ELSEVIER

Physica A 306 (2002) 402–411

PHYSICA A

www.elsevier.com/locate/physa

Aggregation kinetics of popularity

S. Redner

Center for BioDynamics, Center for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215, USA

Abstract

The tools of aggregation kinetics are applied to the "popularity" phenomena of single-lane traffic clustering, and to the growth of a network that mimics citations of scientific publications. In the latter, the network is built by introducing papers (new nodes) one at a time, with preferential linking to more popular previously existing nodes. From the rate equations, the distribution of node degree, as well as various global properties, can be determined easily. A simple extension of the model appears to describe the degree distributions of the world-wide web.

PACS: 02.50.Cw; 05.40.−a; 05.50.+q; 87.18.Sn

Keywords: Traffic clustering; Citation distribution; Growing networks; Degree distribution; Rate equations

1. Introduction

In aggregation, clusters A_i of mass i evolve according to

$$A_i + A_j \stackrel{K(i,j)}{\rightarrow} A_{i+j}\,, \tag{1}$$

where $K(i,j)$ is the rate at which clusters of mass i and mass j form a cluster of mass $k = i + j$. Assuming spatial homogeneity, the system is characterized by the concentrations $c_k(t)$ of aggregates of mass k at time t. Under the law of mass action, these concentrations evolve according to the rate equations

$$\frac{\mathrm{d}c_k}{\mathrm{d}t} = \frac{1}{2}\sum_{i+j=k} K_{ij}c_i c_j - c_k \sum_{j=1}^{\infty} K_{kj}c_j\,. \tag{2}$$

The first term on the right accounts for processes that increase $c_k(t)$, while the second term accounts for loss processes.

E-mail address: redner@buphy.bu.edu (S. Redner).

0378-4371/02/$ - see front matter
PII: S0378-4371(02)00518-6

Due to its fundamental appeal, as well as its broad range of applications [1], there has been intense effort to solve the rate equations for physically relevant reaction rates (see e.g. [2]). Generally, the nature of these solutions depends on: (i) the *homogeneity index* λ of the reaction rate; this is defined by $K(ai, aj) \sim a^{\lambda} K(i, j)$, and (ii) a secondary index v, defined by $K(1, j) \sim j^{v}$, that characterizes the relative importance of large–large and large–small interactions. In many such situations the cluster size distribution exhibits scaling, that is

$$c_k(t) \sim s(t)^{-2} f(k/s(t)) \tag{3}$$

with typical size $s(t) \sim t^{1/(1-\lambda)}$ for $\lambda < 1$. Further, the scaled size distribution is a power law, when $\lambda > v$, and is a localized peak, for $\lambda < v$ [3].

We now investigate two measures of popularity: traffic clustering and the growth of citation-driven networks. The basic message is that the tools of aggregation kinetics are both convenient and powerful in determining the time evolution of these systems.

2. Traffic clustering on suicide alley

Consider traffic on a single-lane road with no passing. Each vehicle has an intrinsic speed that is drawn from a distribution $P_0(v)$. When a faster vehicle overtakes a slower one, the former then moves at the speed of the latter (Fig. 1). The inspiration for this model comes from an infamous 13-mile stretch of Massachusetts highway on Cape Cod known as "suicide alley". Here, two lanes in both directions each are constricted to a single lane, along which no passing is allowed.

As anyone who drives on this type of road has experienced, homogeneous traffic entering such a constriction becomes strongly clustered upon reaching the far end. We can find the typical cluster size $n(v)$ and speed $v(t)$ at time t by a simple dimensional argument [4]. Since the typical distance ℓ between clusters varies as $\ell \sim vt$, the typical number of cars in a cluster is proportional to this distance, yielding $n \sim \ell \sim vt$.

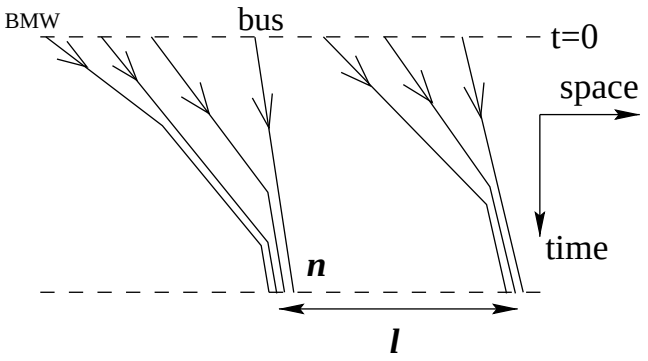

Fig. 1. Space–time evolution of single-lane traffic with no passing. When a slow cluster is overtaken, the combined cluster moves with the speed of the slow cluster.

