Nonuniversal power law scaling in the probability distribution of scientific citations

George J. Peterson[a], Steve Pressé[b], and Ken A. Dill[b,1]

[a]Biophysics Graduate Group, and [b]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94158

Contributed by Ken A. Dill, July 26, 2010 (sent for review June 26, 2010)

We develop a model for the distribution of scientific citations. The model involves a dual mechanism: in the *direct mechanism*, the author of a new paper finds an old paper *A* and cites it. In the *indirect mechanism*, the author of a new paper finds an old paper *A* only *via* the reference list of a newer intermediary paper *B*, which has previously cited *A*. By comparison to citation databases, we find that papers having few citations are cited mainly by the direct mechanism. Papers already having many citations ("classics") are cited mainly by the indirect mechanism. The indirect mechanism gives a power-law tail. The "tipping point" at which a paper becomes a classic is about 25 citations for papers published in the Institute for Scientific Information (ISI) Web of Science database in 1981, 31 for *Physical Review D* papers published from 1975–1994, and 37 for all publications from a list of high *h*-index chemists assembled in 2007. The power-law exponent is not universal. Individuals who are highly cited have a systematically smaller exponent than individuals who are less cited.

graph theory | master equation | h-index | preferential attachment | cumulative advantage

Commonly observed in nature and in the social sciences are probability distribution functions that appear to involve dual underlying mechanisms, with a "tipping point" between them. Examples of such probability distributions include the distributions of city sizes (1, 2); fluctuations in stock market indices (3, 4); U.S. firm sizes (5, 6); degrees of Internet nodes (7, 8); numbers of followers of religions (8); gamma-ray intensities of solar flares (9); sightings of bird species (8); and citations of scientific papers (10–13). In these situations, a distribution $p(k)$ may have exponential behavior for small k and a power-law tail for large k. Here we develop a generative model for one such dual-mechanism process, scientific citations, for which databases are large and readily available. Here, k represents the number of citations a paper receives, ranging from zero to hundreds or, sometimes, thousands. $p(k)$ is the distribution of the relative numbers of such citations, taken over a database of papers.

There have been several important studies of power-law tails of distributions, including those involving scientific citations. Price noted that highly cited scientific papers accumulate additional citations more quickly than papers that have fewer citations (14). He called this "cumulative advantage" (CA): the probability that a paper receives a citation is proportional to the number of citations it already contains. Price showed that this rule asymptotically gives a power law for large k. Power-law tails have been widely explored in various contexts and under different names —"the rich get richer," the Yule process (15, 16), the Matthew effect (17), or preferential attachment (18). Barabási and Albert noted that networks, such as the World Wide Web, often have power-law distributions of vertex connectivities, called "scale-free" behavior (18). Their model, called preferential attachment, leads to a fixed power-law exponent of −3. Because many properties of physical systems near their critical points also display power-law behavior, and because such exponents are often *universal* (i.e., independent of microscopic particulars of the system), it raises the question of which power-law distributions have universal exponents and which do not.

The tail of the scientific citations distribution has been fit by various distributions, including power law (10, 19), log-normal (20), and stretched exponential (21). Recently, Clauset, Shalizi, and Newman proposed detailed statistical tests for determining whether various datasets have true power-law tails (8). In agreement with Redner's earlier analysis (10), Clauset et al. confirm that the 1981 dataset studied by Redner is indeed well fit by a power law.

Our interest here is not just in the large-k tails of such distribution functions. We are interested also in the small-k behavior and the tipping point between the two different regions. After all, the preponderance of scientific papers are not cited very commonly. Some previous models have explored both small-k and large-k regimes of citations. In 2001, Krapivsky and Redner developed a rate equation method to obtain solutions for several generalizations of the CA model, including results for nonlinear connection probabilities (22). Krapivsky and Redner proposed a "growing network with redirection" (GNR) for the citations network. They proposed that new papers could randomly cite existing papers, or could be *redirected* to one of the papers in its reference list. The GNR mechanism leads to a distribution with a *nonuniversal* scaling exponent, depending on the value of the redirection parameter. An analysis of this mechanism for arbitrary out-degree distribution was carried out by Rozenfeld and ben-Avraham (23). Recently, Walker et al. proposed a redirection algorithm to rank traffic to *individual* papers, which, instead of an initial random attachment probability, used an exponentially decaying probability of citation, according to the age of the paper (24). There have been many variations proposed of the basic CA model, including CA with error tolerance (25), with an attractiveness parameter (26), with a fitness parameter (27), with memory effects (28), with hierarchical organization (29), with aging nodes (30), and a number of others. A useful overview of CA models, and power laws in general, is by Newman (9).

