Jointly published by Elsevier Science Ltd, Oxford and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 44, No. 2 (1999) 135–155

LOTKA'S LAW RECONSIDERED: THE EVOLUTION OF PUBLICATION AND CITATION DISTRIBUTIONS IN SCIENTIFIC FIELDS

NICOLE J. SAAM,[1] L. REITER[2]

[1]*Institut für Soziologie, Universität München, Konradstr. 6, D-80801 München (Germany)*
[2]*Universitätsklinik für Tiefenpsychologie und Psychotherapie Wien, Universität Wien, Währinger Gürtel 18-20, A-1090 Wien (Austria)*

(Received September 29, 1998)

This paper reports early steps in research that seeks to clarify how publications of scientists interact dynamically with citations and reputation in shaping the evolution of scientific fields. We assume that Lotka's modified law holds for scientific fields. A primary approach to model publication productivity was published by Yablonsky. In contrast to Yablonsky's unfinished mathematical approach, our simulation approach is not predominantly driven by insight into the formal generation mechanisms of certain processes but more theory driven. It considers the evolution of publication and citation distributions over the histories of scientific fields using both simulated and real historical data.

Introduction

This paper reports early steps in research that seeks to clarify how publications of scientists interact dynamically with citations and reputation in shaping the evolution of scientific fields. It considers the evolution of publication and citation distributions over the histories of scientific fields using both simulated and real historical data.

The first important bibliometric study was published by *Lotka* in 1926. He had discovered that the productivity of scientists follows a power law. He as well as his followers believed that he had found a distribution describing the life time productivity of scientists. Reanalyzing his data *MacRoberts* and *MacRoberts* (1982) showed that Lotka's Law did not apply to the life time productivity of scientists but to the productivity within scientific fields ("areas"). Nevertheless, the "somewhat remarkable" (Lotka) distribution is not yet understood.

0138–9130/99/US $ 15.00

Meanwhile, bibliometrics and scientometrics have advanced tremendously in collecting data, e.g., by foundation of the *Science Citation Index* (SCI) and the *Social Sciences Citation Index* (SSCI) by Garfield, and in means to evaluate these data, especially by help of modern computer technologies. They have led to the discovery of another remarkable distribution, the distribution of citations, which seems to follow a power law as well. Here as there, the distribution is not yet understood, nor, if and to what extent there is a connection between both distributions. For recent attempts to understand why these distributions might appear see *Bookstein* (1990ab, 1997; *Bookstein* and *Wright,* 1997). But today, knowing this seems to be more desirable than ever before.

Whereas Lotka's study reflected the interest of scientists in analyzing their own work – an activity which nowadays would be called self-reflexive – and his results had no effect on scientists' work, today, as governments and universities increasingly try to compare and evaluate the performance of scientists, using productivity and citations as indicators, misunderstandings in the point of reference as well as the underlying causes of productivity and citation distributions may lead to undesired effects, at worst to an unjust distribution of ressources.

Current prospects for complete analytic characterizations of the dynamics of publications and citations in scientific fields are poor because the processes involved are complicated. In particular, they involve linkages between microdynamics and macrodynamics. Although we have begun to understand various aspects at each level, we do not yet know how to integrate them analytically. Therefore, we use a combination of empirical analysis and multilevel simulation.

It is costly and difficult to obtain data suitable for estimating models of publications and citations in a scientific field over long historic periods, as is appropriate for dynamic studies of scientific fields. Yet, available data sometimes provide snapshots of a series of distributions in the evolution of scientific fields even when panel data on the publications and citations of individual scientific authors are not available. We think that we can learn about both the dynamics of scientific fields and individual scientists by studying the dynamics of the distributions. By learning the consequences of the features of individual processes for the evolution of distributions in cases for which rich data are available, we seek to uncover patterns that can be used to infer dynamics when distributions are available but microdata are not.

Classical and recent models of publication distributions and citations

Publications: Lotka's Law

Lotka's Law describes the asymmetrical distribution of the frequency of publications by scientists. He assumed that the empirical distribution may be reproduced best by a Pareto distribution: the relative portion of scientists having n publications is proportional to the quotient $1/n^2$ (Lotka, 1926). This indicates that many scientists publish very few articles, books, etc., whereas very few scientists publish many. Lotka was in search for an explanation of this "law", but did not succed. Reevaluating Lotka's data *MacRoberts* and *MacRoberts* (1982) suggested to interpret this distribution not as life time productivity of scientists as Lotka did, but as applying to the productivity of scientists within scientific fields ("areas"). The distribution of the life time productivity of scientists is not yet known. *MacRoberts* and *MacRoberts* discuss the methodological problems that are in hand here. An analysis by *Reiter, Steiner* and *Werner* (1997) is based on the publications of five different journals in the field of psychoanalysis, family therapy, and systemic therapy (*Psyche, Familiendynamik, Journal of Family Psychiatry, Forum der Psychoanalyse, System Familie*). The results correspond by and large to Lotka's Law, and confirm *MacRoberts* and *MacRoberts*' suggestion that Lotka's Law applies to the publication distribution of scientific fields.

Lotka's Law reconsidered

Although Lotka's Law postulates a Pareto distribution of publications, later studies by *Price* (1963, 1976, 1986a) have demonstrated that this law should be modified. The distribution approaches to $1/n^3$ for highly productive scientists. Therefore *Price* has developed his own – somewhat more complicated – algorithm to reproduce the empirical distribution more correctly. His final discussion even suggests that one may assume a lognormal distribution of publications. Anyhow, most interestingly, if the lognormal distribution is cut up to the mode (the value with the highest frequency) the remaining distribution is very similar to the Pareto distribution. Recently *Sen* and *Chattopadhyay* (1996) have presented a linear equation for bibliometric distributions.

