Chapter 5
Selected Models for Dynamics of Research Organizations and Research Production

Dedicated to the memory of Anatoly Yablonsky. His studies on mathematical models of science contributed much to my interest in mathematical modeling of social and economic systems.

Abstract The understanding of dynamics of research organizations and research production is very important for their successful management. In the text below, selected deterministic and probability models of research dynamics are discussed. The idea of the selection is to cover mainly the areas of publications dynamics, citations dynamics, and aging of scientific information. From the class of deterministic models we discuss models connected to research publications (SI-model, Goffmann–Newill model, model of Price for growth of knowledge), deterministic model connected to dynamics of citations (nucleation model of growth dynamics of citations), deterministic models connected to research dynamics (logistic curve models, model of competition between systems of ideas, reproduction–transport equation model of evolution of scientific subfields), and a model of science as a component of the economic growth of a country. From the class of probability models we discuss a probability model connected to research publications (based on the Yule process), probability models connected to dynamics of citations (Poisson and mixed Poisson models, models of aging of scientific information (death stochastic process model and birth stochastic process model connected to Waring distribution)). The truncated Waring distribution and the multivariate Waring distribution are described, and a variational approach to scientific production is discussed. Several probability models of production/citation process (Paretian and Poisson distribution models of the h-index) as well as GIGP model distribution of bibliometric data are presented. A stochastic model of scientific productivity based on a master equation is described, and a probability model for the importance of the human factor in science is discussed. The chapter ends by providing information about some models and distributions connected to informetrics: limited dependent variable models for data analysis and the generalized Zipf distribution and its connection to the Waring distribution and Yule distribution.

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_5

5.1 Introductory Remarks

The interest in models of research dynamics and research production has increased greatly since the publication of the book *Little Science, Big Science* [1] by Derek de Solla Price in 1963, in which the first systematic approach to the structure of modern science was presented. One began to construct models for the growth of the scientific literature, and this growth was assumed to be exponential (for all of science) but could be also logistic or even linear for some scientific disciplines. In addition, models of aging and obsolescence of scientific information appeared [2–4]. At approximately the same time as Price, Goffman and Newill [5] developed an intellectual epidemics model of scientific communication. From the point of view of this model, the diffusion of ideas in a population of scientists could be compared to the spreading of a virus in some population, causing an epidemic. The model of Goffman and Newill was followed by other models that connected science dynamics to dynamics of populations. Several such models will be discussed below.

The number of models in the area of research dynamics grows continuously. There are many mathematical models connected to the dynamics of research organizations that may supply useful information for support of assessment of research production. The focus of this book is mainly on science dynamics and on results obtained by research on publications and citations. This focus limits the set of models for discussion and determines the selection of the models presented below. In principle, two kinds of models may be developed: deterministic models and probability models. The discussion below begins with models for dynamics of research publications. First of all, several forms of growth function are described. Then two deterministic models of a kind epidemic (SI model and the Goffman–Newill model) are presented. As an example of a deterministic nonepidemiological model, the Price model of knowledge growth is discussed. The nucleation model of Sangwal for citations dynamics follows, and this is the only deterministic model connected to citation dynamics. The reason for this limited coverage is as follows. A citation may be considered a unit of importance of scientific information. But this unit is small, and in addition, citations may arise more frequently than the larger units of scientific information (research publications). Finally, citations may arise quite irregularly. Thus more attention to citation dynamics is given from the point of view of probability models. The presentation of deterministic models continues with a model of competition of ideas, which is important for the evolution of research structures and systems. Further, the reproduction transport equation model of dynamics of scientific fields is discussed. The part devoted to deterministic models ends with a model of science as a component of the economic growth of a country.

The greater part of the chapter is devoted to probability models. This part begins with several general remarks on Poisson processes and their connection to the distributions of Yule and Waring and to the GIGP distribution. Then a probability model of research publications based on the Yule stochastic process is described. After that, attention is focused on models connected to citations of research publications. These models are for citation dynamics of a set of simultaneously appearing research pub-

lications and citation behavior of sets containing subsets of publications published at the same time. The discussion is based on the Poisson distribution and on the mixed Poisson distribution, which will be related to the Yule distribution. Models for aging of scientific information follow (the aging of information is an important topic connected to the dynamics of citations of research publications). Two probability models of the aging of scientific information are considered: a model based on a death stochastic process and a model based on a nonstationary birth process. The last model leads to the Waring distribution and to the negative binomial distribution. The Waring distribution is discussed in greater detail: the truncated Waring distribution and multivariate Waring distribution are described. On the basis of the truncated Waring distribution, a model of brain drain in the case of massive migration through migration channels is mentioned. A description of a variational approach to research production and two models of a production–citation process follows. The GIGP model distribution for bibliometric data is discussed. A master equation model of scientific productivity follows. The chapter ends with a probability model for the importance of the human factor in science.

5.2 Deterministic Models Connected to Research Publications

5.2.1 Simple Models. Logistic Curve and Other Models of Growth

One may consider simple exponential or logistic models of the growth of a number of items. For the case of the exponential model, the assumption is that the growth is proportional to the number of existing items,

$$\frac{dN}{dt} = kN, \tag{5.1}$$

where k is a parameter. The solution of (5.1) is $N(t) = N_0 \exp(kt)$, where N_0 is the number of available items at $t = 0$. It is of interest to know in many cases when the initial number of items N_0 will double. This time is $t^* = \ln(2)/k$ for the case of the exponential model. The exponential model, e.g., may be considered an approximation of the initial increase in the number of research publications in a newly established research field (more details follow below).

If we consider a longer time interval, then the initial exponential increase of the number of items may cease. In this case, one may consider another model, the logistic model of growth:

$$\frac{dN}{dt} = kN(a - N), \tag{5.2}$$

where k and a are (positive) parameters. The solution of the logistic equation (5.2) is

$$N = \frac{a}{1 + \left(\frac{a}{N_0} - 1\right) \exp(-kat)}. \tag{5.3}$$

This solutions has regions of almost exponential growth (when $N \ll a$, a region of almost linear growth around $N = a/2$, and a region of saturation (almost negative exponential growth) around $N \approx a$.

Logistic curves are frequently applied for modeling a variety of processes, e.g., the growth of scientific publications [6–10]. In order to describe trajectories of growth or decline in socio-technical systems, one generally uses the following three-parameter logistic curve [11]:

$$x(t) = \frac{K}{1 + \exp[-\alpha t - \beta]}, \tag{5.4}$$

where the quantities are as follows:

- $x(t)$: number of units in the species or growing variable to study,
- K: the asymptotic limit of growth,
- α: growth rate, which specifies the "width" of the curve for $x(t)$,
- β: specifies the time t_m when the curve reaches the midpoint of the growth trajectory such that $x(t_m) = 0.5\ K$.

The parameters K, α, and β are usually obtained after fitting the available data. It is well known that many cases of epidemic growth can be described by parts of an appropriate logistic curve. But not every interaction scheme leads to logistic growth [12]. The evolution of systems in such regimes may be described by more complex curves such as a combination of two or more simple three-parameter functions [11, 13].

Let us consider in more detail the logistic growth of knowledge and aging of scientific information. The appearance of the logistic curve in this case is a consequence of two processes: an increase in the amount of scientific information and the aging of scientific information. If only increasing of scientific information exists, then the increase may be proportional to the amount of the available information,

$$\frac{dx}{dt} = \alpha x \rightarrow x = x_0 \exp(\alpha t), \tag{5.5}$$

where α is a coefficient (the assumption is that each element produces a new element with a constant intensity α). This leads to exponential growth of scientific information. Such a situation can be observed for new areas of research in which the information is relatively new (and not aged). For more mature research areas, the coefficient α depends on the amount of information $x : \alpha = f(x)$ and decreases with

the aging of the scientific information. A simple assumption is that the decrease in α is proportional to x. Then

$$\frac{dx}{dt} = (a - bx)x. \tag{5.6}$$

Equation (5.6) is the logistic equation. Its solution is

$$x(t) = \frac{a}{b[1 + \sigma \exp(-at)]}, \tag{5.7}$$

where σ is a coefficient that can be determined from the initial conditions. From (5.7), it follows that the speed of the increase of scientific information is

$$\text{Eff} = \frac{dx}{dt} = \frac{\sigma a^2}{b} \frac{\exp(-at)}{\{1 + \exp[\sigma \exp(-at)]\}}. \tag{5.8}$$

The quantity Eff can be considered a measure of the effectiveness of the scientific field. This effectiveness (i) increases when the scientific field is new; (ii) passes through a maximum at $t = \ln(\sigma/a)$ (the maximum "expectation" of the scientific field; (iii) tends to 0 as $t \to \infty$ (the scientific field is exhausted).

In general, the growth can be described by the relationship

$$\frac{dx}{dt} = \alpha(x)x. \tag{5.9}$$

If we are interested in the growth around some value $x = x_0$, then we can represent $\alpha(x)$ by a Taylor series,

$$\alpha(x) = \alpha(x_0) + \frac{1}{1!}\frac{d\alpha}{dx}\Big|_{x=x_0}(x - x_0) + \frac{1}{2!}\frac{d^2\alpha}{dx^2}\Big|_{x=x_0}(x - x_0)^2 + \frac{1}{3!}\frac{d^3\alpha}{dx^3}\Big|_{x=x_0}(x - x_0)^3 \ldots. \tag{5.10}$$

If we use only the first term from (5.10), then the local growth around $x = x_0$ is exponential. If we have to use the first two terms in (5.10), then the local growth can be logistic. If we have to use the first three or more terms from (5.10), then the local growth is more complicated.

Logistic growth is not the only possible growth connected to the evolution of scientific information. The study of Menard [14] revealed three types of research fields with respect to the type of growth of the total number of publications in a given research field: stable fields (linear or exponential growth at small rates); exponentially growing fields (rapidly growing fields); cyclic fields: cyclic change of periods of stable and fast growth [15, 16]. Let us note the mathematical relationships for several kinds of growth functions that may be of interest to readers who encounter growth phenomena in their research:

1. *Gompertz growth function* [10]

$$x(t) = DA^{B^t}, \tag{5.11}$$

where $D > 0$ and $\log(A)\log(B) > 0$.

2. *Ware growth function* [17]

$$x(t) = \delta(1 - \varphi^{-t}), \tag{5.12}$$

where $\delta > 0$ and the constant φ is greater than 1.

3. *Power law growth function* [16]

$$x(t) = a + bt^{\gamma}, \tag{5.13}$$

where $a > 0$ and $b > 0$. For $0 < \gamma < 1$, the growth is concave and without an upper limit; for $\gamma = 1$, the growth is linear: for $\gamma > 1$, the growth is convex.

5.2.2 Epidemic Models

Below, we discuss two epidemic models of diffusion of knowledge by research publications. Epidemic models were used originally in population dynamics [18–24]. And for many years, most models of population dynamics were of interest only to biologists [25–30]. Today, these models are applied in many more areas of science [26–40]. For the area of research on scientific systems, the epidemic models are of great interest, too. This is so because some stages of processes by which ideas spread within a population, e.g., of scientists, has features that are like those of the spread of epidemics [41–43].

Epidemic models are a subclass of the more general class of Lotka–Volterra models [44–49] that are used in research on systems in the fields of biological population dynamics, social dynamics, economics, as well as for modeling processes connected to the spread of knowledge, ideas, and innovations [50–53].

The central concept of the epidemic models is the concept that scientific results spread to scientific communities by an epidemic diffusion process whereby more and more members of the scientific community are "infected" by the new scientific ideas and results. An important channel for spreading of this "infection" is research publications.

5.2.3 *Change in the Number of Publications in a Research Field. SI (Susceptibles–Infectives) Model of Change in The Number of Researchers Working in a Field*

Three basic classes of populations are important in epidemic research: [54]:

- **The susceptibles** S, who can become infectives on coming in contact with infectious material (the infectious material in our case is the scientific ideas).
- **The infectives** I who host the infectious material.
- **The recovered** R who are removed from the epidemic.

Because of this, the name of a class of epidemic models is the SIR-model (susceptibles–infectives–recovered (removed)). Nowakowska [55] discussed several discrete epidemic models for predicting changes in the number of publications in a given scientific field. The main assumption of the models is that the number of publications in the next period of time (say one year) will depend on the number of publications that have recently appeared and on the degree to which the subject has been exhausted. The behavior of the number of publications is considered to be as follows. The numbers of publications appearing in successive periods of time should first increase, then reach a maximum, and as the problem becomes more and more exhausted, the number of publications should decrease. A mathematical relationship that reflects such behavior was proposed by Daley [56]:

$$p_{t+1} = c_t p_t \left(N - \sum_{i=1}^{t} p_i \right), \tag{5.14}$$

where

- p_t: number of publications written in the period t;
- N: number of publications that have to appear in order to exhaust the problems in the research field.
- c_t: coefficient that can be connected to the number of researchers x_t working in the field: $c_t = 1 - (1 - d)^{x_t}$, where d is a parameter.

The epidemic part of the model is connected to the researchers who produce publications in the corresponding research field. There are researchers who produce publications in the field, and the number of these researchers may change. Some factors contribute to a decrease in the number of researchers (they retire or are no longer interested in the corresponding research problems). And there is a factor that contributes to an increase in the number of authors in the research field: new authors may begin to write publications (young researchers that begin their research career or researchers who became interested in the problems from the corresponding research field). We shall treat the last increase in the number of authors as infection and the entire process as an epidemic.

Let us assume that at a certain moment t, the epidemic's state is (x_t, y_t), where

- x_t is the number of infectives: authors who write publications in the corresponding scientific field;
- y_t is the number of susceptibles.

Then:

1. for a sufficiently short time interval Δt, one may expect that the number of infectives $x_{t+\Delta t}$ will be equal to $x_t - ax_t\Delta t + bx_t y_t \Delta t$,
2. while the number of susceptibles $y_{t+\Delta t}$ will be equal to $y_t - bx_t y_t \Delta t$ (a and b are suitable constants).

Let the expected number of individuals who either "die" or "recover" during the interval $(t, t + \Delta t)$, be $ax_t\Delta t$, and let $bx_t y_t \Delta t$ be the expected number of new infections. The equations of this model are

$$\begin{aligned} x_{t+\Delta t} &= x_t - ax_t\Delta t + bx_t y_t \Delta t, \\ y_{t+\Delta t} &= y_t - bx_t y_t \Delta t. \end{aligned} \tag{5.15}$$

The coefficients a and b may depend on the attractiveness of research field, on its being exhausted, etc. After setting appropriate relationships for a and b, one may investigate numerically the dynamics of the infectives x and susceptibles y, i.e., the dynamics of researchers producing publications in the corresponding research field.

5.2.4 *Goffman–Newill Continuous Model for the Dynamics of Populations of Scientists and Publications*

The model discussed above is an example of a discrete model. Now let us consider a continuous epidemic model connected with the dynamics of researchers and publications. Such a model is the Goffman–Newill model.

The Goffman–Newill model of intellectual epidemics is based on the Reed–Frost epidemic model [57–59], which was developed during the 1930s by Lowell Reed and Wade Frost, of the Johns Hopkins University. In the Reed–Frost model, one assumes a fixed population of size N. At each time, there is a certain number of cases of disease, C, and a certain number of susceptibles, S. One assumes that each case is infectious for a fixed length of time, and ignores the latent period: when individuals recover, one assumes that they are immune to further infection. During the infectious period of each case, one assumes that susceptibles may be infected and the disease may propagate further. The Goffman–Newill model [5, 60, 61] exploits the idea that the spreading of scientific ideas within a population of scientists can be studied on the basis of the publications of the members of that population. The main process in the model is the transfer of infectious materials (ideas) between humans by means of an intermediate host (a written article).

Let a scientific field be F and SF a subfield of F. We shall use the following notation: N_0, the number of scientists writing papers in the field F at t_0; I_0, the number of scientists writing papers in SF at t_0 (the number of infectives). Thus $S_0 = N_0 - I_0$ is the number of susceptibles; there is no removal (i.e., no scientists move out of the corresponding population) at t_0, but there is removal $R(t)$ at later times t. In addition, N_0' is the number of papers produced on F at t_0, and I_0' is the number of papers produced in SF at this time.

The process of intellectual infection takes place as follows:

1. A member of F is infected by a paper from I';
2. After some latency period, this infected member produces "infected" papers in N', i.e., the infected member produces a paper in the subfield SF citing a paper from I';
3. These'"infected" papers may infect other scientists from F and its subfields, such that the intellectual infection spreads from SF to the other subfields of F.

Let β be the rate at which the susceptibles from class S become "intellectually infected" from class I and let β' be the rate at which the papers in SF are cited by members of F who are producing papers in SF. As the infection process develops, some susceptibles and infectives are removed, i.e., some scientists are no longer active, and some papers are no longer cited. In addition, let γ and γ' be the rates of removal of infectives from the populations I and I' respectively, and let δ and δ' be the rates of removal from the populations of susceptibles S and S'. Moreover, there can be a supply of infectives and susceptibles in F and SF. Let the rates of introduction of new susceptibles be μ and μ' (these are the rates at which new authors and new papers are introduced in F) and let the rates of introduction of new infectives be υ and υ' (these are the rates at which new authors and new papers are introduced in SF). In addition, within a short interval of time, a susceptible can remain susceptible or can become an infective or be removed; the infective can remain an infective or can be removed; the removed remains removed; the immunes remain immune and do not return to the population of susceptibles.

Let us impose also the condition that the populations are homogeneously mixed. Then the system of model equations is

$$\frac{dS}{dt} = -\beta SI' - \delta S + \mu; \qquad \frac{dI}{dt} = \beta SI' - \gamma I + \upsilon \tag{5.16}$$

$$\frac{dR}{dt} = \gamma I + \delta S; \qquad \frac{dS'}{dt} = -\beta' S'I - \delta S' + \mu' \tag{5.17}$$

$$\frac{dI'}{dt} = \beta' S'I - \gamma' I' + \upsilon'; \qquad \frac{dR'}{dt} = \gamma' I' + \delta' S'. \tag{5.18}$$

The conditions for development of an epidemic are as follows:

1. If as an initial condition at t_0, a single infective is introduced into the populations N_0 and N_0', then for an epidemic to develop, the change in the number of infectives must be positive in both populations.

2. Thus for $\rho = \frac{\gamma - \upsilon}{\beta}$ and $\rho' = \frac{\gamma' - \upsilon'}{\beta'}$, the threshold for the epidemic arises from the conditions $\beta SI' > \gamma I - \upsilon$ and $\beta' S' I' > \gamma' I' - \upsilon'$, so that the threshold is

$$S_0 S_0' > \rho \rho'. \tag{5.19}$$

3. The development of an epidemic is given by the equation for $\frac{dI}{dt}$.
4. The peaks of the epidemics occur at time points where $\frac{d^2 I}{dt^2} = 0$, while the epidemic's size is given by $I(t \to \infty)$.

The Goffman–Newill model stimulated much research in the area of modeling of processes in science by models from population dynamics and epidemiology. Let us mention here just the models of the growth of mathematics specialties [62] and of the growth of papers in a specialty [63–67]. One can add additional categories of researchers to the SIR type of models. One example of this is the adding of the class of researchers exposed to the corresponding scientific ideas. In such a way, one obtains a class of epidemic SEIR models of research production [68, 69].

5.2.5 *Price Model of Knowledge Growth. Cycles of Growth of Knowledge*

An example of nonepidemic model of knowledge growth is the model of Price [70, 71]. The model is based on the following assumptions:

1. The growth is measured by the number of important publications appearing at a given time.
2. The growth has a continuous character, and a finite time period $T = \text{const}$ is needed to build up a result of fundamental character.
3. The interactions between various scientific fields are neglected.

Let in addition the number of scientists publishing results in this field be constant. Then the rate of scientific growth (of the publications x) is proportional to the number of important publications at time t minus the time period T required to build up a fundamental result. The model equation is

$$\frac{dx}{dt} = \alpha x(t - T), \tag{5.20}$$

where α is a constant, and the initial condition $x(t) = \phi(t)$ is defined on the interval $[-T, 0]$.

Often, the population of researchers is varying. Then for consideration of the evolution of the average number of papers per researcher instead of the linear right-hand side (5.20), the following nonlinear model is used:

$$\frac{dx}{dt} = f(x(t - T), x(t)), \tag{5.21}$$

where f is a homogeneous function of degree one. The simplest form of such a function is a linear function. Let us assume that the population of researchers L grows at the constant rate $n = \frac{1}{L}\frac{dL}{dt}$ and let $z = x/L$ be the mean number of papers written by a researcher. Then the evolution of the number of papers written by a researcher has the form

$$\frac{dz}{dt} = \alpha z(t - T) - nz(t). \tag{5.22}$$

We note the following:

1. If $n = 0$ and $T = 0$, the Price model of exponential growth is recovered.
2. Equation (5.22) is linear, but cyclic behavior may appear because of the feedback between the delayed and nondelayed terms.

The Price model was criticized along the following points: the quality of research is omitted, and many scientific products that seem to be new are not really new; creativity and innovation are confused, and creative papers with new ideas and results have the same importance as trivial duplications. Price answered by formulating the hypothesis that one may study only the growth of *important* discoveries, inventions, and scientific laws, rather than all important and trivial things. Then every growth will follow the same pattern as that mentioned above, but the growth will be much slower.