Here, we develop a model to address three points of particular interest to us. First, existing models focus on the power-law tail. We are interested here in the full distribution function and the nature of the transition, or the tipping point, from one mechanism to the other. Second, we seek a mechanism that illuminates why the rich get richer in scientific citations. Third, a strictly linear attachment rule predicts a single fixed exponent, $\gamma = 3$, where $p(k) \propto k^{-\gamma}$. Here, we ask whether the power-law exponent for scientific citations is a universal constant, as is often observed in the physics of critical phenomena, or whether the power-law exponent for citations is a nonuniversal parameter which varies from one dataset to another.

The two-mechanism model we propose here is similar to the GNR model studied in (22), generalized for an out degree greater than one. A general treatment of the GNR model with arbitrary out-degree distribution was given in (23). Here, we derive $p(k)$

Author contributions: G.J.P., S.P., and K.A.D. designed research; G.J.P. performed research; G.J.P. and S.P. contributed new reagents/analytic tools; G.J.P. and K.A.D. analyzed data; and G.J.P. and K.A.D. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: dill@maxwell.ucsf.edu.

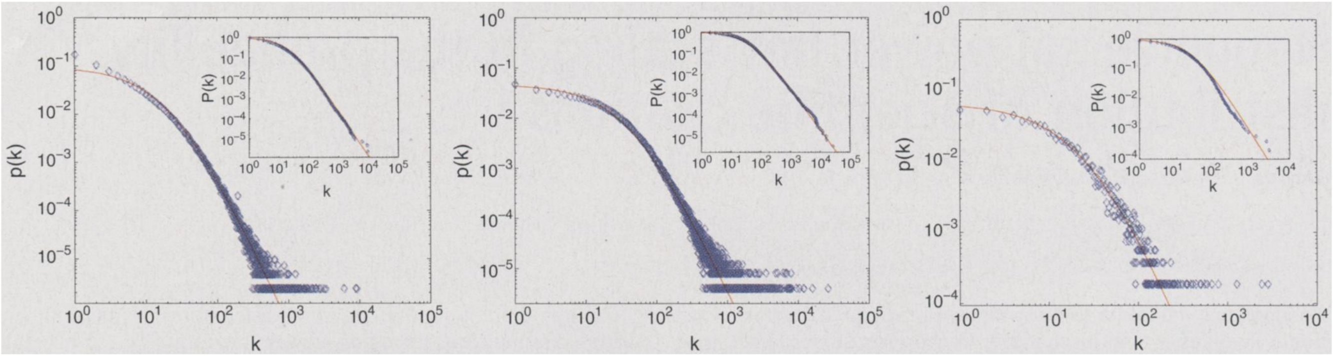

Fig. 1. Probability of receiving exactly k citations (PDF) and at least k citations (CDF, inset) for datasets 1 (left), 2 (center), and 3 (right). Empirical data points are shown as blue diamonds, and best-fit curves as solid red lines.

explicitly for the specific case of a *fixed* out degree, and analyze the tipping-point transition between the two mechanisms. We then fit our $p(k)$ to several citations datasets, and examine how the interactions between the two mechanisms produces different distributions (with different tipping points) for each dataset. By sorting our datasets according to h-index, we show that the scaling exponent, γ, *decreases* systematically with increasing values of h. We interpret the changes in the scaling exponent using a parameter of our model as an increasing bias towards *indirect* citation of well known scientists.