From the modeler's point of view who in the first place investigates in the empirical distribution of publications, and not in Lotka's Law, the exact algorithm that *describes* the distribution is not that much important because it is the modeler's first task to reproduce empirically observed distributions by modeling and simulation, and not Lotka's Law. Decisive is that one has to reproduce and *explain* a right skewed

distribution which shows some empirical variance by modeling and simulating the causal relations that are assumed to underly the phenomenon. Having advanced in this one may come back to Lotka and reconsider the "law" he had found.

Static explanations of Lotka's Law

Classical approaches give static explanations to the productivity of scientists. The productivity of scientists is deduced to individual or structural factors. Individual factors may be the biographic history, the personality, the creativity, the motivation, the identification, the organizational talent, the age or the sex of the publishing scientist. Structural factors may be the scientific socialization, the selection and stratification of scientists, the portion of teamwork, the degrees of freedom in chosing one's topics and problems, institutional constraints like "publish or perish" and other organizational features (e.g., reward systems). A discussion of these arguments may be found in *Fox* (1983) and *Mayer* (1993).

Dynamic explanations of Lotka's Law

Actually, this distribution is the product of a dynamical process. There are three hypotheses which give a dynamic explanation: Due to the *Matthew effect* (derived from Matthew **25**,29, *Merton* 1968) those who already own a portion of something shall be given more. *Zuckerman* (1993) used but not introduced the term *cumulative advantage* to describe the same phenomenon. It was also discussed by *Simon* and *Mandelbrot.* The Simon-Mandelbrot debate* involved several rounds of replies and new criticism. *Price* (1976), *Tague* (1981), and *Glänzel* and *Schubert* (1990) have developed mathematical models to describe this phenomenon. Moreover, *Tague* used also the term *success-breeds-success phenomenon* synonymously. Although this hypothesis may describe what actually happens it does not give an overall causal explanation. In scientometrics cumulative advantage may be explained through conditional expectations (*Glänzel* and *Schubert,* 1990). *Price* (1986b, originally published 1975) has discovered the phenomenon and scope of transience. In his analysis, he found that in each year authors who have never been heard of before and never are heard of again amount to about 25 % of those recorded for that year. They are called transients. These people cannot have migrated to a different field of research, for the corpus he investigated in included

*The debate took place in the journal *Information and Control* and started with *Mandelbrot* 1959. The final conclusion from Simon's point of view seems to be found in *Ijiri* and *Simon,* 1977.

all publishable fields of science, all institutions and countries. For those beginning a publishing career the mortality is very high. He calculated a total birthrate of 45% and a death rate of 35% which overlap to give the transients. In his opinion, this is a result of the Matthew effect.

Reiter, *Steiner* and *Werner* (1997) have analyzed the temporal evolution of publications of five different journals in the field of psychoanalysis, family therapy, and systemic therapy (*Psyche, Familiendynamik, Journal of Family Psychiatry, Forum der Psychoanalyse, System Familie*). In the beginning years, the (successful) journals perceive a comparatively high concentration of authors which is decreasing continuously, finally reaching an equilibrium (unsuccessful journals do not decrease their concentration to the same extent and stop being published). A possible dynamical explanation of this process is that initially, a leading group of authors which has organized the foundation of the journal predominates the publications (*concentration hypothesis*). In the course of time, the journal becomes established which is indicated by other scientists increasingly submitting papers for publication. Finally, the equilibrium point is reached indicating the maturity of the journal (*equilibrium hypothesis*).

Citations

The frequency of citations of scientists (self citation excluded) seems to follow a distribution which is very close to that of the publications. We want to join *Stephan* and *Levin* (1991) who assume that citations may be more right skewed than publications.

The higher the reputation of the scientist, the quality of the article, or the quality of the journal the article was published in the higher the probability of receiving citations more frequently. Furthermore, citation networks are made responsible for high citation frequencies. *Whitley* (1972) adduces the *bandwagon effect*: As soon as one article of an author is quoted, other articles may follow.

A dynamical model of publications and citations

This section describes our strategy of building a dynamical model to reproduce and explain the evolution of publication and citation distributions over time. We assume that Lotka's modified law holds for scientific fields. Some primary approaches to model publication productivity may be found in *Yablonsky* (1980: 16-17). In contrast to *Yablonsky*'s unfinished mathematical approach our simulation approach is not

predominantly driven by insight into the formal generation mechanisms of certain processes but more theory driven.

In a nutshell, our dynamical model of publications and citations incorporates 8 variables (W[rite], U[tility], P[ublications], C[itations], F[ield], R[eputation], S[tate], A[ctive]) which are calculated on the micro level of the scientists and 11 parameters (U^*, α, ω, χ, β, γ, λ, μ, κ, η, σ). On the macro level we have calculated the aggregations and distributions of the micro variables' values, as well as coefficients of skewness.

Publications

Before being able to get published a certain paper or book a scientist has to produce (write) it. This is the only activity in the whole process that is accomplished by the scientist himself. Being published, cited, or attributed any degree of reputation are passive operations executed by publishers, and other scientists of the scientific field. In the previous section, we have given a short review of factors which may rely to the productivity of scientists. We concentrate on the Matthew effect for three reasons. First, as explained above, we do not have available those microdata which would be necessary to evaluate the more detailed hypotheses, concerning the individual's background, his personality, creativity etc. The production of papers and books is a phenomenon which is very difficult to observe. Second, it is a most simple assumption which may already produce a lognormal distribution as it is equal to the "law of proportionate effect" (*Kapteyn,* 1903). Last but not least, it seems improbable to us (see also *Reiter*, *Steiner* and *Werner,* 1997) that the above mentioned variety of individual and structural factors of comparatively low complexity (level) interact without the influence of higher level structures and produces a distribution so stable that it is called a law.