5.3 A Deterministic Model Connected to Dynamics of Citations

Sangwal [72–75] proposed a model of the growth of citations of a scientist based on the progressive nucleation mechanism known from chemistry [76]. In chemistry, this mechanism describes simultaneous nucleation and growth of a nucleus to crystallites of visible size. If the initial volume of the crystallizing phase is V and the crystallized volume is $V(t)$, then one has the following relationship for the ratio V_c/V:

$$\alpha(T) = \frac{V_c(t)}{V} = \left\{1 - \exp\left[-\left(\frac{t}{\Theta}\right)^q\right]\right\}, \tag{5.23}$$

where the relationships for the time constant Θ and for the exponent q are

$$q = 1 + \nu d; \quad \Theta = \left(\frac{q}{kG^{q-1}J_s}\right)^{1/q},$$

and the parameters are as follows:

- $\nu > 0$: a constant;
- d: dimension of the growing nucleus (can be 1, 2, 3);
- k: shape factor of the nucleus ($k = 4\pi/3$ for a spherical nucleus);

- $G = \frac{r^{1/v}}{t}$;
- r: radius of the growing nucleus;
- J_s: rate of stationary nucleation.

When $kJ_s = G$, then $\Theta = \frac{q^{1/q}}{kJ_s}$, which will be the case of interest for us. In this case, the nuclear radius grows in time as $r(t) \propto t^v$.

The process of nucleation can also be used to describe the growth of citations of a paper written by scientist. In this case,

$$\alpha(t) = \alpha(t) = \frac{C(t)}{C_{max}} = \left\{1 - \exp\left[-\left(\frac{t}{\Theta}\right)^q\right]\right\}, \tag{5.24}$$

where C is the maximum number of citations that a paper can receive, and $C(t)$ is the cumulative number of citations of the paper in the time t. The other parameters are defined as above (we recall that ($\Theta = \frac{q^{1/q}}{kJ_s}$). The nucleation model can be transferred to a description of the accumulation of citations of a paper if several conditions are met:

- Citations received by a paper and the paper earning these citations compose a closed system in which the process of occurrence of citations is stationary.
- Occurrence of citations of a paper continues in time and finally approaches a constant value C_{max}, which is the maximum number of citations received by the paper at time T.
- The dependence of the cumulative number of citations $C(t)$ of the paper at time t is determined by the maximum number of citations C_{max}, a time constant Θ, and an exponent q. The citation pattern of different papers of an author is characterized by different values of $C(t)$, Θ, and q for each paper.

If a researcher has authored n papers, then the cumulative fraction $\alpha_s(t)$ of the citations of these papers is

$$\alpha_s(t) = \sum_{i=0}^{n} \alpha_i(t). \tag{5.25}$$

If we assume that the researcher publishes papers at equal time intervals ΔT, then

$$\alpha_s(t) = \sum_{i=0}^{n} \alpha_i[t - (i-1)\Delta T] = \sum_{i=1}^{n} \left\{1 - \exp\left[-\left(\frac{t - (i-1)\Delta T}{\Theta_i}\right)\right]\right\}. \tag{5.26}$$

One can fit the model parameters for the data of the researcher whose production is evaluated. In most cases, the fit describes very well the process of accumulation of citations [75].

5.4 Deterministic Models Connected to Research Dynamics

5.4.1 *Continuous Model of Competition Between Systems of Ideas*

Ideas can diffuse not only among scientists in one organization but also in space (e.g., from scientists from one country to scientists from other countries). Thus one may include spatial variables in the models describing the diffusion of ideas. Such models can be of great interest during periods of globalization of economies, knowledge, and technology [77–82]. Below, we describe a model closely connected to the space–time models of migration of populations [83, 84].

The diffusion of ideas is often accompanied by competition between systems of ideas. Let a population of N individuals occupy a two-dimensional plane. We assume that:

- there exists a set of ideas $P = \{P_0, P_1, \ldots, P_n\}$;
- N_i members of the population are followers of the set P_i of ideas;
- members N_0 of the class P_0 are not supporters of any set of ideas.

In such a way, the population is divided into $n + 1$ subpopulations of followers of different sets of ideas, and $N = N_0 + N_1 + \cdots + N_n$. Let a small region $\Delta S = \Delta x \Delta y$ be selected in the plane. In this region, there are ΔN_i individuals holding the ith set of ideas, $i = 0, 1, \ldots, n$. If ΔS is sufficiently small, the density of the ith population can be defined as $\rho_i(x, y, t) = \frac{\Delta N_i}{\Delta S}$. Further, we assume that members of the ith population are capable of moving through the borders of the area ΔS. Let $\mathbf{j}_i(x, y, t)$ be the current of this movement. The total change in the number of members of the ith population is

$$\frac{\partial \rho_i}{\partial t} + \mathrm{div}\mathbf{j}_i = C_i, \tag{5.27}$$

where the changes are summarized by the function $C_i(x, y, t)$.

The first term in (5.27) describes the net rate of increase of the density of the ith population. The second term describes the net rate of immigration into the area. The right-hand side of (5.27) describes the net rate of increase exclusive of immigration. The quantities $\mathbf{j}_i$ and C_i are as follows: $\mathbf{j}_i$ is assumed to have two parts, a nondiffusion part $\mathbf{j}_i^{(1)}$ and a diffusion part $\mathbf{j}_i^{(2)}$ that is assumed to have the general form of a linear multicomponent diffusion [77] (D_{ik} is the coefficient of diffusion):

$$\mathbf{j}_i = \mathbf{j}_i^{(1)} + \mathbf{j}_2^{(2)} = \mathbf{j}_i^{(1)} - \sum_{k=0}^{n} D_{ik}(\rho_i, \rho_k, x, y, t)\nabla\rho_k. \tag{5.28}$$

A further assumption is that some of the followers of the set of ideas P_i are capable of changing to another set of ideas, e.g., they can change P_i for P_j. It can be assumed that the following processes can occur with respect to the members of the subpopulations:

- **Deaths**: described by a term $r_i\rho_i$. We assume that the number of deaths in the ith population is proportional to its population density. In general, $r_i = r_i(\rho_\nu, x, y, t; p_\mu)$, where ρ_ν stands for $(\rho_0, \rho_1, \ldots, \rho_N)$ and p_μ stands for $(p_1, \ldots, p_M)$, containing parameters of the environment.
- **Noncontact conversion**: in this class are included kinds of changes between P_i and P_j exclusive of changes after interpersonal contact between the members of populations. A reason for noncontact conversion can be the existence of different kinds of mass communication media (scientific books, influence of mass media, etc.). For the ith population, the change in the number of members by this kind of conversion is $\sum_{j=0}^{n} f_{ij}\rho_j, f_{ii} = 0$. In general, $f_{ij} = f_{ij}(\rho_\nu, x, y, t; p_\mu)$.
- **Contact conversion**: this happens by interpersonal contacts among the members of the population. Such contacts can happen between members in groups consisting of two members (binary contacts), three members (ternary contacts), four members, etc. As a result of the contacts, members of each population can change their sets of ideas. For binary contacts, let it be assumed that the probability of change for a member of the jth population is proportional to the probability of, for instance, the number of contacts, i.e., proportional to the density of the ith population. Then the total number of "conversions" from P_j to P_i is $a_{ij}\rho_i\rho_j$, where a_{ij} is a parameter.
  Next, a change in the set of ideas can take place by ternary contact. For this, one must have a group of three members. We assume that such a group exists with a probability proportional to the corresponding densities of the concerned populations. In a ternary contact between members of the ith, jth, and kth populations, members of the jth and kth populations can change their sets of ideas to $P_i = b_{ijk}\rho_i\rho_j\rho_k$, where b_{ijk} is a parameter. In general, $a_{ij} = a_{ij}(\rho_\nu, x, y, t; p_\mu)$; $b_{ijk} = b_{ijk}(\rho_\nu, x, y, t; p_\mu)$; etc.

On the basis of all of the above the C_i term can be written as

$$C_i = r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j + \sum_{j=0}^{n} a_{ij}\rho_i\rho_j + \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \ldots. \tag{5.29}$$

Hence the model system of equations is

$$\frac{\partial\rho_i}{\partial t} + \operatorname{div}\mathbf{j}_i^{(1)} - \sum_{j=0}^{n} \operatorname{div}(D_{ij}\nabla\rho_j) = r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j + \sum_{j=0}^{n} a_{ij}\rho_i\rho_j + \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \ldots. \tag{5.30}$$

The density of the entire population is $\rho = \sum_{i=0}^{n} \rho_i$. This density can change over time. One possible assumption is that ρ changes over time according to the Verhulst law

$$\frac{\partial\rho}{\partial t} = r\rho\left(1 - \frac{\rho}{C}\right), \tag{5.31}$$

where $C(\rho_v, x, y, t; p_\mu)$ is the carrying capacity of the environment and $r(\rho_v, x, y, t; p_\mu)$ is a positive or negative growth rate.

Now let us consider the case in which the current $\mathbf{j}_i^{(1)}$ is negligible, i.e., $\mathbf{j}_i^{(1)} \approx 0$. In addition, we consider only the case in which all parameters are constants. The model system of equations becomes

$$\frac{\partial \rho_i}{\partial t} - D_{ij} \sum_{j=0}^{n} \Delta \rho_j = r_i \rho_i + \sum_{j=0}^{n} f_{ij} \rho_j + \sum_{j=0}^{n} a_{ij} \rho_i \rho_j + \sum_{j,k=0}^{n} b_{ijk} \rho_i \rho_j \rho_k + \ldots, \tag{5.32}$$

where

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \quad i = 0, 1, 2, \ldots, n. \tag{5.33}$$

Next we shall separate the dynamics of averaged quantities from the dynamics of fluctuations. If $q(x, y, t)$ is a quantity defined in an area S, then the corresponding plane averaged quantity is

$$\overline{q} = \frac{1}{S} \int \int_S dxdy\, q(x, y, t). \tag{5.34}$$

The fluctuations are denoted by $Q(x, y, t)$:

$$q(x, y, t) = \overline{q}(t) + Q(x, y, t). \tag{5.35}$$

We assume that territory S is large enough; every plane averaged combination of fluctuations vanishes; $\int \int_S dxdy \Delta Q_k$ is finite. Then $\overline{\Delta Q_k} = \frac{1}{S} \int \int_S dxdy \Delta Q_k \to 0$. On the basis of these assumptions, the dynamics of the averaged quantities are separated from the dynamics of fluctuations by means of a plane averaging of (5.32). The result is

$$\overline{\rho}_0 = \overline{\rho} - \sum_{i=1}^{n} \overline{\rho}_i; \quad \frac{d\overline{\rho}}{dt} = r\overline{\rho} \left(1 - \frac{\overline{\rho}}{C}\right) \tag{5.36}$$

$$\frac{d\overline{\rho}_i}{dt} = r_i \overline{\rho}_i + \sum_{j=0}^{n} f_{ij} \overline{\rho}_j + \sum_{j=0}^{n} a_{ij} \overline{\rho}_i \overline{\rho}_j + \sum_{j,k=0}^{n} b_{ijk} \overline{\rho}_i \overline{\rho}_j \overline{\rho}_k + \ldots. \tag{5.37}$$

Instead of (5.36), we can write an equation for $\overline{\rho}_0$ of the type of (5.37). Then the total population density $\overline{\rho}$ will not follow the Verhulst law.

Equations (5.36) and (5.37) represent the model of competition among sets of ideas proposed in [85]. There also exists a discrete version of this model [86], and it can be applied to competition between different sets of ideas (scientific, political, religious, technological, etc.).

5.4.2 *Reproduction–Transport Equation Model of the Evolution of Scientific Subfields*

By means of migration, people can move from one territory to another. The change of the field of research by a scientist may also be considered a migration process [82, 87]. In order to study this, let us map research problems by sequences of signal words or macro-terms $P_i = (m_i^1, m_i^2, \ldots, m_i^k, \ldots, m_i^n)$, which are registered according to the frequency of their occurrence in the texts. Then:

- Each point of the problem space, described by a vector $\mathbf{q}$, corresponds to a research problem, with the problem space containing all scientific problems (no matter whether they are under investigation or not).
- The scientists distribute themselves over the space of scientific problems with density $x(\mathbf{q}, t)$. Thus there is a number $x(\mathbf{q}, t)d\mathbf{q}$ of scientists working at time t in the element $d\mathbf{q}$.
- The field mobility processes correspond to a density change of scientists in the problem space, i.e., instead of working on problem $\mathbf{q}$, a scientist may begin to work on problem $\mathbf{q}'$.
- As a result, $x(\mathbf{q}, t)$ decreases and $x(\mathbf{q}', t)$ increases.

This movement of scientists can be described by means of a reproduction–transport equation:

$$\frac{\partial x(\mathbf{q}, t)}{\partial t} = x(\mathbf{q}, t)\, w(\mathbf{q} \mid t) + \frac{\partial}{\partial \mathbf{q}} \left(f(\mathbf{q}, x) + D(\mathbf{q}) \frac{\partial \mathbf{q}}{\partial x} \right). \tag{5.38}$$

In (5.38), self-reproduction and decline are represented by the term $w(\mathbf{q} \mid x)\, x(\mathbf{q}, t)$; for the reproduction rate function $w(\mathbf{q} \mid x)$, one can write the relationship

$$w(\mathbf{q} \mid x) = a(\mathbf{q}) + \int d\mathbf{q}' b(\mathbf{q}, \mathbf{q}' x(\mathbf{q}, t). \tag{5.39}$$

The local value of $a(\mathbf{q})$ is an expression of the rate at which the number of scientists in field $\mathbf{q}$ is growing through self-reproduction and decline. The function $b(\mathbf{q}, \mathbf{q}')$ describes the influence exerted on the field $\mathbf{q}$ by the neighboring field $\mathbf{q}'$. The field mobility is modeled by means of the term $\frac{\partial}{\partial \mathbf{q}} \left(f(\mathbf{q}, x) + D(\mathbf{q}) \frac{\partial}{\partial \mathbf{q}} x(\mathbf{q}, t) \right)$.

In order to use this equation, we need initial conditions and determination of the coefficients on the basis of statistical data for the distribution of the scientists with respect to the research problems.

5.4.3 Deterministic Model of Science as a Component of the Economic Growth of a Country

Below we discuss a component of the model of evolution of the GDP (gross domestic product) of a country. This component is connected to the role of technology for increasing GDP [88–90].

The GDP of a country may grow extensively by inflow of workforce or capital to the national economic structures and systems [91]. But the GDP of a country may grow also intensively by advancement in science and technology. Let us discuss a simple model in which the GDP Y has the form

$$Y(t) = Y(L(t), C(t), T(t)). \tag{5.40}$$

The quantities in (5.40) are as follows:

- $L(t)$: labor (human resources);
- $C(t)$: production resources;
- $T(t)$: technology level.

Note that the above quantities are not chosen arbitrarily. They represents important factors that may influence the GDP of a country.

The change in the GDP over time is given by

$$\frac{dY}{dt} = \frac{\partial Y}{\partial L}\frac{dL}{dt} + \frac{\partial Y}{\partial C}\frac{dC}{dt} + \frac{\partial Y}{\partial T}.\frac{dT}{dt}. \tag{5.41}$$

The term $(\partial Y/\partial T)(dT/dt)$ describes the change in the GDP because of the evolution of technology. This component of the change of the GDP will be of interest for us below. Let us note that if technology advances, $((dT/dt) > 0)$, this is a contribution to the growth of the GDP. If technology for some reason deteriorates, $((dT/dt) < 0)$, then it can contribute to a decrease in the GDP.

The change in the GDP due to technology may be assumed to be [92]

$$\frac{\partial Y}{\partial T} = \frac{Y}{T}. \tag{5.42}$$

Equation (5.42) means that the increase in the technology level leads to a proportional increase of the GDP. Then the studied term from (5.41) becomes

$$\frac{\partial Y}{\partial T}\frac{dT}{dt} = Y\left(\frac{1}{T}\frac{dT}{dt}\right). \tag{5.43}$$

Next we shall discuss how the term $(1/T)(dT/dt)$ depends on S_T: the growth in knowledge about technology. Then the growth in knowledge about technology will be connected to the growth in scientific knowledge, which will be denoted by S.

We adopt the following notation:

- I_T: the investment directed to applications of the results of new technologies (machines, processes, etc.);
- I_0: the investments in older technologies;
- γ: coefficient of proportionality between the growth of knowledge about technology S_T and growth of scientific knowledge S.

Then the relationship between T and S is

$$\frac{1}{T}\frac{dT}{dt} = \gamma \frac{I_T}{I_O}\frac{1}{S}\frac{dS}{dt}. \tag{5.44}$$

Equation (5.44) leads to the following conclusions:

1. **Importance of the fundamental research**: Research and especially fundamental research lead to an increase in scientific knowledge. If there is no growth in scientific knowledge, $((dS/dt) = 0)$, then there is no technological evolution, $((1/T)(dT/dt) = 0)$, and an important factor for the growth of the national GDP is lost.
2. **Importance of the transfer of scientific knowledge to knowledge about technology**: If $\gamma = 0$, i.e., there is no transfer, then $((1/T)(dT/dt) = 0)$ (no technology evolution) even if scientific knowledge grows. Thus what is important for a country is to increase γ (by strengthening engineering sciences by creating new engineering institutes, for example). The value of γ for developed countries is about 0.5 (1 % growth in scientific knowledge results in 0.5 % growth in the number of patents).
3. **Importance of investment in new technologies**: If there is no such investment $(I_T = 0)$, then there is no evolution of technology, $((1/T)(dT/dt) = 0)$, even if there is growth of scientific knowledge and an intensive transfer of knowledge about technology.

The rate of growth of scientific knowledge $(1/S)(dS/dt)$ is assumed to depend on two main factors: the funding of (investment in) science I and the labor L ("human resources" or the number of qualified scientists). Let us set

$$\frac{1}{S}\frac{dS}{dT} = \phi(I, L). \tag{5.45}$$

Let us assume that $\phi(I, L)$ is a homogeneous function of degree α with respect to the funding I and a homogeneous function of the factor β with respect to the human resources L. Then we can obtain the relationship

$$\phi = aI^{\alpha}L^{\beta} = \frac{1}{S}\frac{dS}{dt}, \tag{5.46}$$

where a is a coefficient of integration. Hence a power-law relationship may exist between the rate of growth of scientific knowledge and investment and the number

of qualified scientists. We stress the words *power law*, since such laws arise frequently in studies of research systems (for examples, see Chap. 5).

Equation (5.46) leads to interesting conclusions.

1. **Exponential growth of knowledge in an established research area**. Let us consider an established research area with constant investment in science: $I =$ const and a constant number of qualified scientists $L =$ const. From (5.46), we obtain the relationship
$$S = S_0 \exp[aI^\alpha L^\beta t] \tag{5.47}$$
(S_0 is a constant of integration), which means that the scientific knowledge in this area is growing exponentially.
2. **Double-exponential growth of scientific knowledge in a new research area**. Let us now consider a new research area in which the number of scientists grows exponentially over time, $L = \exp(\mu t)$, and the funding is constant: $I =$ const and large enough. Then the growth of scientific knowledge in this area is double-exponential,
$$S = S_0 \exp\left[\frac{aI^\alpha}{\mu\beta}\exp(\mu\beta t)\right]. \tag{5.48}$$

The substitution of (5.44)–(5.46) in (5.43) leads to the following relationship for the influence of science on the change of GDP of a country:

$$\frac{\partial Y}{\partial T}\frac{dT}{dt} = \gamma a \frac{I_T}{I_O} I^\alpha L^\beta Y. \tag{5.49}$$

Equation (5.49) shows that countries that have a large GDP possess advantages (since $\frac{\partial Y}{\partial T}\frac{dT}{dt} \propto Y$), and in addition, the human factor and investment in science are very important. Thus every nation should try to build a community of qualified researchers and should invest sufficiently in the national research system. If this is not the case, then the process of global competition among the nations will lead inevitably to a brain drain.

The model above represents a global point of view of the importance of science as a component of economic growth of a country. There exists also a local point of view regarding this importance. A local point of view means that one considers the growth of the output of a worker with advancing technology. A mathematical model of this relationship may be based on the Cobb–Douglass production function and on the Solow model. The form of the Cobb–Douglass production function is [93, 94]

$$Y = AK^\alpha L^{1-\alpha}, \tag{5.50}$$

where

- Y: output per worker;
- K: physical capital per worker;
- L: human capital per worker (labor);

- A: productivity;
- α: output elasticity of the physical capital;
- $\beta = 1 - \alpha$: output elasticity of the human capital.

Looking at (5.50), we can conclude that technological advance allows (by increasing productivity) given quantities of physical and human capital to be combined to produce more output than was possible when older technology was used. Hence changes in technology directly affect economic growth. In addition, human capital L per worker cannot grow infinitely. Then in order to increase the output Y, one has to increase the physical capital K per worker (there are also limits to this increase), or one can increase productivity A by advancing technology. Thus even when K and L have reached their maximum values, *as long as A (productivity) continues to grow as a consequence of technological advance, income per capita will continue to grow too*.

