil 9(2):111–119
210. Ott A, Magnasco M, Simon A, Libchaber A (1993) Measurement of the persistence length of polymerized actin using fluorescence microscopy. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 48(3):R1642–R1645
211. Paluch E, Sykes C, Prost J, Bornens M (2006) Dynamic modes of the cortical actomyosin gel during cell locomotion and division. Trends Cell Biol 16(1):5–10
212. Paluch E, van der Gucht J, Joanny J-F, Sykes C (2006) Deformations in actin comets from rocketing beads. Biophys J 91(8):3113–3122
213. Pampaloni F, Lattanzi G, Jonas A, Surrey T, Frey E, Florin E-L (2006) Thermal fluctuations of grafted microtubules provide evidence of a length-dependent persistence length. Proc Natl Acad Sci USA 103(27):10248–10253
214. Pantaloni D, Le Clainche C, Carlier MF (2001) Mechanism of actin-based motility. Science 292(5521):1502–1506
215. Parent CA, Devreotes PN (1999) A cell's sense of direction. Science 284(5415):765–770
216. Parmeggiani A, Franosch T, Frey E (2003) Phase coexistence in driven one-dimensional transport. Phys Rev Lett 90(8):086601
217. Parmeggiani A, Jülicher F, Ajdari A, Prost J (1999) Energy transduction of isothermal ratchets: Generic aspects and specific examples close to and far from equilibrium. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 60(2 Pt B):2127–2140
218. Peskin CS, Ermentrout GB, Oster GF (1994) Cell mechanics and cellular engineering. Springer, New York
219. Peskin CS, Odell GM, Oster GF (1993) Cellular motions and thermal fluctuations: the brownian ratchet. Biophys J 65(1):316–324
220. Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization – A unified approach to nonlinear science. Cambridge University Press, Cambridge
221. Plastino J, Lelidis I, Prost J, Sykes C (2004) The effect of diffusion, depolymerization and nucleation promoting factors on actin gel growth. Eur Biophys J 33(4):310–320
222. Plastino J, Sykes C (2005) The actin slingshot. Curr Opin Cell Biol 17(1):62–66
223. Pollard TD (2003) The cytoskeleton, cellular motility and the reductionist agenda. Nature 422(6933):741–5
224. Pollard TD, Blanchoin L, Mullins RD (2000) Molecular mechanisms controlling actin filament dynamics in nonmuscle cells. Annu Rev Biophys Biomol Struct 29:545–576
225. Pollard TD, Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. Cell 112(4):453–465
226. Porter ME (1996) Axonemal dyneins: Assembly, organization, and regulation. Curr Opin Cell Biol 8(1):10–17

227. Prost J, Chauwin JF, Peliti L, Ajdari A (1994) Asymmetric pumping of particles. Phys Rev Lett 72(16):2652–2655

228. Purcell EM (1977) Life at low reynolds number. Am J Phys 45:3

229. Raftopoulou M, Hall A (2004) Cell migration: Rho GTPases lead the way. Dev Biol 265(1):23–32

230. Ramaswamy S, Simha RA, Toner J (2003) Active nematics on a substrate: Giant number fluctuations and long-time tails. Europhys Lett 62(2):196–202 Apr

231. Ramaswamy S, Toner J, Prost J (2000) Nonequilibrium fluctuations, traveling waves, and instabilities in active membranes. Phys Rev Lett 84(15):3494–3497

232. Raychaudhuri D, Park JT (1992) Escherichia coli cell-division gene FtsZ encodes a novel GTP-binding protein. Nature 359(6392):251–254

233. Reimann P (2002) Brownian motors: Noisy transport far from equilibrium. Phys Rep 361:57–265

234. Revenu C, Athman R, Robine S, Louvard D (2004) The co-workers of actin filaments: From cell structures to signals. Nat Rev Mol Cell Biol 5(8):635–646

235. Ridley AJ (2001) Rho family proteins: Coordinating cell responses. Trends Cell Biol 11(12):471–477

236. Ridley AJ, Schwartz MA, Burridge K, Firtel RA, Ginsberg MH, Borisy G, Parsons JT, Horwitz AR (2003) Cell migration: Integrating signals from front to back. Science 302(5651): 1704–1709

237. Rieder CL, Khodjakov A (2003) Mitosis through the microscope: advances in seeing inside live dividing cells. Science 300(5616):91–96

238. Risler T, Prost J, Jülicher F (2004) Universal critical behavior of noisy coupled oscillators. Phys Rev Lett 93(17):175702

239. Riveline D, Ott A, Jülicher F, Winkelmann DA, Cardoso O, Lacapere JJ, Magnusdottir S, Viovy JL, Gorre-Talini L, Prost J (1998) Acting on actin: The electric motility assay. Eur Biophys J 27(4):403–408

240. Roberts TM, Stewart M (2000) Acting like actin. The dynamics of the nematode major sperm protein (MSP) cytoskeleton indicate a push-pull mechanism for amoeboid cell motility. J Cell Biol 149(1):7–12