A Two-Mechanism Model

Consider a directed graph on which each node represents a scientific paper. Each edge represents a citation of one paper by another. An outgoing edge indicates *giving* a citation, and an incoming edge indicates *receiving* a citation. At a given time, the graph has N nodes, representing *old* papers that are already part of the graph. At each time step, a new paper is published (a node is added to the graph). Each new paper gives a fixed number of citations, n, distributed among the N old papers. Hence the total number of citations given is Nn, and the total number of citations received is also Nn. In general, we consider situations in which N is large. Let k be the number of incoming links (citations) that a paper has received. For example, a paper that has received no citations from other papers has $k = 0$. Some "classic" papers have attracted more than $k = 1{,}000$ citations. A given collection of papers will have a distribution, $p(k)$, of papers that have received $k = 0,1,2,\ldots$ citations.

We first focus on a particular old paper, paper A. The probability that a new paper will randomly link to paper A is

$$r_{\text{direct}} = \frac{1}{N}. \qquad [1]$$

We call Eq. **1** the *direct mechanism* of citations.*

In addition, scientific papers are also cited by an *indirect mechanism*: the author of the new paper may first find a paper B and learn of paper A *via* B's reference list. On the citation graph, searching through B's reference list is a nearest-neighbor-link mechanism. Suppose there are already k incoming links to paper A. Because there are a total of nN incoming links to all papers, the probability that the author of the new paper randomly finds paper A, *via* the reference list of some other paper is

$$r_{\text{indirect}}(k) = \frac{k}{Nn}. \qquad [2]$$

Given that the author of the new paper has found old paper A, the author will either cite a paper from A's reference list with probability c, or cite A itself with probability $1-c$. If paper A currently has k citations, then the number of citations, $R(k)$, to paper A from a new paper, through either the direct or indirect mechanism, is

$$R(k) = n[(1-c)r_{\text{direct}} + cr_{\text{indirect}}(k)] = \frac{n(1-c)}{N} + \frac{kc}{N}. \qquad [3]$$

Next, we compute the in-link distribution $p(k)$, the fraction of the N papers that have k incoming citations. The total number of papers having k citations is $Np(k)$.[†] We calculate $p(k)$ using a difference equation to express the flows into and out of the bin of papers having k citations for each time step (each time a new node is added). The population of the bin of papers with k citations increases every time a paper with $k-1$ citations receives another citation and decreases every time a paper that already has k citations receives another citation,

$$\begin{aligned} p(k) &= N[R(k-1)p(k-1) - R(k)p(k)] \\ &= [n(1-c) + c(k-1)]p(k-1) - [n(1-c) + ck]p(k). \qquad [4] \end{aligned}$$

Eq. **4** rearranges to:

$$p(k) = \frac{\alpha - 1 + k}{\alpha + 1/c + k} \cdot p(k-1), \qquad [5]$$

where, to simplify the notation, we have defined

$$\alpha = \frac{n}{c} - n. \qquad [6]$$

The equation for $p(0)$ involves no inflow from a lesser bin. Instead, the inflow comes from the addition of a new paper per time step, which is 1 by definition. The outflow term is calculated as for other values of k. Therefore, $p(0) = 1 - n(1-c)p(0)$, which rearranges to:

$$p(0) = \frac{1}{n - nc + 1}. \qquad [7]$$

Substituting in Eq. **7** and applying Eq. **5** recursively gives[‡]

$$p(k) = \frac{1}{\alpha c + 1} \cdot \frac{(\alpha - 1 + k)!(\alpha + 1/c)!}{(\alpha - 1)!(\alpha + 1/c + k)!}. \qquad [8]$$

*Because each new paper will not cite an old paper more than once, the direct probability, Eq. **1**, of the first citation is $1/N$, for the second citation is $1/(N-1)$, and so on, and for the nth citation is $1/(N-n+1)$. For real-world graphs, however, N is of the order of 500,000 and n is around 20. So, we assume $N \gg n$, and $1/(N-n+1) \sim 1/N$. Similarly, the indirect probability, as $Nn \gg n$, Eq. **2** is approximately $k/(Nn-n+1) \sim k/(Nn)$. Note also that, perhaps unrealistically, no special weight is given to the possibility of simultaneously citing both paper A and one of its references.