The result of the mathematical theory is that the rate of growth of the total output $Y^* = (1/Y)(dY/dt)$ per worker (in the steady state of the production system) is connected to the growth of productivity A (which means that there is a strong connection between the growth of the total output and technological progress). Namely, if the rate of advance of technology is $A^* = (1/A)(dA/dt)$, then

$$Y^* = A^* \left(\frac{1}{1-\alpha} \right). \tag{5.51}$$

Equation (5.51) tells us that technological advance (by research and development) is extremely important for economic growth.

5.5 Several General Remarks About Probability Models and Corresponding Processes

In many cases, in the mathematical models of mechanisms of production of scientific information, one uses the concept of population of "sources" producing "items" observed over time [95]. *The observation of the items produced by a source is equivalent to the observation of a stochastic point process: a sequence of events occurring randomly in time*. The modeling of the corresponding process requires specification of the probabilistic mechanism producing the observed events.

> The simplest available point process is the Poisson process, which corresponds to the situation that events occur completely at random over time with the overall average rate of occurrence remaining constant, so that the expected number of events occurring increases linearly with time.

In order to model more realistic situations, the rate of the Poisson process may:

1. vary in time deterministically [96]. In this case, the number of occurring events may have nonlinear variation in time, and the process is called an inhomogeneous Poisson process;
2. vary in time stochastically [97, 98]. Such a process is called a doubly stochastic Poisson process or Cox process.

Each of the three Poisson processes described above has independent increments. The Poisson process and the doubly stochastic Poisson process have stationary increments. Thus they are able to model situations in which the probability distribution of the number of events in a period of time depends only on the length of the period and not on the time at which it begins.

When the entire population of sources is studied, it may happen that some variability in the rate of production between different items exists. The observed process is then a mixture of the individual processes, and it can be modeled mathematically by mixing the parameters determining the rates of production of the individual sources. The resulting mixed process may still have stationary increments, but because of the mixing, the increments are no longer independent.

We are going to describe briefly three kinds of Poisson processes that will arise in the models discussed below: the Greenwood–Yule process (gamma–Poisson process), GIGP (generalized inverse Gaussian–Poisson process), and Waring process (a negative binomial process) [95]. Let us consider a source that produces X_t ($t \geq 0$) items in the interval $[0, t]$. The process of production of items (the point process) is specified by a parameter θ, and we know the form of the process $\{X_t \mid \theta\}$ for a given value of θ. For given θ, the increments of the process are stationary but not independent, and

$$p(X_t = r) = E_\theta P(X_t = r \mid \theta) = \int dx f_\theta(x) p(X_t = r \mid \theta = x). \tag{5.52}$$

The above-mentioned three processes will be obtained by specifying the probability distribution function $f_\theta(x)$ and the form of the conditional process $\{X_t \mid \theta\}$. For example, in order to obtain the Greenwood–Yule process (called also gamma–Poisson process), we have to assume that each source produces items as a Poisson process and the probability distribution function is for the gamma distribution. In detail,

$$p(X_t = r \mid \lambda) = \exp(-\lambda t) \frac{(\lambda t)^r}{r!}; \quad r = 0, 1, \ldots, \tag{5.53}$$

where λ is the rate of the Poisson process; λ has a gamma distribution with scale parameter β and index ν:

$$f_\lambda(x) = \frac{\beta^{-\nu} x^{\nu-1}}{\Gamma(\nu)} \exp(-\frac{x}{\beta}); \quad x > 0. \tag{5.54}$$

As a result of substituting (5.53) and (5.54) in (5.52), one obtains the negative binomial distribution of index ν and parameter $p_t = 1/(1+\beta t)$:

$$p(X_t = r) = \binom{r+\nu-1}{r}\left(\frac{1}{1+\beta t}\right)^{\nu}\left(\frac{\beta t}{1+\beta t}\right)^{r}; \quad r = 0, 1, \ldots. \tag{5.55}$$

The GIGP (generalized inverse Gaussian–Poisson process) is obtained when the probability distribution function for the rate λ of the Poisson process (5.53) is

$$f_{\lambda}(x) = c(\alpha, \gamma, \theta)x^{\gamma-1}\exp\left[-x\left(\frac{1}{\theta}-1\right)-\frac{\alpha^2\theta}{4x}\right], \tag{5.56}$$

where $x > 0$; $-\infty < \gamma < \infty$; $\alpha \geq 0$, and the constant ensuring the normalization is

$$c(\alpha, \gamma, \theta) = \frac{(1-\theta)^{\gamma/2}}{2(\alpha\theta/2)^{\gamma}}K_{\gamma}\{\alpha(1-\theta)^{1/2}\}, \tag{5.57}$$

where $K_{\gamma}\{\alpha(1-\theta)^{1/2}\}$ is the modified Bessel function of the second kind of order γ. The substitution of the density (5.56) in (5.52) leads to the distribution

$$p(X_t = r) = \frac{(1-\theta_t)^{\gamma/2}}{K_{\gamma}\{\alpha(1-\theta)^{1/2}\}}\frac{(\alpha_t\theta_t/2)^r}{r!}K_{r+\gamma}(\alpha_t); \quad r = 0, 1, \ldots, \tag{5.58}$$

where $\theta_t = (t\theta)/[1+\theta(t-1)]$ and $\alpha_t = \alpha[1+(t-1)\theta]^{1/2}$. This distribution is reduced to the GIGP distribution when $t = 1$ (then $\theta_t = \theta$ and $\alpha_t = \alpha$). Because of this, the process X_t described by (5.58) will be called a GIGP process and may be denoted by GIGP$(\alpha_t, \theta_t, \gamma)$. Sichel [99, 100] used $\gamma = -1/2$, i.e., the GIGP$(\alpha_t, \theta_t, -1/2)$ distribution

$$p(X_t = r) = \left(\frac{2\alpha_t}{\pi}\right)^{1/2}\exp[\alpha(1-\theta)^{1/2}]\frac{(\alpha_t\theta_t/2)^r}{r!}K_{r-1/2}(\alpha_t); \quad r = 0, 1, \ldots, \tag{5.59}$$

in many practical applications.

Finally, we consider the Waring process (which will be much discussed below in the text). For this process, each source produces items as a negative binomial process of parameter q and index ψ:

$$p(X_t = r \mid q) = \binom{r+\psi t-1}{r}q^{\psi t}(1-q)^r; \quad r = 0, 1, \ldots, \tag{5.60}$$

and the parameter q has a beta distribution with parameters a and b:

$$f_p(x) = \frac{1}{B(a,b)}\frac{\psi^a x^{b-1}}{(x+\psi)^{a+b}}. \tag{5.61}$$

The substitution of (5.60) and (5.61) in (5.52) leads to

$$p(X_t = r) = \frac{\Gamma(\psi t + a)}{B(a,b)\Gamma(\psi t)} \frac{\Gamma(r + \psi t)\Gamma(r + b)}{r!\Gamma(r + \psi t + a + b)}. \tag{5.62}$$

Equation (5.62) describes the generalized Waring distribution [101–103]; Γ is the gamma function, and B is the beta function.

Some remarks about the moments of the obtained distributions follow. Moments of all orders exist for the gamma–Poisson distribution and for the GIGP distribution. For the existence of moments of the generalized Waring distribution, one has to impose some requirements on the parameters of the distribution. For the gamma–Poisson distribution, the mean $E[X_t]$ and the variance $V[X_t]$ are

$$E[X_t] = \nu\beta t, \tag{5.63}$$

$$V[X_t] = \nu\beta t(1 + \beta t). \tag{5.64}$$

For the GIGP distribution with $\gamma = -1/2$,

$$E[X_t] = \frac{\alpha\theta t}{2(1-\theta)^{1/2}}, \tag{5.65}$$

$$V(X_t) = \frac{\alpha\theta t}{4(1-\theta)^{3/2}}[2(1-\theta) + t\theta]. \tag{5.66}$$

For the generalized Waring distribution,

$$E[X_t] = \frac{\psi bt}{a-1}; \quad a > 1, \tag{5.67}$$

$$V(X_t) = \frac{\psi b(a+b-1)}{(a-1)^2(a-2)}(a - 1 + \psi t); \quad a > 2. \tag{5.68}$$

5.6 Probability Model for Research Publications. Yule Process

Probability models are very interesting and powerful tools for the study of the dynamics of research systems and characteristics of research production. Let us demonstrate this with a discussion of a probability model of dynamics of research publications [104] that will lead us to the famous statistical distribution of Yule.

Let us now consider scientific publications from the following point of view. A researcher has x publications. Then he/she writes one more publication, and we shall consider this as a transition to another state characterized by $x + 1$ publications. The

occurrence of a new publication is a rare event, and because of this, we shall consider the process of the occurrence of a new publication to be a Poisson pure multiplicative random process where the probability of transition to a new state in the time interval $(t, t + \Delta t)$ depends on the state of the system at time t.

5.6.1 *Definition, Initial Conditions, and Differential Equations for the Process*

We begin our study at the point in time where a studied researcher has one publication. Let $p_x(t)$ be the probability that a researcher has x publications at time t. Then the initial condition is $p_x(0) = 1$ if $x = 1$ and $p_x(0) = 0$ if $x \neq 1$. The process evolves according to the following two rules:

1. The probability of a transition from state x to state $x + 1$ in the interval $(t, t + \Delta t)$ is proportional to the interval Δt. We denote this probability by $\lambda(x)\Delta t$.
2. The probability of two or more transitions for the interval Δt is negligibly small.

Because of the above rules, the probability of a lack of transition between the states x and $x + 1$ in the time interval $(t, t + \Delta t)$ is $1 - \lambda(x)\Delta t$.

The probability that our system (the researcher) is in the state x (has x publications) for the interval $(t, t + \Delta t)$ is the sum of the probability that the system jumped there from the state $x - 1$ within the time interval and the probability that the system has not jumped to the next state $x + 1$ within the time interval. In symbols, this reads

$$p_x(t + \Delta t) = [1 - \lambda(x)\Delta t]p_x(t) + \lambda(x - 1)p_{x-1}(t)\Delta t. \tag{5.69}$$

This can be written as the following system of differential equations for the probability:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t), \\ \frac{dp_x(t)}{dt} &= -\lambda(x)p_x(t) + \lambda(x - 1)p_{x-1}(t). \end{aligned} \tag{5.70}$$

5.6.2 *How a Yule Process Occurs*

In order to continue analysis of (5.70), we have to determine $\lambda(x)$. We shall use the linear hypothesis for the parameter $\lambda(x)$:

The probability of a transition increases proportionally to the number of publications:

$$\lambda(x) = \lambda x, \tag{5.71}$$

where λ *is a constant.*

In other words, there is a linear hypothesis of the following kind: *If an author has many publications, he/she doesn't need much time to produce another one. In this way, our stochastic process becomes a linear pure multiplicative process* (**Yule process**) [105–109].

Using (5.71), one obtains the following solution of the system of equations (5.70): $p_x(t) = 0$ when $x = 0$ and

$$p_x(t) = [1 - \exp(-\lambda t)]^{x-1} \exp(-\lambda t). \tag{5.72}$$

Let us recall that in the case under discussion, the distribution (5.72) gives the probability that a researcher will have x publications at time t if at time $t = 0$, he had one publication.

5.6.3 Properties of Research Production According to the Model

1. *Expected value.*
   The expected value is the mean number of publications that are expected to be written for time t. Then
   $$E[x(t)] = \exp(\lambda t), \tag{5.73}$$
   which is often observed in practice and is called the **law of exponential growth of science**.
2. λ: *a measure of the publication activity of the researchers.*
   After a "differentiation" of (5.73), one obtains
   $$\lambda = \frac{dx_t/dt}{x_t}, \tag{5.74}$$
   which means that λ is the rate of growth of the number of publications, i.e., a measure of the intensity of publication (and partially of the scientific) activity of a researcher.
3. *Research work in a research area for some finite time.*
   Usually, a researcher works for some (finite) time on problems from some research area and then changes the research area of work (or retires). This time depends

on the potential of the research area, on the talent of the researcher, on the age of the researcher, on the work conditions, etc. The finite time of work is different for different researchers and is a random variable whose distribution can be obtained from queuing theory. The distribution is

$$p(t) = \nu \exp(-\nu t), \tag{5.75}$$

where $\nu = 1/t^*$ and t^* is the average value of t. This random distribution of the time of activity in a research area can be incorporated in the Yule distribution as $p_x(t) = p(x/t)$. Then in order to obtain the probability distribution of the publications that are observed in a database, we have to calculate the following integral:

$$p(x) = \int_0^\infty dt\, p(x/t)p(t) =$$

$$\int_0^\infty dt\, [1 - \exp(-\lambda t)]^{x-1} \exp(-\lambda t)\nu \exp(-\nu t). \tag{5.76}$$

The integration of (5.76) leads to the *Yule distribution*

$$p(x) = \alpha B(x, \alpha + 1), \tag{5.77}$$

where:

- $B(x, \alpha + 1) = \frac{\Gamma(x)\Gamma(\alpha+1)}{\Gamma(x+\alpha+1)}$ is the beta function;
- $\Gamma(x) = (x - 1)!$ is the gamma function;
- $\alpha = \nu/\lambda$.

The Yule distribution obtained above leads to several interesting conclusions about research production.

1. **Asymptotic behavior**: For large x, one obtains $\frac{\Gamma(x)}{\Gamma(x+\alpha+1)} \approx \frac{1}{x^{\alpha+1}}$ (the Stirling approximation was used). Let us in addition assume that α has small values. Then $\Gamma(\alpha + 1) \approx 1$, and the Yule distribution is reduced to

$$p(x) \approx \alpha\Gamma(\alpha + 1)\frac{1}{x^{\alpha+1}} \approx \frac{\alpha}{x^{\alpha+1}}, \tag{5.78}$$

which is the law of Pareto for $x_0 = 1$ and small values of α. Thus on the basis of the hypothesis that the scientific activity is a random branching multiplicative process with linear increase of effectiveness of the researchers (Yule process), we have obtained one of the basic laws of research production.

2. **Evaluation of the parameter** α: This can be done on the basis of the Yule distribution for researchers who have just one publication. For these researchers,

$$p(1) = \frac{\alpha\Gamma(1)\Gamma(\alpha+1)}{\Gamma(\alpha+2)} = \frac{\alpha}{\alpha+1} \tag{5.79}$$

(we have used $\Gamma(1) = 1$ and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$). Then taking into account that $p(1) = N_1/N$ is the proportion of the number N_1 of researchers with one publication in a group of N researchers, we obtain

$$\alpha = \frac{p_1}{1-p_1} = \frac{N_1}{N-N_1}. \tag{5.80}$$

Thus we can evaluate α by taking N and N_1 from a large enough database.

5.7 Probability Models Connected to Dynamics of Citations

5.7.1 Poisson Model of Citations Dynamics of a Set of Articles Published at the Same Time

Citation analysis is one of the frequently used methods of assessment of research impact [110–114]. An important topic in the research on citations is the investigation of citation distributions. This research may follow two paths [115]:

1. *Path 1*: Take a particular source—book, article, journal issue, journal volume, etc.—and study the age distribution of the cited articles in the studied source [116].
2. *Path 2*: Take a collection of sources (articles published in a journal, or articles from some scientific field) at a given time and then follow up and note the times at which each source from the collection is cited [117, 118].

Below, we present a probability model obtained by following *Path 2* and assuming continuous time as well as the presence of aging of published material (in the course of time, the material becomes obsolescent (and less frequently cited)) and the existence of publications that are never cited. The model is as follows [115]. Let us consider a population of sources that produces items over time. The population (for the case of citation analysis) consists of a collection of articles published at the same time $t = 0$. The items produced by the papers are their citations. The assumption is that citations are received randomly over time. Since different articles are in different scientific areas (with different popularity) and have different relevance, etc., their citation rates are also different. We assume that these rates of a randomly chosen source are characterized by a random variable Λ that has probability distribution F_Λ over the population of sources. Let X_t be the number of citations to a *randomly chosen*

source (article) in the interval $[0, t]$. The probability that this number of citations will be equal to r is

$$p(X_t = r) = \int_0^\infty dF_\Lambda(\lambda^*)\, P(X_t = r \mid \Lambda = \lambda^*). \tag{5.81}$$

We can recognize the process $\{X_t, t \geq 0\}$ as a counting process, and the model (5.81) is a mixture of counting processes with mixing distribution F_Λ and mixing parameter λ. Next, one has to assume the nature of the process connected to the conditional term $P(X_t = r \mid \Lambda = \lambda^*)$. The initial assumption can be that this process is a Poisson process [119–122] with stationary and independent increments. This will lead us to the distribution

$$P(X_t = r \mid \Lambda = \lambda^*) = \exp(-\lambda^* t)\frac{(\lambda^* t)^r}{r!}; \quad r = 0, 1, 2, \ldots. \tag{5.82}$$

In (5.82), $\lambda^* = \text{const}$, and the mean of the Poisson distribution is $\lambda^* t$. We note here that numerous models of citation distribution have been proposed based on different probability distribution functions $f(\lambda^*)$, $(dF_\Lambda(\lambda^*) = f(\lambda^*)d\lambda^*)$ [123].

Let us now consider the case in which λ^* depends on time. Since λ^* can be associated with the citation rate of a given paper, it can vary with the time t. If $\lambda^* = \lambda^*(t)$, then (5.82) has to be substituted by the more complicated equation [124]

$$P(X_t = r \mid \Lambda = \lambda^*) = \exp[-M(\lambda^*, t)]\frac{M(\lambda^*, t)^r}{r!}; \quad r = 0, 1, 2, \ldots, \tag{5.83}$$

where

$$M(\lambda^*, t) = \int_0^t ds \lambda^*(s).$$

In the case of citations of articles, an almost universal citation pattern in time $c(t)$ can be observed. Then we can assume that the citation rate $\lambda^*(t)$ of a paper has the particular form

$$\lambda^*(t) = \lambda c(t), \tag{5.84}$$

where $\lambda = \text{const}$. Then

$$M(\lambda^*, t) = \int_0^t ds\, \lambda c(s) = \lambda C(t); \quad C(t) = \int_0^t ds\, c(s) \tag{5.85}$$

and

$$P(X_t = r \mid \Lambda = \lambda^*) = \exp[-\lambda C(t)]\frac{[(\lambda C(t)]^r}{r!}; \quad r = 0, 1, 2, \ldots. \tag{5.86}$$

The mean of the Poisson process is $\lambda C(t)$; $c(t)$ is called the *obsolescence density function*; and $C(t)$ is called the *obsolescence distribution function* ($t > 0$). We assume that $\lim\limits_{t\to\infty} C(t) < \infty$.

The substitution of (5.86) in (5.81) leads to the final relationship for the citation production distribution:

$$p(X_t = r) = \int\limits_0^\infty dF_\Lambda(\lambda)[\lambda C(t)]^r \left[\frac{\exp[-\lambda C(t)]}{r!}\right], \; r = 0, 1, 2, \ldots. \tag{5.87}$$

This can also be written as the expected value

$$p(X_t = r) = E_\Lambda[P(X_t = r \mid \Lambda)]. \tag{5.88}$$

From (5.87), one can obtain the first citation distribution. Let T be the time after publication of the first citation of a randomly chosen source (article). We can consider T a random variable. For times $t < T$, the number of citations of a paper is 0. Then let $F_T(t)$ be the cumulative distribution function of the first citation time: $F_T(t) = p(T \le t)$. Since $p(T \le t) = 1 - p(T > t)$ and $p(T > t)$ is the same as the probability $p(X_t = 0)$, we have

$$F_T(t) = 1 - p(X_t = 0) = 1 - \int\limits_0^\infty dF_\Lambda \exp[-\lambda C(t)]. \tag{5.89}$$

An interesting consequence obtained on the basis of the first citation distribution (5.89) is as follows.

There will be publications that will be never cited.

This feature follows from the relationship $\lim\limits_{t\to\infty} F_T(t) < 1$. Indeed, we can see that

$$\int\limits_0^\infty dF_\Lambda \exp[-\lambda C(t)] = L_\Lambda[C(t)]$$

is the Laplace transformation of Λ, which has the property $L_\Lambda(1) > 0$. Then

$$\lim_{t\to\infty} F_T(t) = 1 - \lim_{t\to\infty} p(X_t = 0) = 1 - \lim_{t\to\infty} L_\Lambda(C(t)) = 1 - L_\Lambda(1) < 1.$$

The model developed above can be used for obtaining the nth citation distribution [125]. The result for the nth citation distribution is

$$F_n(t) = p(T_n < t) = \int_0^{C(t)} ds \, \frac{s^{n-1}}{(n-1)!} E_\Lambda[\Lambda^n \exp(-\Lambda s)]; \, t < \infty, \tag{5.90}$$

$$p(T_n = \infty) = \int_1^{\infty} ds \, \frac{s^{n-1}}{(n-1)!} E_\Lambda[\Lambda^n \exp(-\Lambda s)]. \tag{5.91}$$

5.7.2 *Mixed Poisson Model of Papers Published in a Journal Volume*

The accumulation of citations has varying dynamic behavior over the lifetime of a paper, and among other things, this behavior is also influenced by the reputation of the journal in which the paper was published. In most cases, immediately after publication, the number of citations grows slowly, usually because it may take some time for citing papers to appear in print and to be entered in the citations databases. After this initial period, citations increase faster as citations lead to new readers who may also cite the publication. Finally, the material of the paper becomes outdated and/or obsolete. Then the number of citations per year decreases. This is the typical behavior, but there exist other patterns of behavior such as "sleeping beauties," "shooting stars," etc. [126, 127].