Due to the Matthew effect we assume that the paper production increases continuously. Let W_{it} denote the paper production of scientist i in period t. Assume that the paper production of each scientist in each period be one except he has died during the last time intervall ($S_{i,t-1} = 1$, see section *birth-death processes*).

$$W_{it} = \begin{cases} 0 \text{ if } S_{i,t-1} = 1 \\ 1 \text{ else} \end{cases} \qquad (1)$$

We make assumptions on the utility of the papers that are produced. We do not use the concept of quality, here, because there are enough examples in history that quality was not recognized when a paper should be published for the first time. The utility of a paper (*Reiter,* 1995) shall indicate its utility for other scientists at the time of its first publication. Each person is given a personal average utility ω_i which is a uniform random number between 0.4 and 0.7 reflecting her talent as it is perceived by the scientific community of the field. The utility of each persons' papers varies: It is exponentially distributed around the personal average utility. Let U_{it} denote the utility of the paper produced by scientist i in period t:

$$U_{it} = \exp(\omega) \tag{2}$$

Whether a paper is accepted for publication is a discrete problem. Equation 3 describes this decision process. Let P_{it} denote the overall number of papers published by scientist i in period t.

$$P_{it} = \begin{cases} P_{i,t-1} + W_{it} \text{ if } U_{it} > U^* \vee R_{i,t-1} > R^*_{t-1} \\ P_{i,t-1} \text{ else} \end{cases} \tag{3}$$

In case that a scientist has completed a paper ($W_{it} = 1$) it is being published either if the utility is higher than a threshold utility (if ($U_{it} > U^*$)) or if her personal reputation in this scientific field is higher than a threshold value of reputation (if ($R_{i,t-1} > R^*_{t-1}$), see section *reputation*). We assume that the threshold value of utility U^* is a constant, whereas the threshold value of reputation R^* derives dynamically from the overall distribution of reputation in this scientific field, and from the maximum value of reputation that a single scientist held in this field during the last period ($0 < \alpha < 1$):

$$R^*_t = \alpha\, R_{\max,t-1} \tag{4}$$

Citations

Let C_{it} denote the overall number of citations of scientist i's papers in period t by other scientists. It depends on the overall citations $C_{i,t-1}$ at time t-1, the product of the overall utility of his published articles $P^*_{i,t-1}$ which is subject to a certain degree of depreciation ($P^*_{it} = \chi^* P^*_{i,t-1} + U_{i,t-1}$ with $0 < \chi < 1$), and on the proximity F_i of his papers to the scientific field. The term is multiplied with parameter γ with $0.0 < \gamma < 1.0$

which is randomly distributed indicating the randomness of being quoted and weighted by parameter $\beta \geq 1$. There exist two different values of β: a slightly greater one for authors who publish and are quoted in the field (β_1) and a smaller one for those who are quoted only (β_2). To obtain whole numbers of citations C is truncated.

$$C_{it} = C_{i,t-1} + [\, \beta \, \gamma \, P^*_{i,t-1} \, (\lambda - F_i)] \tag{5}$$

To allow for old papers being quoted, e.i., papers that have been published before the formation of the new field, each individual is initialised with a chisquare distributed random number of publications with degree of freedom 4. This distribution was chosen because a uniform as well as a normal distribution would be counterintuitive. We must start from any kind of right skewed distribution, but we should not start from exactly the same kind of distribution that the simulation should reproduce.

F_i describes the proximity that the scientists who are quoted in a certain scientific field have to this field. E.g., a theorist of social system theory may be quoted by a member of the systemic therapy field without ever having published in this field himself. He may nevertheless be closer to this field than a mathematician who has developed an algorithm which is used and therefore quoted by some other systemic therapy adept to model certain therapeutic processes. F_i is initialised as a random number with $0.0 \leq F_i \leq 1.0$. A distance parameter $\lambda = 1.5$ is used to calculate the proximity.

Reputation

We assume that the reputation R_{it} of a scientist i at period t depends on her reputation during the last period which is subject to a certain degree of depreciation ($0 < \mu < 1$) induced by the natural process of forgetting. She may win new reputation from the citations of all her papers in the last period.

$$R_{it} = R_{i,t-1} + \Delta t \, (-\mu \, R_{i,t-1} + C_{it} - C_{i,t-1}) \tag{6}$$

According to equation 6, R_{it} depends on initial size R_{i0} which is initialized using the same personal constant mean which was used to calculate the utility of the scientist's papers (we plan experiments to initialize R_{i0} as zero).

Birth-death processes

Following *Price* (1986b) we assume birth-death processes. If a person has been dead during the last time step ($S_{i,t-1} = 1$) he is born with probability $\kappa = 0.45$. If he has

been alive ($S_{i,t-1}$ =2) he dies with probability η = 0.35. A new person is born by copying any existing person, taking over his mean utility and field values, and resetting all other variables to zero.

$$\begin{aligned} P(S_{it}=1) &= \eta \quad \text{if} \quad S_{i,t-1}=2 \\ P(S_{it}=2) &= \kappa \quad \text{if} \quad S_{i,t-1}=1 \end{aligned} \tag{7}$$

Finally, we initialize whether an author actually writes in the model scientific field. Then he is active. He is passive if he is only being quoted.

$$\begin{aligned} P(A_i=1) &= \sigma \\ P(A_i=0) &= 1-\sigma \end{aligned} \tag{8}$$

In sum, our dynamical model of publications and citations incorporates 8 variables which are calculated on the micro level of the scientists and 11 parameters. On the macro level we have calculated the aggregations and distributions of the micro variables' values, as well as coefficients of skewness. Table 1 gives an overview of the variables and parameters of the model. Figure 1 is giving a graphic representation of the model (Th indicating thresholds).