[†] The in-link distribution should be considered a function of both k and N, $p(k,N)$. However, we find that in the large N limit, the difference between $p(k,N)$ and $p(k,N-1)$ decreases as $1/N$. It is therefore vanishingly small for very large N, and $\lim_{N\to\infty} p(k,N) = p(k)$.

[‡] The factorials in Eq. **8** are understood to be gamma functions for noninteger $1/c$ values. To show that Eq. **8** is normalized, we use

$$\sum_{k=0}^{\infty} \frac{(\alpha-1+k)!}{(\alpha+1/c+k)!} = (\alpha c+1)\frac{(\alpha-1)!}{(\alpha+1/c)!}.$$

Substituting into Eq. **8**, we find that $\sum_k p(k) = 1$, as required.

Table 1. Fitting parameters for datasets 1–3

Dataset	c	n	γ	α	N
1. All 1981 publications	0.454 ± 0.004	17.3 ± 0.3	3.20 ± 0.02	25.0 ± 0.7	415,229
2. High *h*-index chemists	0.517 ± 0.001	42.0 ± 0.1	2.935 ± 0.005	37.0 ± 0.3	245,461
3. *Phys. Rev. D* publications	0.48 ± 0.03	27 ± 2	3.1 ± 0.1	31 ± 5	5,327

When α is sufficiently large, we apply Stirling's approximation to Eq. **8**, which yields

$$p(k) \approx \frac{(\alpha+1/c)^{\alpha+1/c}}{(\alpha c+1)(\alpha-1)^{\alpha-1}}\left(\frac{\alpha-1+k}{\alpha+1/c+k}\right)^{\alpha+k} \times (\alpha-1+k)^{-1}\left(\alpha+\frac{1}{c}+k\right)^{-1/c}. \quad [9]$$

In the large-k tail ($k \gg \alpha$), we have

$$\left(\frac{\alpha-1+k}{\alpha+1/c+k}\right)^{\alpha+k} \approx e^{-(1+1/c)},$$

and

$$(\alpha-1+k)^{-1}\left(\alpha+\frac{1}{c}+k\right)^{-1/c} \approx k^{-(1+1/c)}.$$

Therefore, **9** becomes, in the large-k tail:

$$p(k) \approx \left[\frac{(\alpha+1/c)^{\alpha+1/c}e^{-(1+1/c)}}{(\alpha c+1)(\alpha-1)^{\alpha-1}}\right]k^{-(1+1/c)}. \quad [10]$$

Expression **9** gives our model's prediction for the distribution of citations, expressing both the direct and indirect citation mechanisms. Expression **10** indicates that once a paper's number of citations, k, is large enough, further citations of that paper undergo a sort of runaway growth because there are so many ways to find it through other papers that have already cited it; for scientific citations, the rich get richer. The tipping point where r_{indirect} overtakes r_{direct} happens at

$$k = \alpha. \quad [11]$$

For example, if $c = 1/2$ and the average paper in the database gives out $n = 15$ citations, then after any particular paper in that database has received 15 citations, it will begin to accumulate citations significantly faster than random—it will have "tipped over" into the power-law scaling region. In this region, the power-law exponent,

$$\gamma = 1 + \frac{1}{c}, \quad [12]$$

is determined by the parameter c. Hence, "cumulative advantage" arises in our model because there are more routes (through the reference lists of other papers) for finding a classic paper than for finding a nonclassic paper.

The Datasets

Fig. 1 shows fits to normalized empirical probability distribution functions (PDFs, the probability of receiving *exactly k* citations) and complementary cumulative distribution functions (CDFs, the probability of receiving *at least k* citations), $P(k) = \int_k^\infty p(k')dk'$, for three datasets:

1. Citations of publications catalogued in the ISI Web of Science database in 1981 (10)
2. Citations of publications by authors on a 2007 list of the living highest h-index chemists (33)
3. Citations of publications in the *Physical Review D* journal from 1975–1994 (10)

Datasets 1 and 3 were downloaded from Sidney Redner's website. We gathered dataset 2 from the ISI Web of Science using a Python script. Parameters for these fits are shown in Table 1, and plots of the datasets and best-fit $p(k)$ distributions are shown in Fig. 1. We also sorted dataset 2 by h-index. Parameters for different h-index ranges are shown in Table 2, and fits are shown in Fig. 2. The relation between our estimates of γ and h is shown in Fig. 3. To obtain estimates and 95% confidence intervals of c and n, we used Matlab's implementation of the iteratively reweighted least squares algorithm, using bisquare weights (32). All curve fitting was applied to the raw (not binned or log-transformed) data.