The investigation of citation behavior in journal volumes can be based on the mixed Poisson distribution [128–131] model of Burrell [115, 125]. A journal volume can be treated as a collections of paper, usually from the same years and with common characteristics. The main assumption is that each paper generates citations at a constant (latent) rate (λ) following the Poisson distribution but that these rates vary across the collection as a random variable Λ. Then the probability that a paper will generate r citations at time t is

$$p(Z_t = r \mid \Lambda = \lambda) = \exp(-\lambda t) \frac{(\lambda t)^r}{r!}. \tag{5.92}$$

The population distribution of randomly chosen papers of unknown λ will be a mixture of the Poisson distributions of the kind (5.92),

$$p(X_t = r \mid \Lambda) = \int_0^{\infty} dF(\lambda) \exp(-\lambda t) \frac{(\lambda t)^r}{r!}, \tag{5.93}$$

where $F_\Lambda(\lambda)$ is the cumulative distribution of λ (of the latent rate), also called the mixing distribution.

There are different possibilities for the form of mixing distribution [132–134], but the most widely used distribution is the gamma distribution of shape parameter ν and size α:

$$\frac{d}{d\lambda}F_\Lambda(\lambda) = \exp(-\alpha\lambda).\frac{\alpha^\nu \lambda^{\nu-1}}{\Gamma(\nu)} \tag{5.94}$$

The appearance of the gamma distribution above is not a coincidence. The gamma mixture of Poisson distributions follows a negative binomial distribution [135–137] (a fact proved by Greenwood and Yule [138]). Yule is the same scientist who first described the preferential attachment process (Yule process). This negative binomial distribution is

$$P(X_t = r) = \binom{r+\nu-1}{\nu-1}\left(\frac{\alpha}{\alpha+t}\right)^\nu\left(1-\frac{\alpha}{\alpha+t}\right)^r, \quad r = 0, 1, 2, \ldots. \tag{5.95}$$

In most cases, citations of a paper do not occur at constant intervals (evenly) in time. Thus in most cases, λ is not a constant. The rate $\lambda(t)$ will be different for different papers. It can be assumed [115] that $\lambda(t)$ may be written in the form

$$\lambda(t) = \lambda c(t), \tag{5.96}$$

where $c(t)$ describes some pattern of citation behavior that is the same for all articles from the discussed collection of articles (i.e., $c(t)$ describes a sort of obsolescence). The function $c(t)$ is the probability density function of obsolescence, and $C(t)$ is the cumulative distribution function of obsolescence.

With the obsolescence distribution, the model discussed above leads to the following negative binomial distribution for the probability that a paper in a collection of papers will have r citations [139]:

$$p(X_r = r) = \binom{r+\nu-1}{\nu-1}\left(\frac{\alpha}{\alpha+C(t)}\right)^\nu\left(1-\frac{\alpha}{\alpha+C(t)}\right)^r, \quad r = 0, 1, 2, \ldots. \tag{5.97}$$

Many assumptions can be made about the form of $C(t)$. Two possibilities are as follows:

- Logistic function: $C(t) = 1/(1 + a\exp(-bt))$;
- Weibull distribution: $C(t) = 1 - \exp[-(t/b)^2]$.

The values of $C(t)$ can be determined by fitting citation data. Additional information about the investigation of citations in several research disciplines can be found in [140], where a Poisson distribution and an exponential distribution are used for describing such data.

5.8 Aging of Scientific Information

As a consequence of the continuous research efforts of scientists, a continuous flow of new scientific information exists, and existing scientific information ages. As a consequence of these two processes, there is a continuous reorganization of the structure of scientific information. For example, suppose a scientist publishes an article. At first, interest in the article may be significant (a large number of citations, for example). Then interest decreases as the information in the article ages and the scientific potential of the obtained results decreases. If one studies closely the number of citations of a publication, three periods can usually be distinguished:

1. **First two years after publication**: with rare exceptions, articles are not cited much in this period (they are not very well known to the corresponding scientific community). The exceptions are extremely important, however: if an article is very much cited within this initial period, it is highly probable that it will become a very influential publication that may contribute much to the development of the corresponding scientific field.
2. **Next five years**: here the publication achieves most of its citations as it becomes well known. If there are no citations, the publication has been judged by the corresponding scientific community to be of little use. This judgment is valid in the general case, but there can be rare exceptions: "sleeping beauties" that suddenly become current many years after publication [141].
3. **More than seven years after publication**: the number of citations usually begins to decrease, and the publication slowly moves toward the scientific archives.

The above considerations show that by their continuous work in obtaining new knowledge, researchers continuously renew the structure of scientific information by opening a place for the new information and compressing the aged information (this information, compressed to citations, arises in some of the new publications). In this process, researchers mainly use the achievements of the previous generation of researchers.

5.8.1 Death Stochastic Process Model of Aging of Scientific Information

The main assumptions of the model are as follows [142]:

1. At the initial moment of the study, there is some portion of the scientific publications that are cited. The number of citations of these publications $x(t)$ decreases with advancing time. The number of citations at $t = 0$ is x_0.

2. The probability that in the interval $(t, t + \Delta t)$ there will be $x - 1$ citations if in the previous interval the number of citations was x is $\mu_x \Delta t$. Thus the probability that the number of citations will not decrease is $1 - \mu_x \Delta t$.

Then the probability that at the moment $t + \Delta t$ there will be x citations of the scientific publications is

$$p_x(t + \Delta t) = (1 - \mu_x \Delta t) p_x(t) + \mu_{x+1} p_{x+1}(t) \Delta t. \tag{5.98}$$

On the basis of the assumption that the intensity with which citations are decreasing is proportional to the number of citations, $\mu_x = \mu x$, and imposing the initial condition $p_x(0) = 1$ for $x = x_0 \geq 1$ and $p_x(0) = 0$ for $x \neq x_0$, one obtains the following solution of (5.98):

$$p_x(t) = \frac{x_0!}{x!(x_0 - x)!} \exp(-\mu x_0 t)[\exp(\mu t) - 1]^{x - x_0} \tag{5.99}$$

for $0 \leq x \leq x_0$. From (5.99), the average number of citations with advancing time is

$$x_t = x_0 \exp(-\mu t), \tag{5.100}$$

which means that

there occurs an aging of scientific information according to an exponential law. This is a rapid pace of aging, and significant scientific efforts are needed in order to compensate it by production of new scientific information.

5.8.2 *Inhomogeneous Birth Process Model of Aging of Scientific Information. Waring Distribution*

Another approach to the aging of scientific information was proposed by Schubert and Glänzel [143] and discussed by Schubert, Glänzel, and Schoepflin [4, 144]. As we shall see below, the model of Schubert and Glänzel is quite interesting, because it is a deterministic one, yet it is connected to the (not much known but very interesting) Waring distribution. We shall see in addition that this model (that can be connected to an inhomogeneous birth process) leads to the same results as the model discussed above that is based on a death process. And the Waring distribution will be of great interest to us, since it is a generalization of several important statistical distributions appearing in the area of research on science dynamics and research production. Below, we describe a simple model that leads to the Waring distribution. Then we

consider a particular case of a stochastic process connected to repetitive events, and finally, we shall consider a particular class of the process with repetitive events (such as publishing papers and obtaining citations), and we shall consider the aging of scientific information (scientific articles) from the point of view of obtained citations.

5.8.2.1 Waring Distribution

The Waring distribution is a distribution with a very long tail. Because of this property, the Waring distribution is quite suitable for describing characteristics of many systems from the areas connected to research on biology and society. We shall see below that the Waring distribution is connected to other interesting distributions that are presented in this book: the Yule and Zipf distributions.

The Waring distribution may be connected to publication activity, and publication activity may be considered a measure of research productivity. Within the context of the epidemic model of Goffman and Newill (discussed above), the susceptible and infected persons have to be continuously replaced by persons entering the system, i.e., the population of researchers should be considered an open population. As we shall see below, the model by Schubert and Glänzel [143] describes similar processes connected to publication activity. The model assumes three groups in the population: a group that is entering the system, a group that is in the system, and a group that is leaving the system. In more detail, we consider an infinite array of cells (boxes) indexed in succession by nonnegative integers. The amount x of some substance can move between the cells. Let x_i be the amount of the substance in the ith cell. Then

$$x = \sum_{i=0}^{\infty} x_i. \tag{5.101}$$

The fractions $y_i = x_i/x$ can be considered probability values of a distribution of a discrete random variable ζ:

$$y_i = p(\zeta = i),\ i = 0, 1, \ldots. \tag{5.102}$$

We assume that the expected value of the random variable ζ is finite and that the content x_i of any cell can change under any of the following three processes:

1. Some amount s of the substance x may enter the system of cells from the external environment through the 0th cell.
2. The rate f_i of the substance x can be transferred from the ith cell into the $(i + 1)$th cell;
3. The rate g_i from the substance x may leak out of the ith cell into the external environment.

The stochastic process connected to the movement of the substance between the cells is formed by a change in the content of the cells, e.g., by a change of papers

published by authors who have entered the system. In this case, $x(t)$ is the (random) number of published papers, and $p(x(t) = i) = y_i$ the probability that an author in the system has published i papers in the period t. The stochastic model is obtained if $x(t)$ is considered the publication activity process of an *arbitrary* author, and $p(x(t) = i) = y_i$ is the probability that this author has published i papers in the time interval between 0 and t.

The three processes mentioned above can be modeled mathematically by a system of ordinary differential equations:

$$\begin{aligned} \frac{dx_0}{dt} &= s - f_0 - g_0; \\ \frac{dx_i}{dt} &= f_{i-1} - f_i - g_i. \end{aligned} \tag{5.103}$$

The following forms of the relationships for the amount of the moving substances are assumed in [143] ($\alpha, \beta, \gamma, \sigma$ are constants):

$$\begin{aligned} &s = \sigma x; \quad \sigma > 0 \rightarrow \text{self-reproducing property}, \\ &f_i = (\alpha + \beta i)x_i; \quad \alpha > 0, \ \beta \geq 0 \rightarrow \text{cumulative advantage of higher cells}, \\ &g_i = \gamma x_i; \quad \gamma \geq 0 \rightarrow \text{uniform leakage over the cells}. \end{aligned} \tag{5.104}$$

Substitution of (5.104) in (5.103) leads to the relationships

$$\begin{aligned} \frac{dx_0}{dt} &= \sigma x - \alpha x_0 - \gamma x_0; \\ \frac{dx_i}{dt} &= [\alpha + \beta(i-1)]x_{i-1} - (\alpha + \beta i + \gamma)x_i. \end{aligned} \tag{5.105}$$

Let us sum the equations from (5.105). The result of the summation is

$$\frac{dx}{dt} = (\sigma - \gamma)x, \tag{5.106}$$

and the solution for x is

$$x = x(0)\exp[(\sigma - \gamma)t], \tag{5.107}$$

where $x(0)$ is the amount of x at $t = 0$. Three regimes of change of $x(t)$ follow from (5.107):

1. Regime of exponential growth ($\sigma > \gamma$).
2. Stationary regime ($\sigma = \gamma$).
3. Regime of exponential decay ($\sigma < \gamma$).

The distribution of y_i will lead us to the Waring distribution. From (5.105) and with the help of (5.107) and the relationship $\frac{dy_i}{dt} = \frac{1}{x^2}\left[x\frac{dx_i}{dt} - x_i\frac{dx}{dt}\right]$, one obtains

$$\frac{dy_0}{dt} = \sigma - (\alpha + \sigma)y_0;$$
$$\frac{dy_i}{dt} = [\alpha + \beta(i-1)]y_{i-1} - (\alpha + \beta i + \sigma)y_i. \tag{5.108}$$

The solution of (5.108) is

$$y_i = y_i^* + \sum_{j=0}^{i} b_{ij} \exp[-(\alpha + \beta j + \sigma)t], \tag{5.109}$$

where y_i^* is the stationary solution of (5.109) given by the relationships

$$y_0^* = \frac{\sigma}{\sigma + a},$$
$$y_i^* = \frac{\alpha + \beta(i-1)}{\alpha + \beta i + \sigma} y_{i-1}^*, \ i = 1, 2, \ldots. \tag{5.110}$$

The coefficients b_{ij} are determined by the initial conditions. In the exponential function there are no negative coefficients, and because of this, when $t \to \infty$, the sum in (5.109) vanishes and the system comes to the stationary distribution from (5.110). Thus the distribution of y_i tends to be stationary despite the fact that the system is in a stationary state only when $\sigma = \gamma$.

Thus starting from any initial distribution, after some time, the system reaches the steady state, where the content of each cell decays exponentially with (the same) characteristic time $\frac{1}{\sigma - \gamma}$ and the distribution of the substance among the cells is given by (5.110).

This distribution is called the Waring distribution.

The form of the Waring distribution is

$$P(\zeta = i) = \frac{ak^{[i]}}{(a+k)^{[i+1]}}; \quad k^{[i]} = \frac{(k+i)!}{k!}, \tag{5.111}$$

with parameters $k = \alpha/\beta$ and $a = \sigma/\beta$.

We note that the words "after some time" above mean that the Waring distribution can be considered a good approximation of the considered process for large enough finite times when the stationary state of distribution of substance among the cells has almost been reached.

5.8.2.2 Parameters and Particular Cases of the Waring Distribution

The Waring distribution is quite interesting, since it contains as particular cases the distributions of Yule and Zipf.

Let $a > 2$. The expected value of the Waring distribution is

$$E[\zeta] = \frac{k}{a-1}; \quad a > 1. \tag{5.112}$$

We note that $a > 1$ is a condition for a finite expected value (such a finite value was assumed above). Then from the definition of a, it follows that $\sigma > \beta$.

The variance of the Waring distribution is

$$D^2[\zeta] = \frac{ka(k+a-1)}{(a-1)^2(a-2)}; \quad a > 2. \tag{5.113}$$

Several special cases of the Waring distribution are

1. $\beta = 0$ (*geometric distribution*).
   In this case (called also the model of Frank and Coleman [145, 146] or case with absence of cumulative advantage because of $f_i = \alpha x_i$),

$$P(\zeta = i) = q(1-q)^i; \quad q = \frac{\sigma}{\sigma + a}. \tag{5.114}$$

2. $k = 0, \alpha = 0, \beta \neq 0$ (*Yule distribution*).
   Let then $k \to 0$. The Waring distribution reduces to the Yule distribution [147],

$$P(\zeta = i \mid \zeta > 0) = aB(a+1, i), \tag{5.115}$$

   where B is the beta function. Let us note that in this case, $f_i = \beta i x_i$, which is known also as Gibrath law, much used in economics for describing size distributions of business systems [148] or size distributions of cities [149].
3. $i \to \infty$ (*Zipf distribution*).
   As $i \to \infty$, the Waring distribution becomes

$$P(\zeta = i) \to \frac{c}{i^{(1+a)}}, \tag{5.116}$$

   which is the frequency form of the Zipf distribution (c is an appropriate constant depending on the parameters of the distribution).

5.8.2.3 Truncated Waring Distribution

For some applications, one may need a model with a finite number of cells. In this case, we consider an array of $N+1$ cells (boxes) indexed in succession by nonnegative integers, i.e., the first cell has index 0, and the last cell has index N. We assume that there exists an amount x of some substance that is distributed among the cells. Let x_i be the amount of the substance in the ith cell. Then

$$x = \sum_{i=0}^{N} x_i. \tag{5.117}$$

The fractions $y_i = x_i/x$ can be considered probability values of the distribution of a discrete random variable ζ,

$$y_i = p(\zeta = i),\ i = 0, 1, \ldots, N. \tag{5.118}$$

The process of transfer of substance between the cells can be modeled mathematically by a system of ordinary differential equations:

$$\begin{aligned}
\frac{dx_0}{dt} &= s - f_0 - g_0;\\
\frac{dx_i}{dt} &= f_{i-1} - f_i - g_i,\ i = 1, 2, \ldots, N-1;\\
\frac{dx_N}{dt} &= f_{N-1} - g_N.
\end{aligned} \tag{5.119}$$

The forms of the amounts of the moving substances are the same as in (5.104). The substitution of (5.104) in (5.119) leads to the relationships

$$\begin{aligned}
\frac{dx_0}{dt} &= \sigma x - \alpha x_0 - \gamma x_0;\\
\frac{dx_i}{dt} &= [\alpha + \beta(i-1)]x_{i-1} - (\alpha + \beta i + \gamma)x_i,\ i = 1, 2, \ldots, N-1,\\
\frac{dx_N}{dt} &= [\alpha + \beta(N-1)]x_{N-1} - \gamma x_N.
\end{aligned} \tag{5.120}$$

Let us now derive the distribution of y_i. From (5.120), we obtain

$$\begin{aligned}
\frac{dy_0}{dt} &= \sigma - (\alpha + \sigma)y_0;\\
\frac{dy_i}{dt} &= [\alpha + \beta(i-1)]y_{i-1} - (\alpha + \beta i + \sigma)y_i,\ i = 1, 2, \ldots, N-1;\\
\frac{dy_N}{dt} &= [\alpha + \beta(N-1)]y_{N-1} - \sigma y_N.
\end{aligned} \tag{5.121}$$

We search for a solution of (5.121) in the form

$$y_i = y_i^* + F_i(t), \tag{5.122}$$

where y_i^* is the stationary solution of (5.122) given by the relationships

$$\begin{aligned} y_0^* &= \frac{\sigma}{\sigma+\alpha}; \\ y_i^* &= \frac{\alpha+\beta(i-1)}{\alpha+\beta i+\sigma} y_{i-1}^*,\ i = 1, 2, \ldots, N-1; \\ y_N^* &= \frac{\alpha+\beta(N-1)}{\sigma} y_{N-1}^*. \end{aligned} \tag{5.123}$$

For the functions F_i, we obtain the system of equations

$$\begin{aligned} \frac{dF_0}{dt} &= -(\alpha+\sigma)F_0; \\ \frac{dF_i}{dt} &= [\alpha+\beta(i-1)]F_{i-1} - (\alpha+\beta i+\sigma)F_i,\ i = 1, 2, \ldots, N-1, \\ \frac{dF_N}{dt} &= [\alpha+\beta(N-1)]F_{N-1} - \sigma F_N. \end{aligned} \tag{5.124}$$

The solutions of these equations are

$$F_0(t) = b_{00} \exp[-(\alpha+\sigma)t], \tag{5.125}$$

$$F_1(t) = b_{10} \exp[-(\alpha+\sigma)t] + b_{11} \exp[-(\alpha+\beta+\sigma)t], \tag{5.126}$$

$$\ldots$$

$$F_i(t) = \sum_{j=0}^{i} b_{ij} \exp[-(\alpha+\beta j+\sigma)t];\ i = 1, 2, \ldots, N-1, \tag{5.127}$$

$$F_N(t) = \sum_{j=0}^{N} b_{Nj} \exp[-(\alpha+\beta j+\sigma)t], \tag{5.128}$$

where

$$\begin{aligned} b_{ij} &= \frac{\alpha+\beta(i-1)}{\beta(i-j)} b_{i-1,j};\ i = 1, \ldots, N-1;\ j = 0, \ldots, i-1; \\ b_{Nj} &= -\frac{\alpha+\beta(N-1)}{\alpha+j\beta} b_{N-1,j},\ j = 0, \ldots, N-1; \\ b_{NN} &= 0. \end{aligned} \tag{5.129}$$

The b_{ij} that are not determined by (5.129) may be determined by the initial conditions. In the exponential function in $F_i(t)$ there are no negative coefficients, and because of this, as $t \to \infty$, we have $F_i(t) \to 0$, and the system comes to the stationary distribution from (5.123). The form of this stationary distribution is

$$P(\zeta = i) = \frac{a}{a+k}\frac{(k-1)^{[i]}}{(a+k)^{[i]}};\ \ k^{[i]} = \frac{(k+i)!}{k!};\ \ i = 0, \ldots, N-1,$$
$$P(\zeta = N) = \frac{1}{a+k}\frac{(k-1)^{[N]}}{(a+k)^{[N-1]}}, \tag{5.130}$$

with parameters $k = \alpha/\beta$ and $a = \sigma/\beta$.

The obtained distribution is called the truncated Waring distribution. The distribution (5.130) has a concentration of substance in the last cell (i.e., in the Nth cell). For the case of the nontruncated Waring distribution, the same substance is distributed in the cells $N, N+1, \ldots$.