To find the typical speed, we relate the cluster size to its speed. The probability of finding a fast car (with speed $\geqslant v$) is $Q_+(v) = \int_v^\infty P_0(v')\,dv'$, while $Q_- = 1 - Q_+$ is the complementary slow car probability. Then the typical size of a cluster of speed v may be obtained from $n(v) = \sum_1^\infty kQ_-Q_+^k = Q_+/Q_- \sim v^{-(1+\mu)}$, under the general assumption that $P_0(v) \sim v^\mu$ as $v \to 0$. Here, the speed of the slowest car has been subtracted off so that the speed distribution extends to $v = 0$. Combining $n(v) \sim v^{-(1+\mu)}$ with $n \sim vt$ gives [4]

$$n \sim t^{(\mu+1)/(\mu+2)}, \quad v \sim t^{-1/(\mu+2)}\,. \tag{4}$$

Thus, a car that enters suicide alley slows down considerably and typically becomes part of a large cluster by the far end.

3. Structure of growing networks

We now investigate a growing network, introduced in Ref. [5], that was inspired by modeling the distribution of scientific citations. The model, however, has wider applications, with the world-wide web [6,7] as one notable example. A characteristic feature of these systems is that the node degree distribution $N_k(t)$—the average number of nodes with k links—is a power law [5,12] (some of the results in Ref. [12] were also obtained in Ref. [13]). We can account for this behavior by the rate equation approach.

Some citation data motivates our discussion. From ISI data of 783,339 papers (with 6,716,198 citations) published in 1981 and cited between 1981 and June 1997 [9], 64 papers are cited $\geqslant 1000$ times, 282 papers are cited $\geqslant 500$ times, and 2103 papers are cited $\geqslant 200$ times. Conversely, 633,391 articles are cited $\leqslant 10$ times and 368,110 are uncited! More relevant for this presentation is that the citation distribution itself appears to be a power law with exponent -3 (Fig. 2) [10]; however, Ref. [11] suggests a stretched exponential form.

A crude but effective model [5] for this citation dynamics is illustrated in Fig. 3. Nodes are introduced one at a time and each links to one earlier node. In terms of citations, nodes are publications, and a link from one paper to an earlier one is a citation. The key ingredient that determines the network structure is the *attachment kernel* A_k, namely, the probability that a new node links to an existing node with k links. While much attention has been focused on the linear attachment kernel $A_k = k$, the rate equation approach easily gives the solution for general attachment kernels.

3.1. The degree distribution

The rate equations for the degree distribution $N_k(t)$ are [12]

$$\frac{dN_k}{dt} = [A_{k-1}N_{k-1} - A_kN_k]/A + \delta_{k1}\,. \tag{5}$$

The first term on the right accounts for processes in which a node with $k-1$ links is connected to the new node, thus increasing N_k by one. This happens with probability A_{k-1}/A, where $A(t) = \sum_{j\geqslant 1} A_jN_j(t)$ is the appropriate normalization factor.

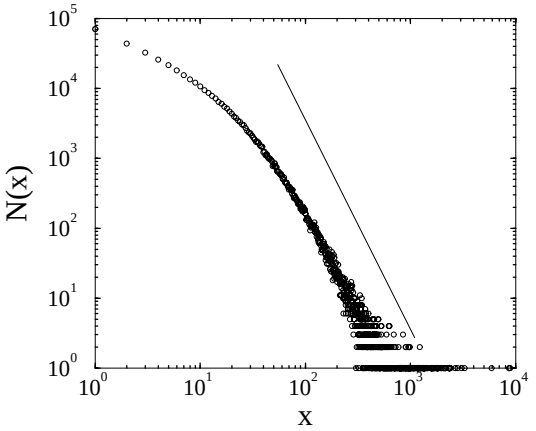

Fig. 2. ISI citation distribution data for the number of papers $N(x)$ with x citations on a double logarithmic scale. The straight line has slope -3.