241. Rogers SL, Gelfand VI (2000) Membrane trafficking, organelle transport, and the cytoskeleton. Curr Opin Cell Biol 12(1): 57–62

242. Rubinstein B, Jacobson K, Mogilner A (2005) Multiscale two-dimensional modeling of a motile simple-shaped cell. Multiscale Model Simul 3(2):413–439

243. Saez A, Ghibaudo M, Buguin A, Silberzan P, Ladoux B (2007) Rigidity-driven growth and migration of epithelial cells on microstructured anisotropic substrates. Proc Natl Acad Sci USA 104(20):8281–8286

244. Sammak PJ, Borisy GG (1988) Direct observation of microtubule dynamics in living cells. Nature 332(6166):724–726

245. Sankararaman S, Menon GI, Sunil Kumar PB (2004) Self-organized pattern formation in motor-microtubule mixtures. Phys Rev E Stat Nonlin Soft Matter Phys 70(3 Pt 1):031905

246. Schliwa M, Woehlke G (2003) Molecular motors. Nature 422(6933):759–765

247. Schmittmann B, Zia RKP (1995) Statistical mechanics of driven diffusive systems. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 17. Academic Press, London

248. Scholey JM, Brust-Mascher I, Mogilner A (2003) Cell division. Nature 422(6933):746–752

249. Schuyler SC, Pellman D (2001) Microtubule "plus-end-tracking proteins": The end is just the beginning. Cell 105(4): 421–424

250. Schütz GM (2001) Exactly solvable models for many-body systems. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 19. Academic Press, London

251. Sekimoto K, Nakazawa H (1998) Contraction of a bundle of actin filaments: 50 years after Szent-Gyorgyi, vol 1. World Scientific, Singapore, p 394

252. Sekimoto K, Prost J, Jülicher F, Boukellal H, Bernheim-Grosswasser A (2004) Role of tensile stress in actin gels and a symmetry-breaking instability. Eur Phys J E Soft Matter 13(3): 247–259

253. Shih Y-L, Rothfield L (2006) The bacterial cytoskeleton. Microbiol Mol Biol Rev 70(3):729–754

254. Simha RA, Ramaswamy S (2002) Hydrodynamic fluctuations and instabilities in ordered suspensions of self-propelled particles. Phys Rev Lett 89(5):058101

255. Small JV, Geiger B, Kaverina I, Bershadsky A (2002) How do microtubules guide migrating cells? Nat Rev Mol Cell Biol 3(12):957–964

256. Small JV, Kaverina I (2003) Microtubules meet substrate adhesions to arrange cell polarity. Curr Opin Cell Biol 15(1): 40–47

257. Small JV, Stradal T, Vignal E, Rottner K (2002) The lamellipodium: Where motility begins. Trends Cell Biol 12(3): 112–120

258. Spudich JA (1994) How molecular motors work. Nature 372(6506):515–518

259. Storm C, Pastore JJ, MacKintosh FC, Lubensky TC, Janmey PA (2005) Nonlinear elasticity in biological gels. Nature 435(7039):191–194

260. Surrey T, Nedelec F, Leibler S, Karsenti E (2001) Physical properties determining self-organization of motors and microtubules. Science 292(5519):1167–1171

261. Täuber UC, Howard M, Vollmayr-Lee BP (2005) Applications of field-theoretic renormalization group methods to reaction-diffusion problems. J Phys A 38(17):R79–R131

262. Taunton J, Rowning BA, Coughlin ML, Wu M, Moon RT, Mitchison TJ, Larabell CA (2000) Actin-dependent propulsion of endosomes and lysosomes by recruitment of N-WASp. J Cell Biol 148(3):519–530

263. Taylor GI (1951) Analysis of the swimming of microscopic organisms. Proc R Soc A 209:447

264. Theriot JA (2000) The polymerization motor. Traffic 1(1): 19–28

265. Theriot JA, Mitchison TJ, Tilney LG, Portnoy DA (1992) The rate of actin-based motility of intracellular Listeria monocytogenes equals the rate of actin polymerization. Nature 357(6375):257–260

266. Tilney LG, Portnoy DA (1989) Actin filaments and the growth, movement, and spread of the intracellular bacterial parasite, Listeria monocytogenes. J Cell Biol 109(4 Pt 1):1597–1608

267. Toner J, Tu YH (1998) Flocks, herds, and schools: A quantitative theory of flocking. Phys Rev E 58(4):4828–4858

268. Tran PT, Marsh L, Doye V, Inoue S, Chang F (2001) A mechanism for nuclear positioning in fission yeast based on microtubule pushing. J Cell Biol 153(2):397–411

269. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans Roy Soc (Lond) 237:37