5.8.2.4 A Nonstationary Birth Process. Negative Binomial Distribution, Papers, and Citations

Let us consider the nontruncated version of the Waring distribution. In addition, let us assume that the system is completely isolated from external influences. This means that no substance enters or leaves the system. Thus the amounts of the moving substances are

$$\sigma = 0;\ g_i = 0;\ f_i = (\alpha + \beta i)x_i;\ \ \frac{\alpha(t)}{\beta(t)} = N > 0. \tag{5.131}$$

The last of the above relationships shows that the process is nonstationary (since the substance flow can depend on time). The governing equations become

$$\frac{dy_0}{dt} = -\beta(t)Ny_0;$$
$$\frac{dy_i}{dt} = \beta(t)[(N+i-1)y_{i-1} - (N+1)y_i]; \tag{5.132}$$

with initial conditions $y_i(0) = 1$ if $i = 0$ and $y_i(0) = 0$ otherwise. What one needs is to obtain the distribution $y_i = p(x(t) = i)$ connected to the process. We recall that $p(x(t) = i)$ is the probability that an author in a system has published i papers in the period t. This distribution can be obtained from (5.132), and its form is very similar to the form of the distribution obtained on the basis of the model of death process above [4]:

$$p(x(t) = k) = \binom{N+k-1}{k} \exp[-N\rho(t)]\{1 - \exp[-\rho(t)]\}^k, \tag{5.133}$$

where $\rho(t) = \int_0^t d\tau\ \beta(\tau)$. Equation (5.133) is the relationship for the *negative binomial distribution*. In addition to the probability $p(x(t))$, one can define also transition probabilities $p_{i,k}(s,t)$ for the probability that at time t, the substance is in the kth unit if at time $s < t$ it was in the ith unit. From the point of view of the case with scientists and articles, $p_{ik}(s,t)$ is the probability that an author will own k articles at time t if at time s he/she owns $i \leq k$ articles. In this case, the evolution of the transition probability [144] is given by

$$\frac{\partial p_{i,k}(s,t)}{\partial t} = \beta(t)[(N+k-1)p_{i,k-1}(s,t) - (N+k)p_{i,k}(s,t)], \tag{5.134}$$

with initial conditions $p_{i,k}(s,s) = 1$ if $k = i$ and $p_{i,k}(s,s) = 1$ otherwise.

Citations are repetitive events exactly like papers. Thus all discussions about the nonstationary birth process connected to papers are the same for the nonstationary birth process connected to citations. In the first case, we have a scientist who publishes papers. In the second case, we have a paper that receives citations. Then (5.133) gives the probability that a paper will have received k citations at time t, and (5.134) gives the transitional probability that a paper will have received k citation at time t if it has i citations at the time s. The distribution connected to the transitional probability $p_{i,k}$ is also a negative binomial distribution. In more detail, the number of received citations for the time $t - s$ when the number of received citations at until time s was i, $p_{i,j}(s,t) = p[x(t) - x(s) = j \mid x(s) = i]$, is

$$p_{i,j}(s,t) = \binom{N+i+j-1}{j} \exp\{-[\rho(t) - \rho(s)](N+i)\}(1 - \exp\{-[\rho(t) - \rho(s)]\})^j, \tag{5.135}$$

i.e., the substance flow during the time period $t - s$ has a negative binomial distribution with parameters $\exp[-r(t) + r(s)]$ and $N + j$, where j is the index of the unit that was reached by the substance at time s [143, 144, 150].

With respect to the aging of scientific information, it is important to study the mean value function $M_i(s,t)$. It will show us that a paper that has received some number of citations during the time s after its publication is expected to receive (during an arbitrary time period $t - s$ after the moment s) a linear expression in what it had received previously:

$$M_i(s,t) = E[x(t) - x(s) \mid x(s) = i] = (N+i)\{\exp[\rho(t) - \rho(s)] - 1\} = c_s(t)i + d_s(t). \tag{5.136}$$

We note that $\frac{d_s(t)}{c_s(t)} = N = \text{const}$ is independent of time, and $c_s(t)$ is a characteristic of the aging process. Large $c_s(t)$ characterizes slowly aging literature.

Let us define

$$M(s,t) = E[x(t) - x(s)] = N \exp[\rho(t) - \rho(s)] \tag{5.137}$$

and

$$q(s,t) = \frac{E[x(s) + N]}{E[x(t) + N]}. \tag{5.138}$$

Then (5.135) can be written as

$$p_{i,j}(s,t) = \binom{N+i+j-1}{j} q(s,t)^{N+i}[1 - q(s,t)]^j, \tag{5.139}$$

and the expected citation rate during the time period $t - s$ under the condition that the corresponding paper has received i citations during the time span s is

$$M_i(s,t) = (N+i)\frac{E[x(t) - x(s)]}{E[x(s)] + N}. \tag{5.140}$$

Finally, from (5.139), one obtains that the probability that an article that has received $i \geq 0$ citations will no longer be cited is

$$p_{i,0}(s,t) = p[x(t) - x(s) = 0 \mid x(s) = i] = q(s,t)^{N+i}. \tag{5.141}$$

The lifetime distribution of a process $\{X(t)\}$ is defined by

$$F(t) = \frac{M(0,t)}{M(0,\infty)}, \quad t \geq 0. \tag{5.142}$$

Let us choose the following particular form of f_i [151]:

$$f_i = (N+i)\alpha^*\beta^* \exp(-\alpha^* t)x_i = \beta^* N(1 + i/N)\alpha^* \exp(-\alpha^* t),\ N > 0,\ \alpha^* > 0, \beta^* > 0. \tag{5.143}$$

The time-invariant part of f_i is proportional to $1 = i/N$, and because of this, increases by transfer from the ith cell to the $(i+1)$th cell (which can be considered a local form reflection of the cumulative advantage principle). The time-dependent component of f_i reflects the local exponential aging of the process (aging of the content relative to an individual unit). Then

$$M(s,t) = N\{\exp[\beta^*(1 - \exp(-\alpha^* t)) - \exp[\beta^*(1 - \exp(\alpha^* s))]]\} \tag{5.144}$$

and

$$F(t) = \frac{\exp[\beta^*(1 - \exp[-\alpha^* t])] - 1}{\exp(\beta^*) - 1}. \tag{5.145}$$

Finally, let us discuss the particular cases in which the model describes articles that obtain citations. One can define the *obsolescence function* $H(s)$: the probability that a paper will not be cited beyond a given time s. The definition is

$$H(s) = p(x(\infty) - x(s) = 0). \tag{5.146}$$

The obsolescence function for our particular case is

$$H(s) = \{1 + \exp(\beta^*) - \exp[\beta^*(1 - \exp(-\alpha^* s))]\}^{-N}. \tag{5.147}$$

We note that $H(\infty) = 1$, i.e., at infinity, every publication is obsolete. We have $H(0) = \exp(-\beta^* N)$, i.e., the probability that a paper is already obsolete at the moment it is published equals the probability that it will never be cited.

5.8.2.5 A Case of Brain Drain: Migration Channel for Research Personnel

Let us now discuss one application of the truncated Waring distribution. We consider a sequence of $N + 1$ countries that form a channel. As a result of a large migration movement, a flow of researchers moves through this channel from the country of entrance to the final destination country that is attractive to them in terms of good conditions for life and work. We may assume a situation of war in some region and motion of a large group of researchers from that region to another (more attractive region). The motion starts from an entry country, and the researchers have to move through a sequence of countries in order to reach a (very attractive from the point of view of the researchers) final destination country. We may think about the sequence of countries as a sequence of boxes (cells). The entry country will be the box with label 0, and the final destination country will be the box with label N. Let us consider a number x of researchers that have entered the channel and are distributed among the countries. Let x_i be the number of researchers in the ith country. This number can change on the basis of the following three processes: (a) A number s of researchers enter the channel from the external environment through the country of entrance (0th cell); (b) A number f_i of researchers move from the ith country to the $(i + 1)$th country; (c) A number g_i of the researchers change their status (e.g., they do not move farther in the direction of the final destination country and they are no longer active in the field of research). Fot the case of a large number of migrating researchers, the values of x_i can be determined by (5.103). The relationships (5.104) mean that (a) the number of researchers s that enter the channel is proportional to the number of researchers in all countries that form the channel; (b) there may be a preference for some countries, e.g., migrants may prefer the countries that are around the end of the migration channel (and the final destination country may be the most preferred one); (c) it is assumed that the conditions along the channel are the same with respect to "leakage" of researchers, e.g., the same proportion γ of researchers move out of the area of research work in every country of the channel.

As can be seen from (5.107), the change in the number of researchers depends on the values of σ and γ. If $\sigma > \gamma$, the number of researchers in the channel increases exponentially. If $\sigma < \gamma$, the number of researchers in the channel decreases exponentially. The dynamics of the distribution of the researchers in the channel is modeled by (5.108). When the time since the beginning of the operation of the channel become large enough, the distribution of the researchers in the countries that form the migration channel becomes close to the stationary distribution described by (5.110). Let us stress that the stationary distribution described by (5.110) is very similar to the Waring distribution, but there is a significant difference between the two distributions due to the finite length of the migration channel: there may be a large concentration of researchers in the final destination country especially, if this country is very attractive for researchers.

The parameters that govern the distribution of researchers in the countries that form the channels are σ, α, β, and γ. The parameter σ is the "gate" parameter, since it regulates the number of researchers that enter the channel. If σ is large, then the number of researchers in the channel may increase very rapidly, and this can lead to problems in the corresponding countries. We note that σ participates in each term of the truncated Waring distribution. This means that the situation at the entrance of the migration channel influences significantly the distribution of researchers in the countries of the channel.

The parameter γ regulates the "absorption" of the channel, since it regulates the change of the status of some researchers. They may settle in the corresponding country and may accept a job that is out of the area connected to their research. A large value of γ may compensate for the value of σ and may even lead to a decrease in the number of researchers in the channel. The parameter α regulates the motion of the researchers from one country to the next country of the channel. A small value of α means that the researchers tend to concentrate in the entry country (and eventually in the second country of the channel). An increase in α leads to an increase in the proportion of researchers that reach the second half of the migration channel and especially the final destination country.

The parameter β regulates the attractiveness of the countries along the channel. Large values of β mean that the final destination country is very attractive to researchers (e.g., has excellent conditions for work and the salaries are large). This increases the attractiveness of the countries in the second half of the channel (researchers are more desirous of reaching these countries because the distance to the final destination country is thereby decreased). If for some reason β is kept at a high value, then almost all the researchers may settle in the final destination country.

5.8.2.6 Multivariate Waring Distribution

One can define the multivariate Waring distribution as follows [152]. Let a and b be positive real numbers. Let $a^{(k)} = \frac{\Gamma(a+b)}{\Gamma(a)}$, where $\Gamma(x) = \int\limits_0^\infty dt\ \exp(-t)t^{x-1}$ is the

gamma function [153]. Let $p(x_1 = k_1, \ldots, x_n = k_n; a, b_1, \ldots, b_n)$ be the probability that $x_1 = k_1, \ldots, x_n = k_n$ with parameters $a, b_1, \ldots, b_n$. The multivariate Waring distribution is given by the relationship

$$p(x_1 = k_1, \ldots, x_n = k_n; a, b_1, \ldots, b_n) = a \frac{\Gamma\left(\sum_{i=1}^{n} k_i - n + 1\right) \Gamma\left(\sum_{i=1}^{n} b_i + a\right)}{\Gamma\left(\sum_{i=1}^{n} k_i + \sum_{i=1}^{n} b_i - n + a + 1\right)} \prod_{i=1}^{n} \frac{\Gamma(k_i + b_i - 1)}{\Gamma(k_i)\Gamma(b_i)}, \tag{5.148}$$

where $k_i = 1, 2, \ldots$ and $i = 1, \ldots, n$, a and b_i are positive real numbers. For $n = 1$, the multivariate Waring distribution is reduced to the univariate Waring distribution

$$p(x = k; a, b) = a \frac{\Gamma(b + k + 1)\Gamma(a + b)}{\Gamma(b)\Gamma(a + b + k)}. \tag{5.149}$$

Let $a^{(b)} = \frac{\Gamma(a+b)}{\Gamma(a)}$. Then the univariate form of the Waring distribution can be written as

$$p(x = k; a, b) = a \frac{b^{(k-1)}}{(a + b)^{(k)}}. \tag{5.150}$$

Two interesting properties of the multivariate Waring distribution are as follows:

1. Let the multivariate random variable $(x_1, \ldots, x_n)$ follow the multivariate Waring distribution (5.148). Then the corresponding expected value is

$$E(x_1, \ldots, x_n) = a \int_0^1 dx\, (1 - x)^{a-n-1} \prod_{i=1}^{n} (1 - x + b_i x). \tag{5.151}$$

2. Every marginal distribution of the multivariate Waring distribution is also a Waring distribution

$$\sum_{k_s=1}^{\infty} \cdots \sum_{k_n=1}^{\infty} p(x_1 = k_1, \ldots, x_s = k_s, x_{s+1} = k_{s+1}, \ldots, x_n = k_n; a, b_1, \ldots, b_n) = p(x_1 = k_1, \ldots, x_s = k_s; a, b_1, \ldots, b_n). \tag{5.152}$$

The simplest case of the multivariate Waring distribution is the bivariate Waring distribution

$$p(x = k, y = j; a, b, c) = a \frac{(k + j - 2)! b^{(k-1)} c^{(j-1)}}{(a + b + c)^{(k+j-1)(k-1)!(j-1)!}}, \tag{5.153}$$

with expected value

$$E(x, y) = 1 + \frac{b+c}{a-1} + \frac{2bc}{(a-1)(a-2)} \tag{5.154}$$

and covariance

$$\mathrm{Cov}(x, y) = 1 + \frac{b+c}{a-1} + \frac{2bc}{(a-1)(a-2)} - \left(1 + \frac{b}{a-1}\right)\left(1 + \frac{c}{a-1}\right). \tag{5.155}$$

If (x, y) follows the bivariate Waring distribution, then the conditional probability $p(x = k \mid y = m)$ is

$$p(x = k \mid y = m) = \frac{1}{(k+1)!} \frac{(a+c)^{(b)}}{(a+c+m)^{(b)}} \frac{b^{(k-1)} m^{(k-1)}}{(a+b+c+m)^{(k-1)}}, \tag{5.156}$$

and the conditional expectation $E(x \mid y = m)$ is

$$E(x \mid y = m) = 1 + \frac{b}{a+c-1} m. \tag{5.157}$$

The multivariate Waring distribution was applied to the study of scientific productivity among authors in six main Chinese journals of information science during the three-year periods 1987–1989 and 1990–1992 [152].

5.8.3 *Quantities Connected to the Age of Citations*

After publication of an article, some tame elapses before the article is cited. Let T be the time between publication of the article and the publication of the citing source. In general, T is a random variable, and one can study distributions of the time to the first citation [115], or to the nth citation [125]. Here we mention several quantities connected to the time of first citation (these quantities can be applied also to the time of second citation, etc.) [154]. Let us assume that T is a continuous quantity, and let $f(t)$ be the probability density function of the distribution of T. Then one can define the age-specific citation rate

$$r(t) = -\frac{d}{dt}[\ln R(t)], \tag{5.158}$$

where

$$f(t) = \frac{dR}{dt},$$

and $R(t) = R_T(t) = p(T > t)$ is called the reliability function of T (here $p(T > t)$ means the probability that $T > t$). From (5.158), it follows that

$$R(t) = \exp\left(-\int_0^t dsr(s)\right). \tag{5.159}$$

Assuming different kinds of distributions for $f(t)$, we can obtain the corresponding relationship for the age-specific citation rate. Since citations (in most cases) can be considered rare events, we can use distributions connected to the theory of extreme events, such as the following:

- The exponential distribution $f(t) = \lambda \exp(-\lambda t)$. In this case, $R(t) = \exp(-\lambda t)$ and

$$r(t) = \lambda. \tag{5.160}$$

  Thus a constant age-specific citation rate implies an exponential distribution of the citation age.
- The Weibull distribution of citation age T with shape parameter $\beta > 0$ and scale parameter $\alpha > 0$. Here the reliability function is $R(t) = \exp[-(t/\alpha)^{\beta}]$, and the age-specific citation rate is

$$r(t) = \frac{\beta t^{\beta-1}}{\alpha^{\beta}}. \tag{5.161}$$

5.9 Probability Models Connected to Research Dynamics

5.9.1 Variation Approach to Scientific Production

The occurrence of laws in the form of hyperbolic relationships (such as the laws of Zipf and Pareto, for example) and the persistence of such laws may lead to the following assumption:

> *A research organization is in an equilibrium state with respect to scientific production if the statistical laws for the characteristic quantities of this productivity are given by hyperbolic relationships.*

We can even extend the above assumption by the additional assumption that the parameters of the statistical laws have selected values (for example, $\alpha = 1$) when the research organization is in an equilibrium state. And if the distributions of the quantities are not described by the appropriate hyperbolic relationships, then the research organization (and its structure and system of functioning) may not be in an equilibrium state.

Equilibrium states of various systems may be studied by variational methods [155]. A hint at the possible applicability of a variational approach in the social sciences is connected to George Zipf, who explained what is now known as Zipf's law in the field of linguistics [156] by means of the principle of least effort:

> *Human communication is based on two opposite tendencies: the one who speaks tries to use the minimum number of words, and this one who hears tries to understand the speaker by investing minimal effort.*

Let the effort $E(x)$ of a researcher to produce x publications be proportional to the time he or she invests for research: $E(x) \propto t$. There is a law for an exponentially growing science that states that scientific production growths exponentially with invested time: $x(t) = \exp(\lambda t)$, where λ is a parameter. From here, $t = \frac{1}{\lambda} \ln(x)$ and

$$E(x) \propto \frac{1}{\lambda} \ln(x) = \rho \ln(x). \tag{5.162}$$

This relationship will be introduced in the relationships for the variational principle of Boltzmann below [104, 157].

The principle of maximum entropy (variational principle of Boltzmann) is for systems whose states x are distributed with probability $p(x)$ ($\int dx\, p(x) = 1$). Then at an equilibrium state with energy

$$E = \int dx\, p(x) E(x), \tag{5.163}$$

the entropy

$$H = -\int dx\, p(x) \ln[p(x)] \tag{5.164}$$

has a maximum value.

The function $p(x)$ above is the probability that a researcher has produced x publications, and we shall treat $E(x)$ below as a measure of the mean effort (mean "energy") spent in the course of the scientific work. The solution of the above variational problem is

$$p(x) = (1/Z) \exp[-\lambda^* E(x)] = (1/Z)(1/x^{\rho\lambda^*}), \tag{5.165}$$

where Z is the statistical sum and λ^* is a parameter that can be determined from the normalization condition and the boundary condition.

Here we shall discuss as the least-value state the state $x_0 = 1$ (researchers must have at least one publication). Then

$$E = \int_1^\infty dx\, p(x)E(x) \tag{5.166}$$

and

$$p(x) = (\rho/E)1/(x^{1+\rho/E}) = \alpha/(x^{1+\alpha});\quad \alpha = (\rho/E). \tag{5.167}$$

This is the law of Pareto (called also the Zipf–Pareto law).

The entropy of a system that obeys the law (5.167) is

$$H = -\int_1^\infty dx\, p(x)\ln[p(x)] = 1 + \frac{1}{\alpha} - \ln(\alpha); \tag{5.168}$$

"Temperature": *The analogy with the thermodynamics may be continued: one may introduce a quantity called "temperature." This quantity is a measure of the external influence on the scientific system.*

"Temperature" can be introduced by comparing the results for Lagrange multipliers in statistical mechanics (where $\lambda^* \propto 1/T$) with the case of scientific production (where $\lambda^* = (1+\alpha)/\rho$). Thus the "temperature" is

$$T \propto \frac{\rho}{1+\alpha}. \tag{5.169}$$

Using (5.169), we can write the Zipf–Pareto law (5.167) as

$$p(x) = \frac{\alpha}{x^{k\rho/T}}, \tag{5.170}$$

where k is a coefficient of proportionality. From (5.169), $\alpha = 1 - \frac{k\rho}{T}$, and the final form of the Zipf–Pareto law (5.170) is

$$p(x) = \frac{1 - \frac{k\rho}{T}}{x^{k\rho/T}}. \tag{5.171}$$

There are two parameters in (5.171):

- k: characteristic of the efforts of the researcher in the publication process. These efforts can depend on the talent of the researcher but also on the conditions of work, salary, etc. Increasing research efforts lead to a decreasing value of k.
- T: characteristic of external influence on research organization. The parameter T can be connected to different flows toward the scientific structures (e.g., to money

flows). Then if the money flow increases, the system is "heated," and if the money flow decreases, the system is "frozen."

Let us analyze (5.171). We shall see the role of better work conditions and increased funding in increasing research production.

1. Let us fix the number of publications x. Thus we can study the influence of ρ and T. Let us fix also T (for example, a fixed quantity of money flows to the scientific organization, and other external conditions are fixed). Then a decrease in ρ will increase the numerator of (5.171) and will decrease its denominator. Hence p will increase. *This means that initiatives to decrease the necessary expenditures of effort by researchers in the publication production process (for example, an initiative for better work conditions or better social networking in the research organization) may increase the probability that researcher will have a larger number of publications*.
2. Let us now fix x and ρ and increase T (for example, by increasing the money flow toward the research organization). The numerator of (5.171) increases, and the denominator decreases. Thus p increases, which means that *one can expect that research production will increase with increased funding*.

Finally, let us note that thermodynamic models are also used in other areas of science such as technological forecasting and the theory of manpower systems [158, 159].

The variational approach can also be applied to the case of discrete distributions (e.g., for studying the circulation of documents) [160]. Let us consider a finite probability distribution $P = \{p_1, \ldots, p_n\}$, where $p_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_i = 1^n p_i = 1$. The entropy attached to this probability distribution is

$$H_n(P) = -\sum_{i=1}^{n} p_i \ln(p_i). \tag{5.172}$$

The entropy is a measure of uncertainty. The uncertainty is maximal when the outcomes are equally likely. Since the uniform distribution maximizes the entropy, it contains the largest amount of uncertainty.