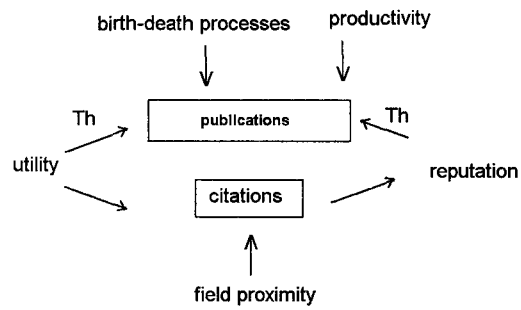

Fig. 1. Graphic representation of the formal model

Table 1
Variables and parameters of the formal model

Variable	Meaning	Initialization at time t=0
W_{it}	*Write*: paper production of scientist i in period t	0 for each scientist
U_{it}	*Utility*: utility of the paper produced by scientist i in period t	0 for each scientist
P_{it}	*Publications*: overall number of papers published by scientist i in period t	chi-square distributed random number with degree of freedom 4
C_{it}	*Citations*: overall number of citations of scientist i's papers in period t	0 for each scientist
F_i	*Field*: proximity that scientist i has to the field of research where he/she is being published or quoted	uniformly distributed random number, $0.0 \leq F_i \leq 1.0$
R_{it}	*Reputation*: reputation of scientist i in period t	chisquare distributed random number with degree of freedom 4.0
S_{it}	*State*: State variable indicating that a scientist is alive or dead	0 for each scientist
A_i	*Active*: personal constant indicating whether a scientist actually writes in a field (then he/she is active) or whether he/she is only being quoted (is passive)	0 for each scientist
Parameters	**Interpretation**	
U^*	threshold utility	$U^* = 0.55$
α	coefficient indicating the relation between maximum reputation of a scientist and the threshold reputation	$0 < \alpha < 1$
ω_i	personal average utility of papers of scientist i	uniformly distributed random number, $0.4 < \omega_i < 0.7$
χ	the degree of depreciation of the overall utility of published papers	$0 < \chi < 1$
β	weighing factor changing the probability of being quoted: there exist two different values of β: a slightly bigger one for authors who publish and are quoted in the field (β_1) and a smaller one for those who are quoted only (β_2).	$\beta \geq 1$, $\beta_1 > \beta_2$
γ	coefficient indicating the randomness of being quoted	uniformly distributed random number, $0.0 < \gamma < 1.0$
λ	distance parameter to calculate the proximity to a field of research	$\lambda = 1.5$
μ	the degree of depreciation of reputation	$0 < \mu < 1$
κ	probability that a new person is born	$\kappa = 0.45$
η	probability that a living person dies	$\eta = 0.35$
σ	proportion of scientists who are active	$\sigma = 0.07$

A measure of skewness

Although there are scientometric indicators which measure certain characteristics of publication productivity (e.g., *Schubert* and *Glänzel,* 1991, *Braun, Glänzel* and *Schubert,* 1990), for instance population growth, transience, renewal, and cumulative advantage, there is no indicator available which can be used to compare the skewness of the simulated versus empirical publication and citation distribution. *Rao* (1988) has discussed measures of inequality which are suitable for circulation data, e.g., the Pearson skew coefficient, the α-measure, Gini's index, the Lorenz coefficient, and Pratt's measure. As his empirical data show circulation distributions are far less skew than publication and citation data. Our empirical distributions are extremely right skewed. The median always belongs to the minimum value of the publications or citations (see Table 4). Therefore, we did not apply one of the discussed measures.

In order to indicate and compare the skewness of the simulated and the empirical data we have experimented with several coefficients of skewness which we adopted from *Sachs* (1992: 167), *Fahrmeir* (1997: 72) and *Hogg* (1974: 918). *Sachs* and *Fahrmeir* introduce several measures of skewness, e.g., a 1-9 decil coefficient of skewness, which is based on the median and difference of decil 1 to decil 9. After some tests with these coefficients we transformed the latter using the median, the minimum and the maximum of x:

$$s = \frac{(x_{max} - m) - (m - x_{min})}{(x_{max} - m) + (m - x_{min})} = \frac{x_{max} + x_{min} - 2m}{x_{max} - x_{min}} \tag{9}$$

m representing the median, and $-1 \leq s \leq 1$. In case of a symmetrical distribution s equals 0. This coefficient may not be robust to extreme activities. As a robust procedure *Hogg* (1974: 918) has introduced the following measure:

$$Q_2 = \frac{\overline{U}(0.05) - \overline{M}(0.25)}{\overline{M}(0.25) - \overline{L}(0.05)} \tag{10}$$

where $\overline{U}(\beta)$ ($\overline{M}(\beta)$, $\overline{L}(\beta)$) is the average of the largest (middle, smallest) $n\beta$ order statistics, and $Q_2 \geq 0$. In case of a symmetrical distribution Q_2 equals 1, in case of a left (right) skewed distribution $0 \leq Q_2 < 1$ ($Q_2 > 1$). Extremely right skewed distributions lead to zero division.

We first used the s-coefficient to measure the skewness of our distributions and then evaluated our results with Hogg's coefficient.

Implementation

The model was implemented in MIMOSE (MIcro and multilevel MOdeling SoftwarE) which is a simulation language that was under development at the Department of Social Science Informatics in Koblenz/Germany (*Möhring,* 1996; *Möhring* and *Ostermann,* 1996). MIMOSE allows the specification of social systems with several levels and several properties respectively. Complex social multilevel models have already been implemented in MIMOSE (*Saam,* 1996, forthcoming).

Simulation results of the baseline model

Simulation time of a multilevel simulation depends crucially on the number of elements and their attributes on the micro level. Almost all variables of our model are located on the micro level. To keep simulation time handy we have simulated a comparatively short time series: the journal *Systeme,* 1988-1993 (*Reiter,* 1994). In 6 years 12 issues had been published. 46 scientists had published articles. 709 different authors had been refered to in overall 1185 citations. The author with the most citations (but no publications in the journal itself) was quoted 18 (19) times.