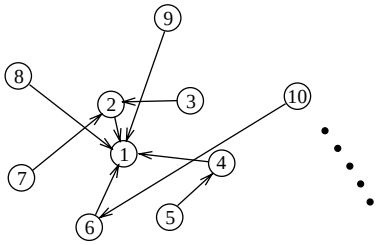

Fig. 3. Growing network. Nodes are added sequentially and a single link joins a new node to an earlier node. Node 1 has degree 5, node 2 has degree 3, nodes 4 and 6 have degree 2, and the remaining nodes have degree 1. Node 1 is the "ancestor" of 6, while 10 is the "descendant" of 6.

A corresponding role is played by the second (loss) term on the right-hand side. The last term accounts for the introduction of new nodes with no incoming links.

It is easy to verify that the moments of the degree distribution, $M_n(t)=\sum_{j\geqslant 1} j^n N_j(t)$, increase linearly with time for $0 \leqslant n \leqslant 1$. In fact, $M_0(t)$ is the total number of nodes and it grows as $M_0(t) = M_0(0) + t$. Similarly, the first moment gives the number of link endpoints which grows as $M_1(t) = M_1(0) + 2t$. The first two moments are thus *independent* of the attachment kernel.

For kernels of the form $A_k = k^\gamma$ with $0 \leqslant \gamma \leqslant 1$, both the degree distribution and $A(t)$ grow linearly with time. By substituting $N_k(t) = t\, n_k$ and $A(t) = \mu t$ into Eq. (5), we obtain the recursion relation $n_k = n_{k-1}A_{k-1}/(\mu + A_k)$ and $n_1 = \mu/(\mu + A_1)$. These yield the formal expression

$$n_k = \frac{\mu}{A_k} \prod_{j=1}^{k} \left(1 + \frac{\mu}{A_j}\right)^{-1} . \tag{6}$$

To complete the solution, we need the amplitude μ which can be found numerically by combining the definition $\mu = \sum_{j \geqslant 1} A_j n_j$ and Eq. (6). The final asymptotic result is [12,13]

$$n_k \sim \begin{cases} k^{-\gamma} \exp[-\mu(\frac{k^{1-\gamma} - 2^{1-\gamma}}{1-\gamma})], & 0 \leqslant \gamma < 1, \\ \begin{cases} k^{-3}, & A_k = k, \\ k^{-\nu}, \ \nu > 2, & A_k \sim k, \end{cases} & \gamma = 1, \\ \text{best seller}, & 1 < \gamma < 2, \\ \text{bible}, & 2 < \gamma. \end{cases} \tag{7}$$

The degree distribution decays exponentially for $\gamma = 0$, while for $0 < \gamma < 1$, this distribution exhibits a robust stretched exponential decay. For the strictly linear kernel $A_k = k$, the solution to Eq. (5) is $n_k = 4/[k(k+1)(k+2)]$. For asymptotically linear attachment kernels $A_k \sim k$, the situation is more delicate, as the exponent of the degree distribution is *non-universal* and depends on microscopic details of A_k. From Eq. (6), we find $n_k \sim k^{-\nu}$, where the exponent $\nu = 1 + \mu$ can be tuned to *any* value larger than 2 [12,14].

For super-linear kernels, one node links to almost every other node. For $\gamma > 2$, all but a finite number of nodes are linked to a "bible" that has the rest of the links. For $1 < \gamma < 2$, the number of nodes with a small number of links grows slower than linearly in time while a "best seller" has the rest of the links. There is also an accompanying infinite sequence of transitions as γ ranges between 1 and 2. Generally for $(m+1)/m < \gamma < m/(m-1)$, the number of nodes with $> m$ links is finite, while $N_k \sim t^{k-(k-1)\gamma}$ for $k \leqslant m$.

3.2. Node degree correlations

An important advantage of the rate equation approach is that we can obtain properties beyond the single-particle degree distribution with minimal additional effort. One such property is the correlation between degrees of connected nodes [14]. These develop naturally because a node with large degree is likely to be old. Thus, its ancestor is also old and hence also has a large degree. Define $C_{kl}(t)$ as the number of nodes of degree k that attach to an ancestor node of degree l. For example, in the network of Fig. 3, there are $N_1 = 6$ nodes of degree 1, with $C_{12} = C_{13} = C_{15} = 2$. There are also $N_2 = 2$ nodes of degree 2, with $C_{25} = 2$, and $N_3 = 1$ nodes of degree 3, with $C_{35} = 1$.