Let $X = \{1, \ldots, n\}$ be a random variable and p_i the probability of the occurrence of the value i. We have the constraint

$$\sum_{i=1}^{n} p_i = 1, \tag{5.173}$$

and we impose an additional constraint about the expected value of the distribution X:

$$E(X) = \sum_{i=1}^{n} i p_i = \mu. \tag{5.174}$$

According to the principle of maximum entropy, we have to find the distribution P that maximizes the entropy (5.172) subject to the constraints (5.173) and (5.174). Introducing two Lagrange multipliers α and β, we have to find a maximum for the functional

$$L = H_n(P) - \alpha\left(\sum_{i=1}^{n} p_i - 1\right) - \beta\left(\sum_{i=1}^{n} ip_i - E(X)\right). \tag{5.175}$$

The Euler equations for L from (5.175) are

$$\begin{aligned} \partial L/\partial p_i &= -\ln(p_i) - 1 - \alpha - \beta i; \quad i = 1, \ldots, n, \\ \partial L/\partial \alpha &= 1 - \sum_{i=1}^{n} p_i, \\ \partial L/\partial \beta &= E(X) - \sum_{i=1}^{n} ip_i. \end{aligned} \tag{5.176}$$

The solution of these equations is

$$p_i = \frac{\exp(-\beta_0 i)}{\sum_{i=1}^{n} \exp(-\beta_0 i)}, \tag{5.177}$$

where β_0 is the solution of the equation

$$\sum_{i=1}^{n} [i - E(X)] \exp[-(i - E(X))] = 0. \tag{5.178}$$

A similar calculation can also be made for the case of more than two constraints.

5.9.2 Modeling Production/Citation Process

Joint modeling of production and citation processes in science attracted considerable attention after the introduction of the h-index of Hirsch. Below, we shall consider two models of the processes connected to the h-index.

5.9.2.1 Model of h-Index Based on Paretian Distributions

Discrete Paretian distributions and the Price distribution are distributions that are widely used for modeling publication activity and citation processes [161]. The properties of these distributions needed for investigation of the Hirsch index are

represented by means of Gumbel's characteristic extreme values [162]. The reason for this is that the Hirsch index can be defined on the basis of Gumbel's rth characteristic values.

Gumbel's rth characteristic values are defined as follows. Let us consider a random variable X that gives the citation rate of a paper. We define

- $p_k = P(X = k)$: probability distribution of X $(k \geq 0)$;
- $F(k) = P(X < k)$: cumulative distribution function of X.

Gumbel's rth characteristic extreme value is then defined as

$$u_r = \max\{k : G(k) \geq r/n\}, \tag{5.179}$$

where

- $G(k) = G_k = 1 - F(k) = P(X \geq k)$;
- n: given sample with distribution F.

The Hirsch index can be defined analogously to Gumbel's rth characteristic extreme value as follows:

$$h = u_h. \tag{5.180}$$

5.9.2.2 Case of Paretian Distribution of the Random Variable X

A distribution of a random variable (in our case, the distribution of citations X) is a Paretian distribution if it obeys asymptotically Zipf's law:

$$\lim_{k\to\infty} \frac{G_k}{k^\alpha} \approx \text{const.} \tag{5.181}$$

Below, we shall use a prominent member of the class of Paretian distributions, namely the Pareto distribution $p_k = P(X = k) \approx \frac{d}{(N+k)^{-(1+\alpha)}}$. This distribution is Paretian as $k \to \infty$. For the case $k \gg N$, we obtain

$$G_k = P(X \geq k) \approx \frac{d_1}{k^\alpha}, \tag{5.182}$$

where d_1 is a positive constant. Then

$$u_r \approx c_1 \left(\frac{n}{r}\right)^{1/\alpha}, \tag{5.183}$$

where c_1 is a positive constant. Equation (5.183) leads to the following equation for the Hirsch index (in the presence of the assumption $n \gg 1$):

$$h = u_h \approx c_1 (n/h)^{1/\alpha} . \tag{5.184}$$

From here, we obtain

$$h \approx c_2 n^{1/(1+\alpha)}, \tag{5.185}$$

where $c_2 = c_1^{\alpha/(1+\alpha)}$.

We can draw the following conclusions from (5.185) (note that we work with the assumption that the citation distribution is a discrete Paretian distribution (with finite expectation)).

1. If the number of underlying papers is large enough, then the Hirsch index h is proportional to the $(1 + \alpha)$th root of the number of publications. Usually α is close to 1. Then h is proportional to the square root of the number of publications.
2. The number of citations of the papers from the Hirsch core (which contains the h-papers: papers that received at least h citations each) is proportional to h^2 for $\alpha > 1$ and a large value of k [161].

5.9.2.3 Case of Price Distribution of the Random Variable X

We recall that in our case, the random variable X is the citation rate of a paper. The Price distribution is [163]

$$p_k = P(X = k) = N\left(\frac{1}{N+k} - \frac{1}{N+k+1}\right) = \frac{N}{(N+k)(N+k+1)}, \tag{5.186}$$

where $k \geq 0$ and N is a positive parameter.

Note that N is a positive parameter. Thus N may be a noninteger. In addition, the Price distribution contains the case $k = 0$ as well as the law of Lotka (for research publications) when $k \gg N$. Moreover, no positive moments of the Price distribution exist. The distribution (5.186) is called the Price distribution, since it contains as a limiting case the square root law of Price (*which states that half of the scientific papers are contributed by the top square root of the total number of scientific authors*) [163]. Let us stress that the Price distribution is a particular case (when $\alpha = 1$) of

the Waring distribution [101, 164]

$$p_k = P(X = k) = \frac{\alpha}{N+\alpha}\frac{N}{N+\alpha+1}\ldots,\frac{N+k-1}{N+\alpha+k} \tag{5.187}$$

where $k \geq 0$ and α and N are positive parameters.

For the case in which the distribution of the citation rate is described by the Price distribution, one obtains

$$G_k = \frac{N}{N+k}. \tag{5.188}$$

Thus the distribution is Paretian (but note that the expected value of X for this distribution is ∞, in contrast to the finite expectation connected to the Pareto distribution discussed above).

The Gumbel rth extreme value is

$$u_r = \left[\frac{N(n-r)}{r}\right], \quad r = 1, 2, \ldots, n, \tag{5.189}$$

where $[\ldots]$ denotes the integer part of the corresponding argument.

The corresponding h index is a solution of the equation

$$h = u_h \approx \frac{N(n-r)}{r}. \tag{5.190}$$

The solution (for $n \gg 1$) can be approximated as

$$h = \left(\frac{N^2}{4} + nN\right)^{1/2} - \frac{N}{2} \approx (nN)^{1/2}, \tag{5.191}$$

which means the following:

The h-index is proportional of the square root of the number of publications (if the citation rate is described by the Price distribution and all other assumptions are valid).

5.9.2.4 Model of *h*-Index Based on the Poisson Distribution

Another model of the h-index is based on the publication–citation model of Burrell [165, 166]. This model is for the publishing record of a scientist who publishes papers at certain times. These papers then attract citations, and both the publication and citation accumulation processes are random. The assumption is that the scientist

starts his/her publishing career at $t = 0$, and by the time $T > 0$, one observes the following:

1. *Poisson process of publishing*
   The author publishes papers according to a Poisson process at rate θ. The distribution of the number of publications Y_T at time T is

$$P(Y_T = r) = \exp(-\theta T)\frac{(\theta T)^r}{r!}, \quad r = 1, 2, \ldots, \tag{5.192}$$

   with expected value $E[Y_T] = \theta T$.
2. *Poisson process of citations receiving*
   Each of the publications receives citations according to a Poisson process of rate Λ, which can vary from paper to paper.
3. *Variation of the rate* Λ
   The citation rate Λ varies over the set of publications of the scientist according to a gamma distribution of index $\nu > 1$ and parameter $\alpha > 0$:

$$f_\Lambda(\lambda) = \frac{\alpha^\nu}{\Gamma(\nu)}\lambda^{\nu-1}\exp(-\alpha\lambda), \tag{5.193}$$

   where $0 < \lambda < \infty$.

The model leads to the following distribution of the citations of a randomly chosen paper of the scientist [166]:

$$P(X_T = r) = \frac{\alpha}{T(\nu-1)}B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right), \quad r = 0, 1, 2, \ldots, \tag{5.194}$$

where

$$B(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x dy\, y^{a-1}(1-y)^{b-1}$$

is the cumulative distribution of the beta distribution of the first kind, and a and b are parameters.

What remains to be calculated is $N(n; T)$: the expected number of papers receiving at least n citations by the time T.

- **Case of $n = 0$ citations**

$$E[N(0; T)] = \theta T, \tag{5.195}$$

  i.e., the number of uncited papers of the scientist is expected to have linear increase over time.

- **Case of $n \neq 0$ citations** In this case [166],

$$E[N(n;T)] = \theta T \left[1 - \frac{\alpha}{T(\nu - 1)} \sum_{r=0}^{n-1} B\left(\frac{T}{\alpha + T}; r + 1, \nu - 1 \right) \right], n = 1, 2, \ldots . \tag{5.196}$$

Equation (5.196) has interesting consequences:

1. *Publish or perish!*: The expected number of papers with n citations is proportional of the publication rate θ.
2. *A long career in science is a good thing!*: The expected number of papers with n citations is increasing in T for every n.
3. *No one is a genius!*: The expected number of papers with n citations is decreasing in n for every T.

Finally, the h-index can be defined as

$$h(T) = \max\{n : n \leq E[N(n, T)]\}, \tag{5.197}$$

and as we have seen just above, the h-index depends on the intensity of publication, the length of the scientific career, and other parameters (such as the parameters α and ν of the beta distribution, which can vary from scientist to scientist).

5.9.3 *The GIGP (Generalized Inverse Gaussian–Poisson Distribution): Model Distribution for Bibliometric Data. Relation to Other Bibliometric Distributions*

Up to now, we have discussed several distributions that may be used to model different aspects of research dynamics and to fit bibliometric data. Sichel [167, 168] argues that there exists a distribution that is very suitable for modeling bibliometric data: the GIGP (generalized inverse Gaussian–Poisson) distribution. The GIGP distribution seems to be complicated, but its goodness of the fit with respect to bibliometric data is usually very good. The GIGP distribution may be obtained as follows. Let us consider a researcher who has an average rate of publishing λ_i papers in unit time. Then the expected number of papers published by this researcher for time t will be $\lambda_i t$. Let us assume that the statistical variability around the average $\lambda_i t$ follows a Poisson distribution. If we have a group of researchers, then within this group, the value of λ_i will vary, since some researchers are more productive than others. Let us assume that the values of λ_i are distributed according to a generalized

inverse Gaussian distribution law (called a GIG distribution).[1] Then we arrive at the compound Poisson distribution called GIGP [170]:

$$p(r,t) = \frac{(1-\theta_t)^{\gamma_t/2}}{K_{\gamma_t}[\alpha_t\sqrt{1-\theta_t}]}\frac{(\alpha_t\theta_t)^r}{2^r r!}K_{r+\gamma_t}(\alpha_t), \tag{5.198}$$

where $r = 0, 1, 2, \ldots$; $0 \le \theta_t \le 1$; $-\infty < \gamma_t < \infty$; $\alpha_t \ge 1$; $K_\nu(z)$ is the modified Bessel function of the second kind of order ν; and t is the length of the considered time period. The time-dependent parameters are as follows:

$$\alpha_t = \alpha\sqrt{1+\theta(t-1)};\quad \theta_t = \frac{\theta t}{1+\theta(t-1)};\quad \gamma_t = \gamma. \tag{5.199}$$

From (5.198), one can calculate the probabilities $p(r)$ by means of a recurrence relation as follows if one knows $p(0)$ and $p(1)$ for $r = 0, 1, 2, \ldots$:

$$p(r) = \left(\frac{r+\gamma-1}{r}\right)\theta_t p(r-1) + \frac{\alpha_t^2\theta_t^2}{4r(r-1)}p(r-2). \tag{5.200}$$

The GIGP is also able do describe the domain $r = 1, 2, , 3, \ldots$. For this purpose, one has to perform zero truncation of the distribution from (5.198). The result is

$$p(r,t) = \frac{(\alpha_t\theta_t)^r K_{\gamma+r}(\alpha_t)}{2^r r!\{(1-\theta_t)^{-\gamma/2}K_\gamma[\alpha_t(1-\theta_t)^{1/2}] - K_\gamma(\alpha_t)\}}. \tag{5.201}$$

The GIGP distribution has been used to describe bibliometric data such as the number of articles published in the area of operations research, the scattering of literature in applied geophysics, the literature on mast cells, publications of a group of chemists several years after receiving their doctoral degrees, in-house journal use in libraries, etc. [167].

The GIGP distribution (5.198) has three parameters. If some of these parameters are known a priori, then the GIGP distribution can be reduced to several different distributions. Some examples of such reduction are as follows:

1. *Negative binomial distribution*: $\alpha = 0$; $\gamma > 0$.
2. *Zero-truncated negative binomial distribution*: $\alpha = 0$; $-1 < \gamma < 1$.
3. *Fisher logarithmic series distribution*: $\alpha =$; $\gamma = 0$.
4. *Inverse Gaussian–Poisson (IGP) distribution*: $\gamma = -1/2$; $r = 0, 1, 2, \ldots$.

[1] The form of this distribution may be written as

$$f(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}x^{(p-1)}\exp[-(ax+b/x)/2],$$

where $x > 0$, K_p is the modified Bessel function of the second kind, $a > 0$, $b > 0$, and p is a real parameter [169].

The upper tail (i.e., for large values of r) of the GIGP distribution is given by the following relationship [168]:

$$p(r) \sim \frac{c\theta^r}{r^{1-\gamma}}, \tag{5.202}$$

where c is a normalizing constant, $0 < \theta \leq 1$, and $-\infty < \gamma < \infty$. Taking the logarithm of both sides of (5.202), one can write

$$Y = A - (1 - \gamma)X - B\exp(X), \tag{5.203}$$

where $Y = \ln p(r); X = \ln r; A = \ln C; B = -\ln\theta$. Thus the tail of the GIGP distribution for $\gamma < 1$ is first linear, and then with increasing value of r, it becomes convex. Let $\theta = 1$. Then the tail of the GIGP distribution described by (5.203) becomes linear, and thus the GIGP distribution for this case corresponds to the distributions of Lotka and Zipf discussed in a previous chapter of this book.

5.9.4 Master Equation Model of Scientific Productivity

We know already that productivity is an important element in the evolution of a research community. It is possible to derive an equation that accounts for the stochastic fluctuations in the productivity of the members of a scientific organization [171]. In order to obtain this model equation, we assume that the main processes of evolution of scientific community are these:

1. the self-reproduction of scientists,
2. aging and death of scientists,
3. departure of scientists from the scientific field due to mobility or abandoning research activities.

Let a be the scientific age (number of years devoted to scientific research) of a researcher, and let a scientific productivity index ξ be incorporated into the researcher state space (ξ and a are assumed to be continuous variables with values in $[0, \infty]$). The dynamics of the research community are described by a number density function $n(a, \xi, t)$, which specifies the age and productivity structure of the scientific community at time t. For example, the number of researchers with age in $[a_1, a_2]$ and scientific productivity in $[\xi_1, \xi_2]$ at time t is given by the integral $\int_{a_1}^{a_2}\int_{\xi_1}^{\xi_2} da\, d\xi\, n(a, \xi, t)$.

The following master equation for this function $n(a, \xi, t)$ can be derived [171]:

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n(a, \xi, t) = -[J(a, \xi, t) + w(a, \xi, t)]n(a, \xi, t) + \int_{-\infty}^{\xi} d\xi' \chi(a, \xi - \xi', \xi', t) n(a, \xi - \xi', t), \tag{5.204}$$

where $w(a, \xi, t)$ denotes the departure rate of community members. If $x(t)$ is a random process describing the scientific productivity variation and $p_a(x, t \mid y, \tau)$ $(\tau < t)$ is the transition probability density corresponding to such a process, then

$$\chi(a, \xi, \xi', t) = \lim_{\Delta t \to 0} \frac{p(\xi + \xi', t + \delta t \mid \xi, t)}{\Delta t}. \tag{5.205}$$

The transition rate $J(a, \xi, t)$ at time t from the productivity level ξ is by definition

$$J(a, \xi, t) = \int_{-\xi}^{\infty} d\xi' \, \chi(a, \xi, \xi', t).$$

The increment ξ' may be positive or negative. The equation for $n(a, \xi, t)$ can be obtained in the following way. First, for the increment we have

$$n(a + \Delta a, \xi, t + \Delta t) = n(a, \xi, t) - J(a, \xi, t) n(a, \xi, t) \Delta t + \int_{-\infty}^{\xi} \chi(a, \xi - \xi', \xi', t) n(a, \xi - \xi', t) d\xi' \Delta t - w(a, \xi, t) n(a, \xi, t) \Delta t, \tag{5.206}$$

where:

- the term on the right-hand side, $[1 - J(a, \xi, t)\Delta t]n(a, \xi, t)$, describes the proportion of individuals whose scientific productivity does not change in $(t, t + \Delta t)$;
- the integral term describes the individuals whose scientific productivity becomes equal to ξ because of increase or decrease in $(t, t + \Delta t)$;
- the last term corresponds to the departure of individuals through stopping research activities or death.

After expanding $n(a + \Delta t, \xi, t + \Delta t)$ around a and t and retaining terms up to the first order in Δt, one obtains the master equation (5.204).

The above master equation is difficult for analysis, and because of this, it is often reduced to an approximation similar to the well-known Fokker–Planck equation. Let

$$\mu_k(a, \xi, t) = \int_{-\xi}^{\infty} d\xi' (\xi')^k \chi(a, \xi, \xi', t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} < (\xi')^k >; \quad k = 1, 2, \ldots, \tag{5.207}$$

where the brackets denote the average with respect to the conditional probability density $p_a(\xi + \xi', t + \Delta t \mid \xi, t)$. In addition, we make the following assumptions:

- $\mu_1, \mu_2 < \infty$;
- $\mu_k = 0$ for $k > 3$;
- $n(a, \xi, t)$ and $\chi(a, \xi, \xi', t)$ are analytic in ξ for all a, t, and ξ'.

The assumption $\mu_k = 0$ for $k > 3$ demands that productivity be continuous, i.e., when $\Delta t \to 0$, the probability of large fluctuations $\mid \xi' \mid$ must decrease so quickly that $<\mid \xi' \mid^3> \to 0$ more quickly than Δt.

When the above assumptions hold, the function n satisfies the equation

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n = -\frac{\partial(\mu_1 n)}{\partial \xi} + \frac{1}{2}\frac{\partial^2(\mu_2 n)}{\partial \xi^2} - wn. \tag{5.208}$$

The following notes are in order here.

1. If $w = 0$, (5.208) is reduced to the Fokker–Planck equation.
2. Equation (5.208) describes the evolution of the scientific community through a drift along the age component and a drift and diffusion with respect to the productivity component.
3. The diffusion term characterized by the diffusivity μ_2 takes into account the stochastic fluctuations of scientific productivity conditioned by internal factors (such as individual abilities, labor motivations, etc.) and external factors (such as labor organization, stimulation systems, etc.).
4. The initial and boundary conditions for (5.208) are:
   - $n(a, \xi, 0) = n^0(a, \xi)$, where $n^0(a, \xi)$ is a known function defining the community age and productivity distribution at time $t = 0$;
   - $n(0, \xi, t) = \nu(\xi, t)$, where the function $\nu(\xi, t)$ represents the intensity of input flow of new members at age $a = 0$ and $\nu(\xi, 0) = n^0(0, \xi)$.
5. In addition, $n(a, \xi, t) \to 0$ as $a \to \infty$.

The general solution of (5.208) with the above initial and boundary conditions is still a difficult task. But for many practical applications, knowledge of the first and second moments of the distribution function $n(a, \xi, t)$ is sufficient. Equation (5.208) can be solved numerically or can be reduced to a system of ordinary differential equations [171].

In a similar way, a model of personal movement can be obtained [172]. The model equation for this case is

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n(a, t) = -n(a, t)[w_1(a, t) + w_2(a, t)] + r(a, t)\nu(t), \tag{5.209}$$

where a is the age variable, t is the time, $n(a, t)$ is the density of researchers having age a at time t, w is the age intensity of researchers' departure, $\nu(t)$ is the intensity of the input flow of new researchers at the moment of time t, $r(a, t)$ is the density

of the input flow age distribution, $w_1(a,t)$ is the intensity of departure due to death, retirement, etc., $w_2(a,t)$ is the intensity of the regulated departure of researchers ($w(a,t) = w_1(a,t) + w_2(a,t)$). Also, a_0 denotes the minimum age of researchers and A denotes the maximum admissible age of researchers; a_0 and A participate in the initial condition

$$n(a,0) = n^0(a), \quad a_0 \leq a \leq A, \tag{5.210}$$

and the boundary condition is

$$n(a_0,t) = 0, \quad t \geq 0. \tag{5.211}$$

5.10 Probability Model for Importance of the Human Factor in Science

Below, we shall discuss a probability model connected to the importance of the human factor in science. One often hears that technological evolution is closely connected to the growth of science and that the growth of science depends heavily on the human factor (number and quality of scientists). Such statements are no surprise, since a connection has been observed between the values of scientometric indicators of the research production of a country's researchers and the corresponding GDP [173–177]. A research organization may have a perfect structure with respect to research positions and research equipment associated with those positions. The research positions may be connected by a perfect system of relations, and the processes in the organization may be carefully planned. But this is not enough. In order to put all the above into effective action, one needs researchers. Researchers of good quality have to fill the research positions. Researchers have to perform actions that contribute to a smooth flow of the processes in a research organization. Only then can the work of this organization be effective. In addition, a researcher does not work alone [178–183]. Teamwork and collaboration among scientists and scientific groups is becoming ever more for solving the scientific problems of today [184–189].