Results

Our model has two parameters: n, the average number of citations given out by all the papers in the database, and c, the chance of citing from a paper's reference list. The model power-law exponent is then fixed by the relationship $\gamma = 1 + 1/c$. Our best fit of dataset 1 gives a value of $n = 17.3 \pm 0.3$, in approximate agreement with the independent estimate of 15.01 found for papers published in 1980 (34). Also, our predicted value of $\gamma = 3.20 \pm 0.02$ agrees with the best-fit power-law exponent previously found by Clauset, of $\gamma = 3.16$ (8). Table 1 shows the best-fit parameter values for the three different datasets.

We explored the $p(k)$ distributions for small groups of scientists, as shown in Fig. 2. We wanted to test an alternate hypothesis that some scientists might publish only low-k papers and others might publish only classic high-k papers. Our limited tests argue against this hypothesis. Fig. 2 indicates that even highly cited scientists have more low-k papers than high-k papers. One reason is that every publication in the scientific literature is new for a while, and requires some time to become highly cited.

Interestingly, the slope of the power-law region differs between the two groups shown in Fig. 2. To examine this difference in more detail, we parsed dataset 2 by h-index (Table 2). The h-index of a scientist is defined as the point where h of the scientist's papers have at least h citations each (31). That is, h is defined by the requirement to satisfy the expression, $Np(h) = h$. There is no simple analytical relationship between a scientist's h-index and the parameters of our model.

From Table 2, we conclude that c increases with h-index, indicating that there is a bias towards selecting papers out of a reference list that were written by scientists who are already very highly cited (Fig. 2). This bias may reflect the tendency of authors who, scanning a paper's references for further information, are more likely to select a paper written by an author of whom they have previously heard. The more highly cited the scientist, the lower his or her power-law exponent (i.e., the fatter the tail); see Fig. 3. The error bars are sufficiently small to indicate that these trends are real, and that there is not a single universal exponent, such as

Table 2. Fitting parameters for *h*-index ranges within dataset 2

h range	c	n	γ	α	N
100+	0.57 ± 0.01	80 ± 3	2.77 ± 0.05	47 ± 4	11,029
90–99	0.54 ± 0.01	77 ± 3	2.86 ± 0.05	57 ± 5	11,476
80–89	0.53 ± 0.01	60 ± 2	2.89 ± 0.04	48 ± 3	15,408
70–79	0.513 ± 0.003	40.6 ± 0.4	2.95 ± 0.01	36.7 ± 0.7	54,236
60–69	0.494 ± 0.002	48.7 ± 0.4	3.02 ± 0.01	51.1 ± 0.9	56,052
54–59	0.493 ± 0.003	34.9 ± 0.3	3.03 ± 0.01	37.0 ± 0.6	44,715
50–53	0.489 ± 0.003	31.3 ± 0.3	3.04 ± 0.01	34.1 ± 0.6	46,421

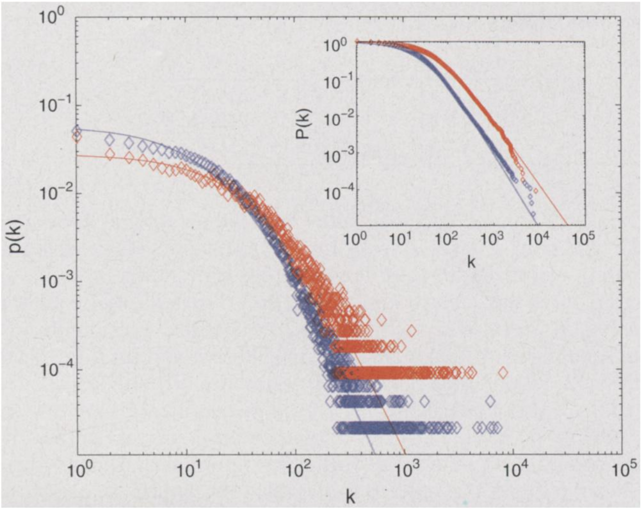

Fig. 2. Comparison of the normalized PDFs and CDFs (inset) for chemists with h = 100+ (red) and chemists with h = 50–53 (blue).