We have initialized our model with 80 scientists on the micro level. 7% of the scientists are chosen randomly to publish *and* be quoted during the whole simulation time. The rest may only be quoted. Due to the initialization of the author's earlier publications and reputation the model starts with a certain degree of skewness in these variables (but, they are by far not as right skewed as our target distributions). As earlier publications and citations must not be included into the newly evolving distributions of the new scientific field we do not have an initial skewness there. All other variables are set to zero.

Figure 2 gives a graphic representation of this model's publication and citation distribution after 13 time steps (one time step to start the system plus 12 to simulate 12 issues). The overall number of publications (citations) is divided into seven classes of equal size located between the minimum and maximum values (pmin, pmax, cmin, cmax) of the simulated publications (citations). The minimum value is one publication (citation). As we have modeled a stochastic linear system of equations the resulting coefficient of skewness shows some variance. The mean simulated skewness coefficient of publications (citations) is located at about $s_p = 0.81$ ($s_c = 0.84$) which is still an average 0.1 to 0.13 deviation from the empirical distributions.

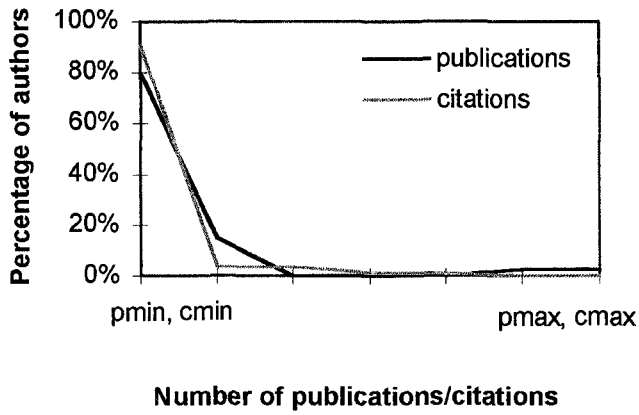

Fig. 2. Simulated distribution of publications and citations (pmin/cmin = 1, pmax/cmax: minimum and maximum values of the simulated publications (citations))

Sensitivity analysis of the baseline model

In order to test the stability of the model's results we ran a small sensitivity analysis*: we calculated elasticity coefficients** indicating the mean % deviation of the skewness output variables (s-measure, see equation (9)) and their standard deviation due to a 10% increase (decrease) of the parameter's values. The results (see Table 2 and 3) prove to be very robust which may be deduced to the linearity of our model. Bold numbers indicate experiments that increase the mean skewness of publications or citations.

Publication and citation distributions in family therapy

We have collected and analyzed data from several German and Austrian scientific journals on family therapy and psychotherapy, e.g., *Psychotherapeut, Psychotherapieforum, Psyche, Systeme,* and *System Familie.* Corrections for mis-spellings and homonymy have been made thoroughly because some of our data sets are relatively small.

*Although we have developed a stochastic linear model a mathematical analysis of the model is not possible due to birth-death processes (*Weidlich* and *Haag*, 1983).

**An overview of sensitivity measures and methods of sensitivity analysis is given in *Chattoe, Saam* and *Möhring,* forthcoming.

Table 2
Sensitivity of publication skewness[a]

experiment elasticity parameter	para. + 10% mean	para. – 10% mean	para. + 10% stand. dev.	para. – 10% stand. dev.
η	-0.22	-0.07	0.09	-0.9
κ	-0.03	-0.35	3.86	11.35
λ	-0.01	-0.02	0.79	0.83
χ	**0.01**	-0.02	-0.36	-0.14
μ	0	0	0	0
β_2	0	0	0	0
β_1	-0.01	-0.01	0.79	-1.09
σ	**0.06**	-0.11	-1.69	1.69
U*	**0.03**	-0.07	0.93	-3.71
α	0	-0.01	0	0.79

[a]Parameter values of the base run are: $\eta = 0.35$, $\kappa = 0.45$, $\lambda = 1.5$, $\chi = 0.9$, $\mu = 0.1$, $\beta_1 = 1.5$, $\beta_2 = 1.0$, $\sigma = 0.07$, $\alpha = 0.8$, and U* = 0.55 (see Table 1).

Table 3
Sensitivity of citation skewness

experiment elasticity parameter	para. + 10% mean	para. – 10% mean	para. + 10% stand. dev.	para. – 10% stand. dev.
η	-0.27	-0.17	3.11	14.21
κ	**0.02**	-0.18	1.43	9.05
λ	**0.03**	-0.02	-2.08	-1.5
χ	**0.02**	-0.01	-2.55	2.18
μ	0	0	0	0
β_2	-0.03	-0.02	0.27	-0.51
β_1	0	**0.01**	1.43	-0.66
σ	0	-0.03	0.44	-0.01
U*	0	0	0.03	0.59
α	0	0	0	-0.07

In general, research on scientometric data has shown that these effects – though they might heavily distort the records of individual contributions – have no substantial influence when large data sets are concerned (*Braun, Glänzel* and *Schubert,* 1990: 37). The publication as well as the citation data are censored to the left: Authors who do not publish cannot be observed. Although our data contain those authors who published but are not quoted these amount only a very small percentage of all authors who could have been quoted but are not. The latter can again not be observed. Therefore we are in lack of the zero-data.

Table 4 and Fig. 3 present four different data sets, two on publications and two on citations. Publications and citations are listed to authors and all co-authors, self citation is excluded. Only proper articles were included. The data sets include

(1) publications by authors of the journal *Psyche* from 1947 to 1992,

(2) publications by authors in the whole field of German-language systemic family therapy (seven journals: *Familiendynamik, Kontext, Partnerberatung, Systhema, Systeme, System Familie, Zeitschrift für systemische Therapie*, 1976-1995),

(3) citations by authors of the journal *Psychotherapeut* (1994), and

(4) citations by authors of the journal *Psychotherapieforum* (1993-1994).