This shows that the human factor is of extreme importance for research organizations. Because of this, we shall discuss below (with the help of mathematics) the importance of the size of the research community.

5.10.1 The Effective Solutions of Research Problems Depend on the Size of the Corresponding Research Community

It is intuitive that larger research communities can solve more complex problems [92]. Let us consider some research problem and let β be the mean probability that a qualified researcher will solve the problem. Then:

- $1-\beta$ is the probability that the researcher will not solve the problem.
- $(1-\beta)^n$ is the probability that a group of n qualified researchers will not solve the problem.

Thus the probability that the same group of n qualified researchers will solve the complex problem (which is not likely to be solved by a single researcher, i.e., $\beta \ll 1$) is

$$p_n = 1-(1-\beta)^n = 1-\exp[n\ln(1-\beta)] \approx 1-\exp(-n\beta). \tag{5.212}$$

If the research group is small, i.e., $n\beta \ll 1$, then from (5.212), we obtain the linear relationship

$$p_n \approx \beta n. \tag{5.213}$$

Then an increase in the size of the group of qualified researchers increases the probability of solving the problem. When the group is small, the probability of solving the problem is proportional of the group size. When the size of the group increases, the nonlinear terms become significant, and the probability p_n increases faster than a linear function.

5.10.2 *Increasing Complexity of Problems Requires Increase of the Size of Group of Researchers that Has to Solve Them*

Scientific organizations evolve and usually become more complex [190, 191]. One factor for such a development is the need to solve research problems of increasing complexity. This increasing complexity leads to a decreasing probability β that a single researcher can solve such a problem. In order to compensate this decrease, one may increase the size of the research group that has to solve the problem.

Let us study the above situation with the help of mathematics. To compensate the decrease of probability means that one has to keep $(dp_n/dt) \geq 0$. Then from (5.212), one obtains

$$\frac{1}{n}\frac{dn}{dt} \geq -\frac{1}{\beta}\frac{d\beta}{dt}. \tag{5.214}$$

Taking into account that $(d\beta/dt) < 0$, the increase in the size of the research group with increasing complexity of the solved problem has to be

$$\frac{dn}{dt} \geq \frac{n}{\beta}\left(-\frac{d\beta}{dt}\right). \tag{5.215}$$

The above simple model leads to the following conclusions. *As the complexity of scientific problems increases with time, one needs larger research collectives in*

order to support a large probability of solving the problems. Thus if a government wants an effective solution of national scientific and technological problems, it has to support a large enough national research community. A decrease in the number of researchers diminishes the national scientific capacity: the probability of solving problems important to the society decreases at least proportionally to the decrease in the size of the corresponding research community.

Note that the value of the parameter β plays an important role in the above model. This value must be kept as large as possible. In other words, an effective scientific community consists of qualified scientists. In addition, let us note that research groups in most cases consist not only of researchers. There are also supporting staff. In connection with this, certain scaling properties may exist for research units [192]. For example, a power law relationship may exist between the number of supporting staff N_s and the number of academic staff N_A of a research institution: $N_s = CN_A^{\beta}$, where C is a constant and β is the exponent of the power law. For the case of the UK National Health System, $C \approx 0.07$ and $\beta \approx 1.3$. The last relationship is an example of a quantitative power law relationship connected to the parts of research (and other) organizations. Such power laws have been discussed in Chap. 4.

5.11 Concluding Remarks

In this chapter, selected classes of deterministic and probability models connected to science dynamics and research production have been discussed. The focus was on the models connected to dynamics of research systems and especially on models for deterministic and statistical properties of the process of publication and the process of citation of research publications. Some of the models have been described very briefly, while for some (probability) models, more discussion has been provided (for the case in which one can obtain interesting conclusions without having to perform long mathematical calculations). This manner of presentation permitted a description of more that twenty models in relatively few pages. We hope that the selected set of models has provided a good impression to the reader about the mathematical tools and methods used for modeling of complex processes and the nonlinear dynamics connected to research systems.

There exist also other deterministic and probability models. For example, there exists a model of science as a part of a global model of a social system. In this model, the scientific system can be treated as a system that has entrances and exits [92]. The input (different flows) comes from the other parts of the social system to the entrances of the science subsystem. At the exit, there are scientific output flows to other parts of the social system. The input flows can be flows of funding or human resources, for example. The main output flow is scientific knowledge. Part of this flow is the flow of publications.

Finally, let us make several remarks on the limited dependent variable models and on the generalized Zipf distribution, since these topics are of significant interest for research in the area of informetrics.

Limited dependent variable models (e.g., binary, ordinal, and count data regression models) may be used for analysis of all kinds of categorical and count data in bibliometrics and scientometrics (such as assessment scores, citation counts, career transitions, editorial decisions, or funding decisions) [193]. The main advantage of limited dependent variable models is that in using them, one may identify the main explanatory variables in a multivariate framework, and in addition, one may estimate the size of the (marginal) effects of these variables.

Let us consider the group of regression models. Limited dependent variable models are a subgroup of this group with a limited range of possible values of the variable of interest. This variable may have a binary outcome (e.g., whether a journal article was cited over a certain period). The variable may take multiple discrete values (e.g., for the case of assessment of research or for the case of peer reviews).

In the case of a binary regression model, we have a variable y_i that can take only the values 0 and 1. We may model the probability that this variable will take value 1 depending on the values of other variables $x_{1i}, \ldots, x_{ki}$ as follows:

$$p(y_i = 1 \mid, x_{1i}, \ldots, x_{ki}) = L(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}), \tag{5.216}$$

where $L(x) = \frac{\exp(x)}{1+\exp(x)}$ is the logistic function (whose range is between 0 and 1). The model (5.216) is called the logit model. The coefficients β_i of the logit model may be estimated by maximizing the likelihood of the data with respect to the coefficients.

The binary logistic model may be used for analyzing or predicting (or for analyzing and predicting) whether articles will be cited [194], for analysis of funding and editorial decisions [195], for analysis of winning scientific awards [196], etc. [197, 198]. One illustration of the application of the model can be seen in [193], in which the dependent variable measures whether an article was cited in another published article during the calendar year following its publication.

For the case of the ordinal regression model, the variable of interest y_i is an ordinal variable that can take only the values $j = 1, 2, \ldots, J$. In this model, the cumulative probability is the probability that an observation i is in the jth category or lower: $p(y_i) \leq j = \delta_{ij}$ can be modeled by the logit relationship

$$\text{logit}(\delta_{ij}) = \alpha_j - \beta_1 x_{1i} - \cdots - \beta_k x_{ki}, \tag{5.217}$$

where $\text{logit}(p) = \log(\frac{p}{1-p}) = \log(p) - \log(1-p)$. Ordinal regression models are applied when we are interested in additional characteristics of the investigated variables with respect to a characteristic modeled by the binary regression model. For example, in the case of binary regression analysis of citations, it was of interest to know whether an article has been cited. If an article has been cited, it may not be of interest how many citations of this article exist. If we are interested in the number of citations, we may use the ordinal regression model above. Such models are used in peer assessment of research groups [199] and for predicting the impact of international coauthorship on citation impact [200].

Finally, one may use count data models if the modeled variable represents the frequency of an event. The count data models can be Poisson models, negative

binomial models, etc. The Poisson model is for a count variable y_i that can take only nonnegative integer values: $0, 1, \ldots$. It is assumed that y_i conditional on the independent variables has a Poisson distribution ($y = 1, 2, \ldots$)

$$p(y_i = y \mid x_{1i}, \ldots, x_{ki}) = \frac{\mu_i^y \exp(-\mu_i)}{y!}, \tag{5.218}$$

where μ_i is the expected value of the distribution that is modeled by

$$\mu_i = \exp[\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}]. \tag{5.219}$$

A limitation of the Poisson regression model is that the Poisson distribution is completely determined by its mean and that the variance is assumed to equal the mean. This restriction may be violated in many applications, since the variance is often greater than the mean. Then there is overdispersion: the variance is greater than the variance implied by assuming a Poisson distribution. One possibility for dealing with overdispersion is to use a negative binomial regression model. This model allows the conditional mean μ_i of y_i to differ from its variance $\mu_i + a\mu_i^2$ by estimating an additional dispersion parameter a.

A Poisson model may be used to identify the effects of coauthorship networks on performance of scholars [201]. Negative binomial regression models can be applied to study citation counts for the purpose of determining the relative importance of authors and journals [202], for comparing sets of papers [203], and for modeling the number of papers [204].

There is a generalization of the Zipf distribution (called the generalized Zipf distribution) that contain as particular cases a family of skew distributions found to describe a wide range of phenomena both within and outside the information sciences and referred to as being of Zipf type. The generalized Zipf distribution is defined as follows [205]. Let

$$d(k \mid f) = \frac{\log[f(k+1)] - \log[f(k)]}{\log(k+1) - \log(k)}, \tag{5.220}$$

where $f(k) > 0$ and the integer k is greater than 1. Let N be the set of natural numbers $1, 2, \ldots$ and Z a random variable defined on N. Let $P(k) = P(X = k)$ and $F(k) = P(X \geq k) = \sum_{i \geq k} P(i)$ be the corresponding distributions connected to Z. A distribution F defined on N is a generalized Zipf distribution with exponent $\alpha > 0$ if and only if $d(k \mid f) \to -\alpha$ as $k \to \infty$, i.e.,

$$\lim_{k \to \infty} d(k \mid f) = \frac{\log[F(k+1)] - \log[F(k)]}{\log(k+1) - \log(k)}. \tag{5.221}$$

It is easily to check that the Waring distribution with

$$F(k) = \frac{\beta^{(k-1)}}{(\alpha + \beta)^{(k-1)}}, \ \beta^{(k)} = \beta(\beta - 1) \ldots (\beta + k - 1) \tag{5.222}$$

is a particular case (belongs to the class) of the generalized Zipf distribution. But the geometric distribution ($P(k) = \theta(1-\theta)^{k-1}$ and $F(k) = (1-\theta)^{k-1}$) does not belong to the class of generalized Zipf distributions.

The class of generalized Zipf distributions has several properties. In order to define the first property, we need to know when a function $\varphi(k)$ varies gradually. Let $\varphi(k)$ be a positive function defined on N. Then $\varphi(k)$ varies gradually if and only if

$$\lim_{k\to\infty} d(k\varphi) = \lim_{k\to\infty} \frac{\log\varphi(k+1) - \log\varphi(k)}{\log(k+1) - \log(k)} = 0; \tag{5.223}$$

$F(k)$ is a generalized Zipf distribution of exponent $\alpha > 0$ if and only if [205]

$$F(k) = \frac{\varphi(k)}{k^{\alpha}}, \tag{5.224}$$

where $\varphi(k)$ is a gradually varying function. An example of a distribution that belongs to the class of generalized Zipf distributions is the Yule distribution, with

$$F(k) = \frac{(k-1)!}{(\alpha+1)^{(k-1)}}. \tag{5.225}$$

We can write this distribution in the form (5.224), where

$$\varphi(k) = \frac{(k-1)!}{(\alpha+1)^{(k-1)}} k^{\alpha}. \tag{5.226}$$

One can define the quantities proportional hazard rate

$$r(k) = \frac{kP(k)}{F(k)}, \tag{5.227}$$

and the conditional expectation

$$e(m) = E[X \mid X \geq m] = \sum_{k\geq m} k \frac{P(x=k)}{P(X \geq m)}. \tag{5.228}$$

Then the following two statements can be proved [205]. First of all, $F(k)$ is a generalized Zipf distribution with exponent $\alpha > 0$ if and only if

$$\lim_{k\to\infty} r(k) \to \alpha. \tag{5.229}$$

Next, $F(k)$ is a generalized Zipf distribution with exponent $\alpha > 1$ if and only if

$$\lim_{k\to\infty} \frac{e(k)}{k} = \lim_{k\to\infty} [e(k+1) - e(k)] = \frac{\alpha}{\alpha - 1}. \tag{5.230}$$

References

1. D. de Solla Price. *Little Science, Big Science*. (Columbia University Press, New York, 1963)
2. D.P. Wallace, The relationship between journal productivity and obsolescence. J. Am. Soc. Inf. Sci. **37**, 136–145 (1986)
3. L. Egghe, On the influence of growth on obsolescence. Scientometrics **27**, 195–214 (1993)
4. W. Glänzel, U. Schoepflin, A bibliometric study on ageing and reception process of scientific literature. J. Inf. Sci. **21**, 37–53 (1995)
5. W. Goffman, V.A. Newill, Generalization of epidemic theory. An application to the transmission of ideas. Nature **204**(4955), 225–228 (1964)
6. P. Nyhius, Logistic curves, in *CIPR encyclopedia of production engineering*, ed. by L. Laperriere, G. Reinhart (Springer, Berlin, 2014), pp. 759–762
7. A. Fernandez-Cano, M. Torralbo, M. Vallejo, Reconsidering Price's model of scientific growth: an overview. Scientometrics **61**, 301–321 (2004)
8. V. Volterra, Population growth, equilibria, and extinction under specified breeding conditions: a development and extension of the theory of the logistic curve, in *The Golden Age of Theoretical Ecology: 1923–1940*, ed. by F.M. Scudo, J.E. Ziegler (Springer, Berlin, 1978), pp. 18–27
9. C.-Y. Wong, L. Wang, Trajectories of science and technology and their co-evolution in BRICS: Insigths from publication and patent analysis. J. Inf. **9**, 90–101 (2015)
10. L. Egghe, I.K. Ravichandra, Rao. Classification of growth models based on growth rates and its applications. Scientometrics **25**, 5–46 (1992)
11. P.S. Meyer, Bi-logistic growth. Technol. Forecast. Soc. Chang. **47**, 89–102 (1994)
12. M. Ausloos, On religion and language evolutions seen through mathematical and agent based models, in *Proceedings of the First Interdisciplinary CHESS Interactions Conference*, ed. by C. Rangacharyulu, E. Haven (World Scientific, Singapore, 2010), pp. 157–182
13. P.S. Meyer, J.W. Yung, J.H. Ausubel, A primer on logistic growth and substitution: the mathematics of the Loglet Lab software. Technol. Forecast. Soc. Chang. **61**, 247–271 (1999)
14. H.W. Menard, *Science: Growth and Change* (Harvard University Press, Cambridge, MA, 1971)
15. G.N. Gilbert, Measuring the growth of science: a review of indicators of scientific growth. Scientometrics **1**, 9–34 (1978)
16. D. Wolfram, C.M. Chu, X. Lu, Growth of knowledge: bibliometric analysis using online database data, in *Informetrics 89/90*, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1990), pp. 355–372
17. G.O. Ware, A general statistical model for estimating future demand levels of data-base utilization within an information retrieval organization. J. Am. Soc. Inf. Sci. **24**, 261–264 (1973)
18. N. Bailey, Some stochastic models for small epidemics in large populations. Appl. Stat. **13**, 9–19 (1964)
19. M.S. Bartlet, *Stochastic Population Models in Ecology and Epidemiology* (Wiley, New York, 1960)
20. W. Goffman, An epidemic process in an open population. Nature **205**, 831–832 (1965)
21. D. Mollison, Dependence of epidemic and population velocities on basic parameters. Math. Biosci. **107**, 255–287 (1991)
22. F.C. Hoppensteadt, *Mathematical Theories of Populations: Demographics, Genetics and Epidemics* (SIAM, Philadelphia, PA, 1975)
23. K. Cooke, P. van den Driessche, X. Zou, Interaction of maturation delay and nonlinear birth in population and epidemic models. J. Math. Biol. **39**, 332–352 (1999)
24. H.W. Hethcote, The mathematics of infectious diseases. SIAM Rev. **42**, 599–653 (2000)
25. V. Colizza, A. Barnat, M. Barthelemy, A. Vespigniani, The modeling of global epidemics: stochastic dynamics and predictability. Bull. Math. Biol. **68**, 1893–1921 (2006)
26. R.M. May, Simple mathematical models with very complicated dynamics. Nature **261**(5560), 459–467 (1976)

27. H. Caswell, *Matrix Population Models* (Wiley, New York, 2001)
28. R.D. Holt, Population dynamics in two-patch environments: some anomalous consequences of an optimal habitat distribution. Theoretical Population Biology **28**, 181–208 (1985)
29. M.P. Hassell, H.N. Comins, R.M. May, Spatial structure and chaos in insect population dynamics. Nature **353**(6341), 255–258 (1991)
30. Z. Ma, J. Li, Basic knowledge and developing tendencies in epidemic dynamics, in *Mathematics for Life Sciences and Medicine*, ed. by Y. Takeuchi, Y. Iwasa, K. Sato (Springer, Berlin, 2007), pp. 5–49
31. N.K. Vitanov, M. Ausloos, Knowledge epidemic and population dynamics models for describing idea diffusion, in *Models for Science Dynamics*, ed. by A. Scharnhorst, K. Börner, P. van den Besselar (Springer, Berlin, 2012), pp. 69–125
32. C. Antonelli, *The Economics of Localized Technological Change and Industrial Dynamics* (Kluwer, Dordrecht, 1995)
33. P. Anderson, Perspective: complexity theory and organization science. Organ. Sci. **10**, 216–232 (1999)
34. M.A. Nowak, Five rules for the evolution of cooperation. Science **314**(5805), 1560–1563 (2006)
35. W. Weidlich, G. Haag, *Concepts and Models of a Quantitative Sociology: The Dynamics of Interacting Populations* (Springer, Berlin, 1983)
36. D. Strang, Adding social structure to diffusion models. Sociol. Methods Res. **19**, 324–353 (1991)
37. P.A. Geroski, Models of technology diffusion. Res. Policy **29**, 603–625 (2000)
38. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. Rev. Mod. Phys. **81**, 591–646 (2009)
39. N.K. Vitanov, Z.I. Dimitrova, Application of the method of simplest equation for obtaining exact traveling-wave solutions for two classes of model PDEs from ecology and population dynamics. Commun. Nonlinear Sci. Numer. Simul. **15**, 2836–2845 (2010)
40. R. Baptista, The diffusion of process innovations: a selective review. Int. J. Econ. Bus. **6**, 107–129 (1999)
41. I.Z. Kiss, M. Broom, P.G. Craze, I. Rafols, Can epidemic models describe the diffusion of topics across disciplines? J. Inf. **4**, 74–82 (2010)
42. H.G. Landau, A. Rapoport, Contribution to the mathematical theory of contagion and spread of information. I: spread through a thoroughly mixed population. Bull. Math. Biophys. **15**, 173–183 (1953)
43. W. Goffman, Mathematical approach to the spread of scientific ideas—the history of mast cell research. Nature **212**, 449–452 (1966)
44. A. Lotka, *Elements of Physical Biology* (Williams and Wilkins, Baltomore, 1925)
45. V. Volterra, Variations and fluctuations of the number of individuals in animal species living together. Journal du Conseil/Conseil Permanent International pour l'Exploration de la Mer **3**, 3–52 (1928)
46. F.J. Ayala, M.E. Gilpin, J.G. Ehrenfeld, Competition between species: theoretical models and experimental tests. Theor. Popul. Biol. **4**, 331–356 (1973)
47. M.E. Gilpin, F.J. Ayala, Global models of growth and competition. PNAS **70**, 3590–3593 (1973)
48. R.D. Holt, J. Pickering, Infectious disease and species coexistence: a model of Lotka-Volterra form. Am. Nat. **126**, 196–211 (1985)
49. Y. Takeuchi, *Global Dynamical Properties of Lotka-Volterra Systems* (World Scientific, Singapore, 1996)
50. A. Castiaux, Radical innovation in established organizations: being a knowledge predator. J. Eng. Technol. Manag. **24**, 36–52 (2007)
51. K. Dietz, Epidemics and rumors: a survey. J. R. Stat. Soc. A **130**, 505–528 (1967)
52. S. Solomon, P. Richmond, Power laws of wealth, market order volumes and market returns. Phys. A **299**, 188–197 (2001)