$\gamma = 3$; rather, the exponent depends on the subset of scientists examined. Note that, here, we consider a scientist to have authored a paper if his or her name appears anywhere in the list of authors. An interesting question for future work might be to examine whether this effect is changed by only considering the h-index of each paper's leading and/or corresponding author.

Our model bears some resemblance to Price's application of CA to scientific citations (14). One key difference is that our two parameters both have physical meaning. To avoid the issue of new papers having a citation probability of zero when $k = 0$, Price proposed that the citation probability should be proportional instead to $k + w$, where w is a constant that he refers to as a "fudge factor." He sets $w = 1$, although as later noted by Newman, there does not seem to be a good reason to choose this value (9). The connection rule for our model is given by Eq. **3**, and suggests a simple interpretation: Price's constant arises from random connections, and the tipping point, Eq. **11**, is determined by the average size of the reference lists given out per paper, and the probability of searching through those reference lists.

This two-mechanism model also provides a justification for a CA mechanism. Barabási and Albert remarked that CA only produced a power-law distribution when the connection probability was linearly proportional to k (18), but it was not clear what was special about linearity. The present model presents a possible explanation for the existence of this mechanism, and why the k dependence should be linear: k appears in r_{indirect} because a paper's k incoming citations are represented by k nearest-neighbor links on the graph.

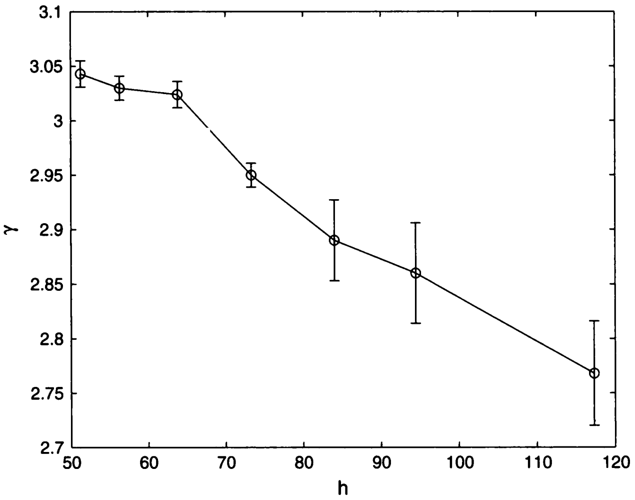

Fig. 3. Power-law exponent γ plotted against h-index for subsets of dataset 2.

Conclusion

We have developed a model of scientific citations, involving both direct and indirect routes to finding and citing papers. This two-mechanism model predicts exponential behavior in the small-k region and power-law tails in the large-k region. One parameter of the model, n, is the average number of citations given out per paper. Our best-fit value of n is consistent with an independent, empirical measure of it made by Biglu (34). Our other parameter, c, defines the power-law exponent, $\gamma = 1 + 1/c$, which is in agreement with data previously evaluated in (8). Two key findings here are: (i) the tipping point for a paper to reach classic-paper status, i.e., its power-law citation region, is about 25 citations for the ISI Web of Science database, and (ii) the power-law exponent is not a universal feature of all scientific citations. The exponent diminishes systematically with increasing h-index of a scientist. Our model describes systems that are governed by random choices in the small-k region, cumulative advantage in the high-k region, and a tipping point between them.