For the linear attachment kernel, the degree correlation $C_{kl}(t)$ evolves according to the rate equation

$$M_1 \frac{\mathrm{d}C_{kl}}{\mathrm{d}t} = [(k-1)C_{k-1,\,l} - kC_{kl}] + [(l-1)C_{k,\,l-1} - lC_{kl}] + (l-1)C_{l-1}\,\delta_{k1}\,. \tag{8}$$

The first two terms on the right account for the change in C_{kl} due to the addition of a link onto a node of degree $k-1$ (gain) or k (loss), respectively, while the second set of terms gives the change in C_{kl} due to the addition of a link onto the ancestor node. Finally, the last term accounts for the gain in C_{1l} due to the addition of the new node.

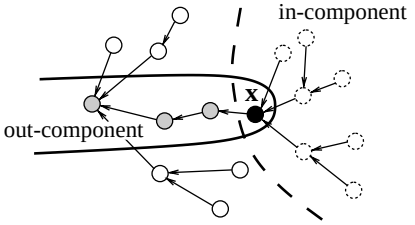

Fig. 4. In-component and out-component of node **x**.

Once again $C_{kl} \to tc_{kl}$; this reduces Eq. (8) to time-independent recursion relations, whose solution in the scaling regime $k \to \infty$ and $l \to \infty$ is [14]

$$c_{kl} \to \begin{cases} 16\,(l/k^5) \text{ when } l \ll k\,, \\ 4/(k^2\,l^2) \text{ when } l \gg k\,. \end{cases} \tag{9}$$

The basic feature is that the degree correlation does not factorize; that is, $c_{kl} \neq n_k n_l = (k\,l)^{-3}$.

3.3. Global properties

In the context of citations, several global properties are of interest. One is obtained by taking the reference list of this paper, plus the reference lists of all these cited papers, etc. In a growing network, this citation ancestry is the *out-component* with respect to a given node **x**—the set of nodes that can be reached by following directed links that emanate from **x** (Fig. 4). In a similar vein, we could track all publications that cite this work, plus all papers that cite these daughter papers, etc. This progeny comprises the in-component to node **x**—the set from which **x** can be reached by following a path of directed links on the network.

The rate equations for these two components can be written and solved in much the same spirit as the degree correlation [14]. From these, the number of in-components with s nodes at time t, $I_s(t)$, has the generic asymptotic behavior

$$I_s(t) = t/[s(s+1)]\,. \tag{10}$$

The salient feature is that there is a robust s^{-2} tail, independent of the form of the attachment kernel. The result agrees with measurements of the web [7].

The complementary out-component from each node is related to an underlying network "genealogy" A genealogical tree may be built by taking generation $g = 0$ to contain the initial node. Nodes that attach to those in generation g form generation $g+1$. For example, the network of Fig. 3 has five nodes in generation $g = 1$ and four in $g = 2$, leading to the genealogical tree of Fig. 5.

By construction, the number O_s of out-components with s nodes equals L_{s-1}, the number of nodes in generation $s-1$ in the genealogical tree. We may compute $L_g(t)$ by noting that $L_g(t)$ increases when a new node attaches to a node in generation $g-1$. For the uniform attachment kernel, this occurs with rate L_{g-1}/M_0, where $M_0(t) = 1 + t$

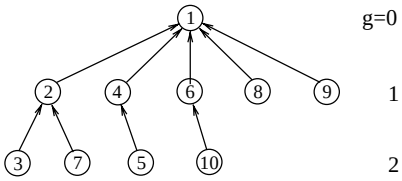

Fig. 5. Genealogy of the network in Fig. 3. The indices indicate when a node is introduced, while the ancestor determines the generation number of the new node.

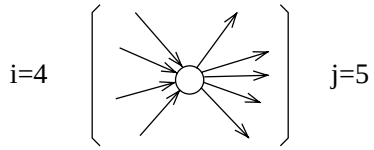

Fig. 6. A node with in-degree $i = 4$, out-degree $j = 5$, and total degree 9.

is the number of nodes. This gives a simple differential equation for $L_g(t)$ with solution $L_g(\tau) = \tau^g/g!$, where $\tau = \ln(1+t)$. Thus at fixed (large) time, the generation size grows with g when $g < \tau$, and then decreases. The number O_s of out-components with s nodes simply equals

$$O_s(\tau) = \tau^{s-1}/(s-1)! \tag{11}$$

As a useful corollary, since the genealogical tree contains approximately $e\tau$ generations at time t, the network diameter $D \approx 2e\tau \approx 2e \ln N$, where N is the number of nodes.