53. S. Solomon, P. Richmond, Stable power laws in variable economics. Lotka—Volterra implies Pareto—Zipf. Eur. Phys. J. B **27**, 257–261 (2002)
54. W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A **115**, 700–721 (1927)
55. M. Nowakowska, Epidemical spread of scientific objects: an attempt of empirical approach to some problems of meta—science. Theory Decis. **3**, 262–297 (1973)
56. D.J. Daley, Concerning the spread of news in a population of individuals who never forget. Bull. Math. Biophys. **29**, 373–376 (1967)
57. A.D. Barbour, S. Utev, Approximating the Reed-Frost epidemic process. Stoch. Process. Appl. **113**, 173–197 (2004)
58. H. Abbey, An examination of the Reed-Frost theory of epidemics. Hum. Biol. **24**, 201–233 (1952)
59. J.A. Jacquez, A note on chain-binomial models of epidemic spread: what is wrong with the Reed-Frost formulation? Math. Biosci. **87**, 73–82 (1987)
60. W. Goffman, V.A. Newill, Communication and epidemic process. Proc. R. Soc. Lond. Ser. A **298**, 316–334 (1967)
61. G. Harmon, Remembering William Goffman: mathematical information science pioneer. Inf. Process. Manag. **44**, 1634–1647 (2008)
62. M. Cohen, A. Blaivas, A model for the growth of mathematical specialties. Scientometrics **3**, 265–273 (1981)
63. B.M. Gupta, L. Sharma, C.R. Karisiddappa, Modelling the growth of papers in a scientific speciality. Scientometrics **33**, 187–201 (1995)
64. M. Kochen, Stability and growth of knowledge. Am. Doc. **20**, 186–197 (1969)
65. A.N. Tabah, Literature dynamics: studies on growth, diffusion and epidemics. Annu. Rev. Inf. Sci. Technol. **34**, 249–286 (1999)
66. B.M. Gupta, P. Sharma, C.R. Karisiddappa, Growth of research literature in scientific specialities. A modeling perspective. Scientometrics **40**, 507–528 (1997)
67. B.M. Gupta, S. Kumar, S.L. Sangam, C.R. Karisiddappa, Modeling the growth of world social science literature. Scientometrics **53**, 161–164 (2002)
68. L.M.A. Bettencourt, A. Cintron-Arias, D.I. Kaiser, C. Castillo-Chavez, The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. Phys. A **364**, 513–536 (2002)
69. L.M.A. Bettencourt, D.I. Kaiser, J. Kaur, C. Castillo-Chavez, D.E. Wojick, Population modeling of the emergence and development of scientific fields. Scientometrics **75**, 495–518 (2008)
70. M. Szydlowski, A. Krawiez, Growth cycles of knowledge. Scientometrics **78**, 99–111 (2009)
71. D.J. de Solla Price, The exponential curve of science. Discovery **17**, 240–243 (1956)
72. K. Sangwal, Progressive nucleation mechanism and its application to the growth of journals, articles and authors in scientific fields. J. Inf. **5**, 529–536 (2011)
73. K. Sangwal, On the growth of citations of publication output of individual authors. J. Inf. **5**, 554–564 (2011)
74. K. Sangwal, Progressive nucleation mechanism of the growth behavior of items and its application to cumulative papers and citations of individual authors. Scientometrics **92**, 643–655 (2012)
75. K. Sangwal, Growth dynamics of citations of cumulative papers of individual authors according to progressive nucleation mechanism: concept of citation acceleration. Inf. Process. Manag. **49**, 757–772 (2013)
76. D. Kashchiev, *Nucleation: Basic theory with applications* (Butterworth-Heinemann, Oxford, 2000)
77. E.H. Kerner, Further considerations on the statistical mechanics of biological associations. Bull. Math. Biophys. **21**, 217–253 (1959)
78. J.C. Allen, Mathematical model of species interactions in time and space. Am. Nat. **109**, 319–342 (1975)
79. A. Okubo, *Diffusion and Ecological Problems: Mathematical Models* (Springer, Berlin, 1980)

80. W.G. Willson, A.M. de Roos, Spatial instabilities within the diffusive Lotka—Volterra system: individual—based simulation results. Theor. Popul. Biol. **43**, 91–127 (1993)
81. Y.F. le Coadic, Information system and the spread of scientific ideas. R&D Manag. **4**, 97–111 (1974)
82. E. Bruckner, W. Ebeling, A. Scharnhorst, The application of evolution models in scientometrics. Scientometrics **18**, 21–41 (1990)
83. N.K. Vitanov, I.P. Jordanov, Z.I. Dimitrova, On nonlinear population waves. Appl. Math. Comput. **215**, 2950–2964 (2009)
84. N.K. Vitanov, I.P. Jordanov, Z.I. Dimitrova, On nonlinear dynamics of interacting populations: coupled kink waves in a system of two populations. Commun. Nonlinear Sci. Numer. Simul. **14**, 2379–2388 (2009)
85. N.K. Vitanov, Z.I. Dimitrova, M. Ausloos, Verhulst-Lotka-Volterra (VLV) model of ideological struggle. Phys. A **389**, 4970–4980 (2010)
86. N.K. Vitanov, M. Ausloos, G. Rotundo, Discrete model of ideological struggle accounting for migration. Adv. Complex Syst. **15**, Article No. 1250049 (2012)
87. W. Ebeling, A. Scharnhorst, Evolutionary models of innovation dynamics, in *Traffic and Granular Flow'99. Social, Traffic and Granular Dynamics*, ed. by D. Helbing, H.J. Herrman, M. Schekenberg, D.E. Wolf (Springer, Berlin, 2000), pp. 43–56
88. E. Borensztein, J. De Gregorio, J.-W. Lee, How does foreigh direct investment affect economic growth? J. Int. Econ. **45**, 115–135 (1998)
89. J. Dedrick, V. Gurbaxani, K.L. Kraemer, Information technology and economic performance: a critical review of the empirical evidence. ACM Comput. Surv. **35**, 1–28 (2003)
90. S.W. Popper, C. Wagner, New foundations of growth: The U.S. innovation system today and tomorrow. RAND MR-1338.0/1-OSTP (2001)
91. E. Mansfield, *Industrial Research and Technological Innovation: An Econometric Analysis* (Norton, New York, 1968)
92. A.I. Yablonskii, *Mathematical Methods in the Study of Science* (Nauka, Moscow, 1986). (in Russian)
93. C.W. Cobb, P.H. Douglas, A theory of production. Am. Econ. Rev. **18**(Supplement), 139–165 (1928)
94. A. Aulin, *The Impact of Science on Economic Growth and its Cycles* (Springer, Berlin, 1998)
95. Q.L. Burell, Predictive aspects of some bibliometric processes, in *Informetrics 87/88*, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1988), pp. 43–63
96. Q.L. Burrell, A note on ageing in a library circulation model. J. Doc. **41**, 100–115 (1985)
97. D.R. Cox, Some statistical methods connected with series of events (with discussion). J. R. Stat. Soc. B **17**, 129–164 (1955)
98. J. Grandell, *Doubly stochastic Poisson processes*, vol. 529, Lecture Notes in Mathematics (Springer, Berlin, 1976)
99. H.S. Sichel, On a distribution representing sentence-length in written prose. J. R. Stat. Soc. A **137**, 25–34 (1974)
100. H.S. Sichel, Repeat-buying and the generalized inverse Gaussian-Poisson distribution. Appl. Stat. **31**, 193–204 (1982)
101. J.O. Irvin, The generalized Waring distribution. Part I. J. R. Stat. Soc. A **138**, 18–21 (1975)
102. J.O. Irvin, The generalized Waring distribution. Part II. J. R. Stat. Soc. A **138**, 204–227 (1975)
103. J.O. Irvin, The generalized Waring distribution. Part III. J. R. Stat. Soc. A **138**, 374–384 (1975)
104. A.I. Yablonsky, *Mathematical Models in Science Studies* (Nauka, Moscow, 1986). (in Russian)
105. G.U. Yule, A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Philos. Trans. R. Soc. B **213**, 21–87 (1925)
106. H.A. Simon, C.P. Bonini, The size distribution of business firms. Am. Econ. Rev. **48**, 607–617 (1958)
107. M. Brown, S. Ross, R. Shorrock, Evacualtion of a Yule process with immigration. J. Appl. Probab. **12**, 807–811 (1975)

108. N. O'Connell, Yule process approximation for the skeleton of a branching process. J. Appl. Probab. **30**, 725–729 (1993)
109. D.J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. **16**, 23–34 (2001)
110. S. Redner, How popular is your paper? An empirical study of the citation distribution. Eur. Phys. J. B **4**, 131–134 (1998)
111. S. Redner, Citation statistics from 110 years of physical review. Phys. Today **58**, 49–54 (2005)
112. C.C. Sarli, E.K. Dubinsky, K.L. Holmes, Beyond citation analysis: a model for assessment of research impact. J. Med. Libr. Assoc. **98**, 17–23 (2010)
113. M.Y. Wang, G. Yu, D.R. Yu, Minimg typical features for highly cited papers. Scientometrics **87**, 695–706 (2011)
114. M. Wang, G. Yu, S. An, D. Yu, Discovery of factors influencing citation impact based on a soft fuzzy rough set model. Scientometrics **93**, 635–644 (2012)
115. Q.L. Burrell, Stochastic modeling of the first-citation distribution. Scientometrics **52**, 3–12 (2001)
116. L. Egghe, I.K. Ravichandra Rao, Citation age data and the obsolescence function: fits and explanations. Inf. Process. Manag. **28**, 201–217 (1992)
117. R. Rousseau, Double exponential models for first-citation processes. Scientometrics **30**, 213–227 (1994)
118. L. Egghe, A heuristic study of the first-citation distribution. Scientometrics **48**, 345–359 (2000)
119. D.R. Cox, V.I. Isham, *Point Processes* (Chapman & Hall, London, 1980)
120. J.F.C. Kingman, *Poisson processes* (Clarendon Press, Oxford, 1992)
121. T. Mikosch, *Non-life Insurance Mathematics. An Introduction with the Poisson Process* (Springer, Berlin, 2009)
122. H.C. Tijms, *A First Course in Stochastic Models* (Wiley, Chichester, 2003)
123. S. Nadarajan, S. Kotz, Models for citations behavior. Scientometrics **72**, 291–305 (2007)
124. S.M. Ross, *Stochastic Processes* (Wiley, New York, 1996)
125. Q.L. Burrell, The *n*-th citation distribution and obsolescence. Scientometrics **53**, 309–323 (2002)
126. A.F.J. van Raan, Sleeping beauties in science. Scientometrics **59**, 467–472 (2004)
127. Q.L. Burrell, Are "sleeping beauties" to be expected? Scientometrics **6**, 381–389 (2005)
128. J. Grandell, *Mixed Poisson processes* (Chapman & Hall, London, 1997)
129. S.A. Klugman, H.H. Panjer, G.E. Wilmot, *Loss Models. From Data to Decisions* (Wiley, Hoboken, NJ, 2008)
130. M. Bennet, *Stochastic Processes in Science, Engineering and Finance* (Chapman & Hall, Boca Raton, FL, 2006)
131. R.-D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values* (Birkhäuser, Basel, 1997)
132. Q.L. Burrell, A simple stochastic model for library loans. J. Doc. **36**, 115–132 (1980)
133. Q.L. Burrell, Predictive aspects of some bibliometric processes, in *Infometrics 87/88*, ed. by L. Egghe, R. Rousseau (Amsterdam, Elsevier, 1988), pp. 43–63
134. Q.L. Burrell, Using the gamma-Poisson model to predict library circulation. J. Am. Soc. Inf. Sci. **41**, 164–170 (1990)
135. J.M. Hilbe, *Negative Binomial Regression* (Cambridge University Press, Cambridge, 2007)
136. N.L. Johnson, A.W. Kemp, S. Kotz, *Univariate Discrete Distributions* (Willey, Hoboken, NJ, 2005)
137. J.H. Pollard, *A Handbook of Numerical and Statistical Techniques: With Examples Mainly from the Life Sciences* (Cambridge University Press, Cambridge, 1977)
138. M. Greenwood, G.U. Yule, An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or repeated accidents. J. R. Stat. Soc. A **83**, 255–279 (1920)
139. J. Mingers, Q.L. Burrell, Modeling citation behavior in management science journals. Inf. Process. Manag. **42**, 1451–1464 (2006)

140. E.S. Vieira, J.A.N.F. Gomes, Citation to scientific articles: its distribution and dependence on the article features. J. Inf. **4**, 1–13 (2010)
141. C. Lachance, V. Lariviere, On the citation lyfecycle of papers with delayed recognition. J. Inf. **8**, 863–872 (2014)
142. A.I. Yablonskii, *Models and Methods of Mathematical Study of Science* (AN USSR, Moscow (in Russian), 1977)
143. A. Schubert, W. Glänzel, A dynamic look at a class of skew distributions. A model with scientometric application. Scientometrics **6**, 149–167 (1984)
144. W. Glänzel, A. Schubert, Predictive aspects of a stochastic model for citation processes. Inf. Process. Manag. **31**, 69–80 (1995)
145. R. Frank, Brand choice as a probability process. J. Bus. **35**, 43–56 (1962)
146. J.S. Coleman, *Introduction to Mathematical Sociology* (Collier-Macmillan, London, 1964)
147. H.A. Simon, On a class of skew distribution functions. Biometrica **42**, 425–440 (1955)
148. Y. Ijiri, H. Simon, *Skew Distributions and the Sizes of Business Firms* (North Holland, Amsterdam, 1977)
149. J. Eeckhout, Gibrath's law for (all) cities. Am. Econ. Rev. **94**, 1429–1451 (2004)
150. W. Glänzel, *Bibliometrics as a Research Field: A Course on Theory and Application of Bibliometric Indicators* (Ungarische Akademie der Wissenschaften, Budapest, 2003)
151. W. Glänzel, U. Schoepflin, A stochastic model for the ageing of scientific literature. Scientometrics **30**, 49–64 (1994)
152. S. Shan, G. Yang, L. Jiang, The multivariate Waring distribution and its application. Scientometrics **60**, 523–535 (2004)
153. M. Abramowitz, I.A. Stegun (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1972)
154. Q.L. Burrell, Age-specific citation rates and the Egghe-Rao function. Inf. Process. Manag. **39**, 761–770 (2003)
155. P. Fronczak, A. Fronczak, J.A. Holyst, Publish or perish: Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. Phys. Rev. E **75**, Art. No.026103 (2007)
156. K.G. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949)
157. A.I. Yablonsky, On fundamental regularities of the distribution of scientific productivity. Scientometrics **2**, 3–34 (1980)
158. L. Hartman, Technological forecasting, in *Multinational Corporate Planning*, ed. by G.A. Steiner, W. Cannon (Crowell-Collier Publishing Co., New York, 1966)
159. G.W. Tyler, A thermodynamic model of manpower system. J. Oper. Res. Soc. **40**, 137–139 (1989)
160. I.K. Ravichandra Rao, Probability distributions and inequality measures for analysis of circulation data, in *Informetrics*, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1988), pp. 231–248
161. W. Glänzel, On the *h*-index—A mathematical approach to a new measure of publication activity and citation impact. Scientometrics **67**, 315–321 (2006)
162. E.J. Gumbel, *Statistics of Extremes* (Dover, New York, 2004)
163. W. Glänzel, A. Schubert, Price distribution. An exact formulation of Price's "Square root law". Scientometrics **7**, 211–219 (1985)
164. H. Boxenbaum, F. Pivinski, S.J. Ruberg, Publication rates of pharmaceutical scientists: application of the Waring distribution. Drug Metab. Rev. **18**, 553–571 (1987)
165. Q.L. Burrell, A simple model for linked infometric processes. Inf. Process. Manag. **28**, 637–645 (1992)
166. Q.L. Burrell, Hirsch's *h*-index: a stochastic model. J. Inf. **1**, 16–25 (2007)
167. H.S. Sichel, A bibliometric distribution which really works. J. Am. Soc. Inf. Sci. **36**, 314–321 (1985)
168. H.S. Sichel, Anatomy of the generalized inverse Gaussian-Poisson distribution with special application to bibliometric studies. Inf. Process. Manag. **28**, 5–17 (1992)

169. L. Perreault, B. Bobee, R. Rasmussen, Halphen distribution system. Mathematical and statistical properties. J. Hydrol. Eng. **4**, 189–199 (1999)
170. H.S. Sichel, Repeat-bying and the generalized inverse Gaussian-Poisson distribution. Appl. Stat. **31**, 193–204 (1982)
171. A.K. Romanov, A.I. Terekhov, The mathematical model of productivity—and age-structured scientific community evolution. Scientometrics **39**, 3–17 (1997)
172. A.K. Romanov, A.I. Terekhov, The mathematical model of the scientific personnel movement taking into account the productivity factor. Scientometrics **33**, 221–231 (1995)
173. P. Vinkler, Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non- EU countries. Scientometrics **74**, 237–254 (2008)
174. L.C. Lee, Y.W. Chuang, Y.Y. Lee, Research output and economic productivity: a Granger causality test. Scientometrics **89**, 465–478 (2011)
175. P.W. Hart, J.T. Sommerfeld, Relationship between growth in gross domestic product (GDP) and growth in the chemical engineering literature in five different countries. Scientometrics **42**, 299–311 (1998)
176. F. de Moya-Anegon, V. Herrero Solana, Science in America Latina: a comparison of bibliometric and scientific-technical indicators. Scientometrics **46**, 299–320 (1999)
177. F. Ye, A quantitative relationship between per capita GDP and scientometric criteria. Scientometrics **71**, 407–413 (2007)
178. J. Sylvan Katz, B.R. Martin, What is research collaboration? Res. Policy **26**, 1–18 (1997)
179. A.F.J. van Raan, Science as an international enterprise. Sci. Public Policy **24**, 290–300 (1997)
180. M. Pezzoni, V. Sterzi, F. Lissoni, Career progress in centralized academic systems: Social capital and institutions in France and Italy. Res. Policy **41**, 704–719 (2012)
181. D.B. de Beaver, R. Rosen, Studies in scientific collaboration: Part I-The professional origins of scientific co-authorship. Scientometrics **1**, 65–84 (1979)
182. D.B. de Beaver, R. Rosen, Studies in scientific collaboration: Part II—Scientific co-authorship, research productivity and visibility in the French scientific elite 1799–1830. Scientometrics **1**, 133–149 (1979)
183. D.B. de Beaver, R. Rosen, Studies in scientific collaboration: Part III—Professionalization and the natural history of modern scientific co-authorship. Scientometrics **1**, 231–245 (1979)
184. T. Luukkonen, O. Persson, G. Sivertsen, Understanding patterns of international scientific collaboration. Sci. Technol. Hum. Values **17**, 101–126 (1992)
185. M. Meyar, O. Persson, Nanotechnology—interdisciplinarity, patters of collaboration and differences in application. Scientometrics **42**, 195–205 (1998)
186. A.E. Andersson, O. Persson, Networking scientists. Ann. Reg. Sci. **27**, 11–21 (1993)
187. G. Melin, O. Persson, Hotel cosmopolitan: a bibliometric study of collaboration at some European universities. J. Am. Soc. Inf. Sci. **49**, 43–48 (1998)
188. P. Mählck, O. Persson, Socio-bibliometric mapping of intra-department networks. Scientometrics **49**, 81–91 (2000)
189. T. Lukkonen, R. Tijssen, O. Persson, G. Sivertsen, The measurement of international scientific collaboration. Scientometrics **28**, 15–36 (1993)
190. C.S. Wagner, L. Leydesdorff, Network structure, self-organization, and the growth of international collaboration in science. Res. Policy **34**, 1608–1618 (2005)
191. R. Stichweh, Science in the system of world society. Soc. Sci. Inf. **35**, 327–340 (1996)
192. B. Jamweit, E. Jettestuen, J. Mathiesen, Scaling properties in European research units. PNAS **106**, 13160–13163 (2009)
193. N. Deschacht, T.C.E. Engels, Limited dependent variable models and probabilistic prediction in informetrics, in *Measuring Scholarly Impact. Methods and Practice*, ed. by Y. Ding, R. Rousseau, D. Wolfram (Springer, Cham, 2014), pp. 193–214
194. H.P. Van Dalen, K. Henkens, Signals in science—the importance of signaling in gaining attention in science. Scientometrics **64**, 209–233 (2005)
195. J.W. Fedderke, The objectivity of national research foundation peer review in South Africa assessed against bibliometric indexes. Scientometrics **97**, 177–206 (2013)

196. L. Rokach, M. Kalech, I. Blank, R. Stern, Who is going to win the next Association for the Advancement of Artificial Intelligence fellowship award? Evaluating researchers by mining bibliographic data. J. Am. Soc. Inf. Sci. Technol. **62**, 2456–2470 (2011)
197. P. Jensen, J.-B. Rouquier, Y. Croissant, Testing bibliometric indicators by their prediction of scientists promotions. Scientometrics **78**, 467–47 (2009)
198. P. Vakkari, Internet use increases the odds of using the public library. J. Doc. **68**, 618–638 (2012)
199. T.C.E. Engels, P. Goos, N. Dexters, E.H.J. Spruyt, Group size, *h*-index and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. Res. Eval. **22**, 224–236 (2013)
200. S.-C.J. Sin, International coauthorship and citation impact: a bibliometric study of six LIS journals, 1980–2008. J. Am. Soc. Inf. Sci. Technol. **62**, 1770–1783 (2011)
201. A. Abbasi, J. Altmann, L. Hossain, Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. J. Inf. **5**, 594–607 (2011)
202. G.D. Walters, Predicting subsequent citations to articles published in twelve crimepsychology journals: author impact versus journal impact. Scientometrics **69**, 499–510 (2006)
203. L. Bornmann, H.D. Daniel, Selecting scientific excellence through committee peer review—a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. Scientometrics **68**, 427–440 (2006)
204. F. Barjak, S. Robinson, International collaboration, mobility, and team diversity in the life sciences: impact on research performance. Soc. Geogr. **3**, 23–36 (2008)
205. S. Shan, On the generalized Zipf distribution. Part I. Inf. Process. Manag. **41**, 1369–1386 (2005)