270. Upadhyaya A, Chabot JR, Andreeva A, Samadani A, van Oudenaarden A (2003) Probing polymerization forces by using actin-propelled lipid vesicles. Proc Natl Acad Sci USA 100(8):4521–4526
271. Upadhyaya A, van Oudenaarden A (2003) Biomimetic systems for studying actin-based motility. Curr Biol 13(18): R734–744
272. Urrutia R, McNiven MA, Albanesi JP, Murphy DB, Kachar B (1991) Purified kinesin promotes vesicle motility and induces active sliding between microtubules in vitro. Proc Natl Acad Sci USA 88(15):6701–6705
273. Vale RD, Milligan RA (2000) The way things move: Looking under the hood of molecular motor proteins. Science 288(5463):88–95
274. Vale RD, Reese TS, Sheetz MP (1985) Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. Cell 42(1):39–50
275. Vallotton P, Danuser G, Bohnet S, Meister J-J, Verkhovsky AB (2005) Tracking retrograde flow in keratocytes: News from the front. Mol Biol Cell 16(3):1223–1231
276. Van den Ent F, Amos LA, Lowe J (2001) Prokaryotic origin of the actin cytoskeleton. Nature 413(6851):39–44
277. Verkhovsky AB, Svitkina TM, Borisy GG (1999) Self-polarization and directional motility of cytoplasm. Curr Biol 9(1): 11–20
278. Vernon GG, Woolley DM (2004) Basal sliding and the mechanics of oscillation in a mammalian sperm flagellum. Biophys J 87(6):3934–3944
279. Vilfan A, Frey E, Schwabl F (1998) Elastically coupled molecular motors. Eur Phys J B 3(4):535–546
280. Vilfan A, Frey E, Schwabl F (1999) Force-velocity relations of a two-state crossbridge model for molecular motors. Europhys Lett 45(3):283–289
281. Vilfan A, Jülicher F (2006) Hydrodynamic flow patterns and synchronization of beating cilia. Phys Rev Lett 96(5):058102
282. Voituriez R, Joanny JF, Prost J (2006) Generic phase diagram of active polar films. Phys Rev Lett 96(2):028102
283. Waterman-Storer CM, Worthylake RA, Liu BP, Burridge K, Salmon ED (1999) Microtubule growth activates Rac1 to promote lamellipodial protrusion in fibroblasts. Nat Cell Biol 1(1):45–50
284. Welch MD, Rosenblatt J, Skoble J, Portnoy DA, Mitchison TJ (1998) Interaction of human Arp2/3 complex and the Listeria monocytogenes ActA protein in actin filament nucleation. Science 281(5373):105–108
285. Wiesner S, Helfer E, Didry D, Ducouret G, Lafuma F, Carlier M-F, Pantaloni D (2003) A biomimetic motility assay provides insight into the mechanism of actin-based motility. J Cell Biol 160(3):387–398
286. Wolgemuth CW, Miao L, Vanderlinde O, Roberts T, Oster G (2005) MSP dynamics drives nematode sperm locomotion. Biophys J 88(4):2462–2471
287. Yasuda K, Shindo Y, Ishiwata S (1996) Synchronous behavior of spontaneous oscillations of sarcomeres in skeletal myofibrils under isotonic conditions. Biophys J 70(4):1823–1829
288. Yeh E, Yang C, Chin E, Maddox P, Salmon ED, Lew DJ, Bloom K (2000) Dynamic positioning of mitotic spindles in yeast: Role of microtubule motors and cortical determinants. Mol Biol Cell 11(11):3949–3961
289. Yoshida K, Soldati T (2006) Dissection of amoeboid movement into two mechanically distinct modes. J Cell Sci 119(Pt 18):3833–3844
290. Yoshida M, Muneyuki E, Hisabori T (2001) ATP synthase – A marvellous rotary engine of the cell. Nat Rev Mol Cell Biol 2(9):669–677
291. Zhou FQ, Cohan CS (2001) Growth cone collapse through coincident loss of actin bundles and leading edge actin without actin depolymerization. J Cell Biol 153(5):1071–1084
292. Ziebert F, Zimmermann W (2004) Pattern formation driven by nematic ordering of assembling biopolymers. Phys Rev E 70(2):022902
293. Ziebert F, Zimmermann W (2005) Nonlinear competition between asters and stripes in filament-motor systems. Eur Phys JE 18(1):41–54
294. Zinn-Justin J (2002) Quantum field theory and critical phenomena, 4th edn. Oxford University Press, Oxford
295. Zumdieck A, Lagomarsino MC, Tanase C, Kruse K, Mulder B, Dogterom M, Jülicher F (2005) Continuum description of the cytoskeleton: ring formation in the cell cortex. Phys Rev Lett 95(25):258103

Books and Reviews

Alberts B et al (2002) The cytoskeleton. The mechanics of cell division. In: Gibbs S (ed) Molecular biology of the cell. Garland, New York

Bray D (2000) In: Day M (ed) Cell movements. Garland, New York

Howard J (2001) Mechanics of motor proteins and the cytoskeleton. Sinauer Associates Inc, Sunderland

Lenz P (ed) (2008) Cell motility. Biological and medical physics, biomedical engineering. Springer, New York

Reviews of special interest and that cover the subjects treated in this article can be found in the following references: [9,57,67,114,129,130,136,163,222,223,233,236,253,257].