All distributions are extremely right skewed. The publication's coefficients of skewness seem to be somewhat smaller (s_p = 0.946 for *Psyche* 1947-1992, s_p = 0.939 for the whole field 1976-1995) than those for the citations (s_c = 0.948 for *Psychotherapeut* 1994, s_c = 0.952 for *Psychotherapieforum* 1993-1994) confirming *Stephan* and *Levin*'s hypothesis (1991). As stated above, the s-coefficient may not be robust to extreme activities. Therefore, we recalculated the skewness using the robust Q-coefficient which is very sensitive to the right (see Eq. (10)). As can be seen from table 4 the Q-coefficients do not confirm *Stephan* and *Levin*'s hypothesis. The publication skewness of systemic family therapy and the quotation skewness of *Psychotherapeut* rank high whereas the other data sets rank low. We deduce this to the short time series of the first two data sets: They cover 10 years or one year, the latter 46 or 2 years. Important for the reduction in Q is that almost all of the middle 50% of the authors have published only one paper during the first years. Only after some years several of these authors rank higher. This also holds for quotations.

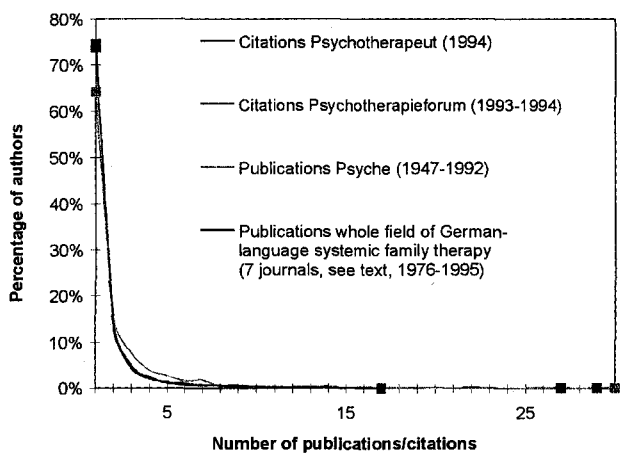

Fig. 3. Empirical distribution of publications and citations: relative numbers

Table 4
Empirical distribution of publications and citations: absolute numbers

A	B	C	D	E
1	668	699	1075	646
2	160	134	192	123
3	81	41	68	44
4	40	23	29	20
5	28	10	19	10
6	17	8	9	10
7	18	6	10	5
8	3	6	8	5
9	1	0	9	3
10	3	5	5	2
11	3	1	4	1
12	2	1	2	2
13	2	1	3	3
14	5	1	1	0
15	0	1	1	0
16	1	0	2	0
17	2	0	1	0
18	1	0	0	0
19	1	1	1	0
20	0	1	0	0
21	0	0	0	0
22	0	0	1	0
23	1	1	1	0
24	0		0	0
25	1		0	0
26	2		0	0
27	0		1	1
28	0			0
29	1			1
30	1			
Sum	1042	940	1442	876
Skewness				
s[a]	0.946	0.939	0.948	0.952
Q_2*[b]	125.5	550.1	835.4	287.2

A: Number of publications/citations per author
B: Publications *Psyche* (1947-1992)
C: Publications whole field of German-language systemic family therapy (7 journals, see text, 1976-1995)
D: Citations *Psychotherapeut* (1994)
E: Citations *Psychotherapieforum* (1993-1994)

[a] See Eq. (9).
[b] See Eq. (10). We used a slightly modified version of Q_2 which we call Q_2*. Q_2 leads to zero division because of the extreme right skewness of the empirical distributions: we exchanged $\bar{M}(0.5)$ for $\bar{M}(0.25)$ in Eq. (10).

Discussion

In the discussion we want to concentrate on three questions: (1) how may our results be evaluated? (2) What are the crucial problems that make modeling and simulation difficult here? (3) What suggestions may be derived for further research?

Results

We have succeded in reproducing right skewed distributions of publications and citations. Applying face validity (Turing test), a standard observer would not be able to find a difference between the simulated and the empirical data, nor to a graphical representation of Lotka's Law. Nevertheless, the simulated coefficients of skewness (s_p = 0.81, s_c = 0.84) are at present somewhat smaller than the empirical ones (s_p = 0.946 for *Psyche* 1947-1992, s_p = 0.939 for the whole field of systemic family therapy 1976-1995, s_c = 0.948 for *Psychotherapeut* 1994, s_c = 0.952 for *Psychotherapieforum* 1993-1994). This difference is not small enough to exclude the change of single hypotheses.

The central question we wanted to answer is how to *explain** the skewness of the distributions of publications and citations. There is a difference between the causal explanation of a phenomenon and its descriptive reproduction. Unfortunately, our model cannot give a causal explanation of the *publication* distribution. All hypotheses that we have incorporated here are descriptive only (transience, Matthew effect). The reasons for transience, e.g., retirement, death, transfer from publishing to teaching, administration or other posts are beyond our model. *Price* is sure that transience is not a result of institutionalization because even in the 17th century the same distribution was prevalent. Instead, institutionalization is supposed to have followed from transience (*Price,* 1986b: 224f). The hyperbolic distribution of publications results from the self-organization of the birth-death process and Matthew's production function. Both productivity and demographic structure result from the mortality rate of the scientists. In case of the *citation* distribution our model is giving some hints to causal explanations. Citations depend on the author's distance to the field and on the accumulated utility of his publications. We had to incorporate a random term into the citation equation which is possibly indicating numerous alternative or additional causal relations.

*Which is a difference to the body of literature which concentrates on the best algorithm to fit the distribution curves (*Ravichandra Rao,* 1995; *Sen* and *Chatterjee,* 1995).