ACKNOWLEDGMENTS. We thank Aéthalie Chabriol for assistance with data acquisition, Kristin Peterson for helpful discussions of curve-fitting methods, and Aaron Clauset, Kingshuk Ghosh, Sergei Maslov, Mark Newman, and Sid Redner for feedback on the manuscript. We thank the ISI Web of Science for their permission to use this data, and Sid Redner for providing a publicly available database of citations. G.J.P. is grateful for financial support from a National Defense Science and Engineering Graduate Fellowship from the Department of Defense, S.P. thanks the Fonds québécois de la recherche sur la nature et les technologies, and K.D. and S.P. appreciate the support from National Institutes of Health GM 34993.

1. George K (1949) Zipf. *Human behavior and the principle of least effort* (Addison-Wesley, Cambridge).
2. Gabaix X (2001) Zipf's law for cities: an explanation. *Q J Econ* 114:739–767.
3. Gopikrishnan P, Plerou V, Amaral LAN, Meyer M, Stanley HE (1999) Scaling of the distributions of fluctuations of financial market indices. *Phys Rev E* 60:5305–5316.
4. Plerou V, Gopikrishnan P, Amaral LAN, Meyer M, Stanley HE (1999) Scaling of the distribution of price fluctuations of individual companies. *Phys Rev E* 60:6519–6529.
5. Okuyama K, Takayasu M, Takayasu H (1999) Zipf's law in income distribution of companies. *Physica A* 269:125–131.
6. Axtell R (2001) Zipf distribution of U.S. firm sizes. *Science* 293:1818–1820.
7. Holme P, Karlin J, Forrest S (2007) Radial structure of the Internet. *P R Soc A* 463:1231–1246.
8. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev* 51:661–703.
9. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46:323–351.
10. Redner S (1998) How popular is your paper? An empirical study of the citation distribution. *Eur Phys J B* 4:131–134.
11. Newman MEJ (2001) The structure of scientific collaboration networks. *Proc Natl Acad Sci USA* 98:404–409.
12. Barabási AL, et al. (2002) Evolution of the social network of scientific collaborations. *Physica A* 311:590–614.
13. Redner S (2004) Citations statistics from 110 years of physical review. *Phys Today* 58:49–54.
14. de Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inform Sci* 27:292–306.
15. Yule GU (1925) A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philos Trans R Soc London B* 213:21–87.
16. Simon HA (1955) On a class of skew distribution functions. *Biometrika* 42:425–440.
17. Merton RK (1968) The Matthew effect in science. *Science* 159:56–63.
18. Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512.
19. Lehmann S, Lautrup B, Jackson AD (2003) Citations networks in high energy physics. *Phys Rev E* 68:026113-1–026113-8.
20. Redner S (2005) Citations statistics from 110 years of Physical Review. *Phys Today* 58:49–54.
21. Laherrere J, Sornette D (1998) Stretched exponential distributions in nature and economy : fat tails with characteristic scales. *Eur Phys J B* 2:525–539.
22. Krapivsky PL, Redner S (2001) Organization of growing random networks. *Phys Rev E* 63:066123-1–066123-14.

23. Rozenfeld HD, ben-Avraham D (2004) Designer nets from local strategies. *Phys Rev E* 70:056107-1–056107-4.
24. Walker D, Xie H, Yan K, Maslov S (2007) Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment* 2007:P06010-1–P06010-10.
25. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382.
26. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of growing networks with preferential linking. *Phys Rev Lett* 85:4633–4636.
27. Bianconi G, Barabási AL (2001) Competition and multiscaling in evolving networks. *Europhys Lett* 54:436–442.
28. Klemm K, Eguíluz VM (2002) Highly clustered scale-free networks. *Phys Rev E* 65:036123-1–036123-5.
29. Ravasz E, Barabási AL (2003) Hierarchical organization in complex networks. *Phys Rev E* 67:026112-1–026112-7.
30. Hajra KB, Sen P (2006) Modeling aging characteristics in citation networks. *Physica A* 368:575–582.
31. Hirsch JE (2005) An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA* 102:16569–16572.
32. Mosteller F, Tukey JW (1977) *Data analysis and regression* (Addison-Wesley, Reading, MA).
33. Peterson A, Schaefer H (2007) H-index ranking of living chemists. *Chemistry World* 4:1–14.
34. Biglu MH (2007) The influence of references per paper in the SCI to impact factors and the Matthew Effect. *Scientometrics* 74:453–470.