In sum, from the explanation point of view, our findings are somewhat disappointing. If we had incorporated those variables on the micro and the macro level that have been refered to in theoretical studies our model would have been tremendously overdetermined. Having renounced to most of them we mainly succed in a descriptive reproduction of the distributions and to a far lesser extent in their explanation. It is well known and often bemoaned that this corresponds to observations of social reality in general.

Independent of the crucial problems described in the next section we would like to stress what our model contributes to the advancement of the field: Our modeling approach is theory driven. Our model is a system model that represents the interdependence of its variables. When scientists publish and are being cited they act within the whole system. The model reminds us of the social processes that take place within science as a social system. Our final model consists of 8 variables on the micro level, and 11 parameters (see Table 1). Apart from the transience parameters κ and η we know very little about the interpretation, operationalization, and empirical values of the parameters that were necessary to relate the variables to build a complete model. To evaluate their significance we suggest further empirical research and data analysis.

Some restrictions on our results ly in the modeling approach we have chosen. We have developed a multilevel model allowing for micro-macro, macro-macro, and macro-micro relations and self-organizing processes. Excluded were explicit micro-micro relations. In order to explain networking effects in citation they would be inevitably necessary. Here, network simulations would be necessary.

Crucial problems

Following *MacRoberts* and *MacRoberts* (1982) and *Reiter, Steiner* and *Werner* (1997) we have decided that our objects of investigation are publication and citation distributions of scientific fields. It proves difficult to delimit scientific fields because even if all journals that may belong to a scientific field are included books that have been published in that field are omitted (*delimitation problem*). Therefore, our implicit hypothesis was that the publication and citation distribution of the journal section of scientific fields is representative of the distributions of complete scientific fields. The journal section of a scientific field was analysed to obtain the publication and citation data. To explain these data not only all the authors that belong to the field should be conceived of. Authors may be quoted who work in neighbouring fields or work on methodological issues. It is a tricky task, therefore, to conceive of the study population (*study population problem*). All actually publishing scientists as well as all actually quoted scientists are only a selection out of this study population. We do have no

empirical data on it nor do we know how to delimit it. Nevertheless, it is our task to explain this process of selection because it is this process that leads to the observed distributions of publications and citations.

Scientists may have published before a new scientific field has developed. When the field starts which is indicated by the establishment of field journals the initial values of the scientists publications, citations, and reputation are not zero, not even in this field. So the modeler has to conceive of the authors' past as well as the relation between the past data and the new one: probably a productive author of a neighbouring established field is productive in the new field as well or she is frequently quoted because of her reputation there (*past present problem*). Again, we have no data.

Finally there is the *co-authorship problem*: Scientists may produce papers together. Their individual as well as structural properties may add, multiply, or neutralize one another. They may be quoted together. But their reputation is not grouped. The artificial separation of co-authorship data (as is the case in our empirical data) conceals these relations.

Summing up all these crucial problems one may question whether modeling and simulation are suitable then. Actually, there is no other method which could investigate into our phenomenon without reflecting and handling these crucial problems. But only modeling or simulation are able to demonstrate that certain hypotheses definitively reproduce the empirical distribution under certain conditions.

Suggestions for further research

Our first suggestion refers to two crucial problems we have discussed above: Advancing data base technologies should enable us to conceive of complete data sets on publications and citations in scientific fields (*delimitation problem*) as well as the authors' past publications (*past present problem*).

Secondly, we encourage empirical research with statistical means in order to advance the theoretical explanation of bibliometric distributions. It is fruitless to improve modeling and simulation unless more of the above mentioned hypotheses are empirically verified.

Finally, we want to direct attention to the formal generation of right skewed distributions. Readers in statistics always explain the evolution of lognormal (and related) distributions as a result of the multiplication of random variables. We have generated the same kind of distribution by combining a linear model including thresholds and birth-death processes. This may be transferable to many other lognormal distributions in social systems.

This article was significantly improved by criticisms and advice of Wolfgang *Glänzel,* Michael *Möhring,* Iris *Pigeot-Kübler* and Klaus G. *Troitzsh.*

The authors want to express their gratitude to Egbert *Steiner* and Victor *Gotwald* who investigated a lot of time in generating the data sets.

References

BOOKSTEIN, A. (1990a), Informetric distributions. Part I: Unified overview. *Journal of the American Society for Information Science,* 41, 368-375.

BOOKSTEIN, A. (1990b), Informetric distributions. Part II: Resilience to ambiguity. *Journal of the American Society for Information Science,* 41, 376-388.

BOOKSTEIN, A. (1997), Informetric distributions. Part III: Ambiguity and randomness. *Journal of the American Society for Information Science,* 48, 2-10.

BOOKSTEIN, A., B. WRIGHT (1997), Ambiguity in measurement. *Scientometrics,* 40, 369-384.

BRAUN, T., W. GLÄNZEL, A. SCHUBERT (1990), Publication productivity. From frequency distributions to scientometric indicators. *Journal of Information Science,* 16, 37-44.

CHATTOE, E., N. J. SAAM, M. MÖHRING (forthcoming): Sensitivity analysis in the social sciences: problems and prospects. In: G. N. GILBERT, U. MUELLER, R. SULEIMAN, K. G. TROITZSCH (Eds.): *Social Science Microsimulation: Tools for Modeling, Parameter Optimization, and Sensitivity Analysis.* Berlin: Springer.

FAHRMEIR, L. (1997), *Statistik – der Weg zur Datenanalyse.* Berlin: Springer.

FOX, M. F. (1983), Publication productivity among scientists. A critical review. *Social Studies of Science,* 13, 285-305.

GLÄNZEL, W., A. SCHUBERT (1990), The Cumulated Advantage Function. A mathematical formulation based on conditional expectations and its application to scientometric distributions. In: L. EGGHE, R. ROUSSEAU (Eds.). *Informetrics 89/90. Proceedings of the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics,* held in London (Canada), 4-7 July, 1989. Amsterdam: Elsevier Science Publishers: 139-147.

HOGG, R. V. (1974), Adaptive robust procedures. A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Assiciation,* 69, 909-927.

IJIRI, Y., H. SIMON (1977), *Skew distributions and the sizes of business firms.* Amsterdam: North-Holland.

LOTKA, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science.* 16, 317-323.

MACROBERTS, M. H., B. R. MACROBERTS (1982), A re-evaluation of Lotka's Law of scientific productivity. *Social Studies of Science,* 12, 449-450.

MANDELBROT, B. (1959), A note on a class of skew distribution functions: analysis and critique of a paper by H. A. Simon. *Information and Control,* 2, 90-99.

MAYER, K. U. Ed. (1993), *Generationsdynamik in der Forschung.* Frankfurt: Campus.

MERTON, R. K. (1968), The Matthew effect in science. *Science,* 159, 56-63.

MÖHRING, M. (1996), Social science multilevel simulation with MIMOSE. In: K. G. TROITZSCH, U. MUELLER, G. N. GILBERT, J. E. DORAN (Eds.). *Social Science Microsimulation.* Berlin: Springer: 123-137.

MÖHRING, M., R. OSTERMANN (1996), *MIMOSE. Eine funktionale Sprache zur Beschreibung und Simulation individuellen Verhaltens in interagierenden Populationen. Einführung in die Modellierung.* Univ. Koblenz.

PRICE, D. J. DE SOLLA (1986a), *Little science, big science – and beyond.* New York: Columbia Univ. Press. 2nd edition. Originally published 1963.

PRICE, D. J. DE SOLLA (1986b), Studies in Scientometrics, Part 1: Transience and Continuance in Scientific Authorship. In: PRICE 1986a: 206-226. Originally published 1975.

PRICE, D. J. DE SOLLA (1976), A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science,* 27, 292-306.

PRICE, D. J. DE SOLLA (1963), *Little science, big science.* New York: Columbia Univ. Press.

RAO, R. I. K. (1995), A stochastic approach to analysis of distribution of papers in mathematics: Lotka's Law revisited. In: M. E. D. KOENIG, A. BOOKSTEIN (Eds). *Proceedings of the 5th International Conference on Scientometrics & Informetrics*, June 7-10, 1995, Chicago/Illinois: 455-464.

RAO, R. I. K. (1988), Probability distributions and inequality measures for analyses of circulation data. In: L. EGGHE, R. ROUSSEAU (Eds.). *Informetrics 87/88. Selected proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval* held at Diepenbeek, Belgium, 25-28 August 1987. Amsterdam: Elsevier Science Publishers: 231-248.

REITER, L. (1995), Das Konzept der „Klinischen Nützlichkeit". Theoretische Grundlagen und Praxisbezug. *Zeitschrift für systemische Therapie,* 13, 193-211.

REITER, L. (1994), Leitfiguren der Familientherapie und systemischen Therapie. *Psychotherapieforum,* 2, 137-143.

REITER, L., E. STEINER, U. WERNER (1997), Ordnungsstrukturen im Wissenschaftsbetrieb. Untersuchungen und Überlegungen zum Lotka'schen Gesetz der Publikationshäufigkeiten am Beispiel der Psychotherapie. In: G. SCHIEPEK, W. TSCHACHER (Eds.). *Synergetik in Psychologie und Psychiatrie,* Braunschweig: Vieweg. S. 328-343.

SAAM, N. J. (1996), *Computergestützte Theoriekonstruktion in den Sozialwissenschaften. Konzeptbasierte Simulation eines theoretischen Modells am Beispiel militärischer Staatsstreiche in Thailand. Unter Anwendung des Mehrebenen-Ansatzes der Synergetik.* San Diego, Erlangen: Society for Computer Simulation International.

SAAM, N. J. Simulating the Micro-Macro Link: New Approaches to an Old Problem and an Application to Military Coups. *Sociological Methodology,* to be published.

SCHUBERT, A., W. GLÄNZEL (1991), Publication dynamics. Models and indicators. *Scientometrics,* 20, 317-331.

SACHS, L. (1992), *Angewandte Statistik,* 7. ed. Berlin: Springer.

SEN, S. K., S. K. CHATTERJEE (1995), Mean Relative Scatter (MRS) – A linear equation for bibliometric distributions: Further empirical tests. In: M. E. D. KOENIG, A. BOOKSTEIN (Eds). *Proceedings of the 5th International Conference on Scientometrics & Informetrics*, June 7-10, 1995, Chicago, Illinois: 505-514.

STEPHAN, P. E., S. G. LEVIN (1991), Inequality in scientific performance: Adjustment for attribution and journal impact. *Social Studies of Science,* 21, 351-368.

TAGUE, J. (1981), The success-breeds-success phenomenon and bibliometric processes. In: *Journal of the American Society for Information Science,* 32, 280-286.

WEIDLICH, W., G. HAAG (1983), *Quantitative Sociology,* Berlin: Springer.

WHITLEY, R. D. (1972), Kommunikationsnetze in der Wissenschaft. Status und Zitierungsmuster in der Tierphysiologie. In: P. WEINGART (Ed.). *Wissenschaftssoziologie.* Frankfurt: Athäneum Fischer: 188-202.

YABLONSKY, A. I. (1980), On fundamental regularities of the distribution of scientific productivity. *Scientometrics,* 2, 3-34.

ZUCKERMAN, H. (1993), Die Werdegänge von Nobelpreisträgern. In: K. U. MAYER (Ed.). *Generationsdynamik in der Forschung.* Frankfurt: Campus: 59-79.