3.4. Joint in- and out-degree distribution

In the world-wide web, link directionality is relevant and the node degree should be resolved into the *in-degree*—the number of incoming links to a node, and the complementary *out-degree* (Fig. 6). Measurements on the web indicate that these two distributions are power laws with different exponents [8]. We now determine these distributions by the rate equation approach.

To generate a non-trivial, out-degree distribution and distinct in- and out-degree distributions spontaneously, we consider the following generalized network (the earliest network model of the type was proposed in Ref. [16]) [17] where growth occurs by two processes (Fig. 7):

(i) With probability p, a new node is introduced and attaches to an earlier node. The attachment probability depends only on the in-degree of the target.
(ii) With probability $q = 1 - p$, a new link is created between already existing nodes. The choices of the originating and target nodes depend on the out-degree of the originating node and the in-degree of the target.

The average node degree can be determined simply. Let $N(t)$ be the total number of nodes, and let $I(t)$ and $J(t)$ be the total in- and out-degree, respectively. According to

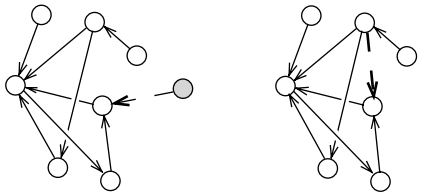

Fig. 7. The growth processes of: (i) node creation (shaded) plus attachment, and (ii) link creation (dashed).

the elemental growth processes, these degrees evolve according to one of the following at each step:

$$(N,I,J) \to \begin{cases} (N+1,I+1,J+1) & \text{with probability } p\,, \\ (N,I+1,J+1) & \text{with probability } q\,, \end{cases} \tag{12}$$

so that $N(t) = pt$, and $I(t) = J(t) = t$. Thus the average in- and out-degrees, $\mathbb{D}_{\text{in}} \equiv I(t)/N(t)$ and $\mathbb{D}_{\text{out}} \equiv J(t)/N(t)$, are both equal to $1/p$.

For the joint degree distribution, we need: (i) the *attachment rate* $A(i,j)$—the probability that a new node links to an existing node with i incoming and j outgoing links, and (ii) the *creation rate* $C(i_1,j_1|i_2,j_2)$—the probability of adding a new link from an (i_1,j_1) to an (i_2,j_2) node. Interesting behavior arises for *linear–bilinear* rates $A_i = i+\lambda$, and $C(j,i) = (i+\lambda)(j+\mu)$, with $\lambda > 0$ and $\mu > -1$. The latter conditions ensure that the rates are positive for all attainable in- and out-degree values, $i \geqslant 0$ and $j \geqslant 1$.

The joint degree distribution, $N_{ij}(t)$, defined as the average number of nodes with i incoming and j outgoing links, obeys the rate equation

$$\frac{\mathrm{d}N_{ij}}{\mathrm{d}t} = (p+q)\left[\frac{(i-1+\lambda)N_{i-1,j} - (i+\lambda)N_{ij}}{I+\lambda N}\right] + q\left[\frac{(j-1+\mu)N_{i,j-1} - (j+\mu)N_{ij}}{J+\mu N}\right] + p\delta_{i0}\delta_{j1}\,. \tag{13}$$

The first group of terms on the right accounts for the changes in the in-degree of target nodes by simultaneous creation of a new node and link (probability p) or by creation of a new link only (probability q). For example, the creation of a link to a node with in-degree i leads to a loss in the number of such nodes. This occurs with rate $(p+q)(i+\lambda)N_{ij}$, divided by the appropriate normalization factor $\sum_{i,j}(i+\lambda)N_{ij} = I+\lambda N$. Similarly, the terms in the second group of terms account for out-degree changes. These occur due to the creation of new links between already existing nodes—hence the prefactor q. The last term accounts for the introduction of new nodes with no incoming links and one outgoing link. This rate equation conserves the total number of nodes, $N = \sum_{i,j} N_{ij}$, while the total in- and out-degrees, $I = \sum_{i,j} iN_{ij}$ and $J = \sum_{i,j} jN_{ij}$, obey $\dot{I} = \dot{J} = 1$.

Because the N_{ij} grow linearly with time, we use $N_{ij}(t) = t\,n_{ij}$, as well as $N = pt$ and $I = J = t$, in Eq. (13) to yield algebraic recursion relations for n_{ij}. The asymptotic

behavior of the in- and out-degree distributions, I_i and O_j, respectively, are found to be the distinct power laws [15],

$$I_i \sim i^{-\nu_{\rm in}}, \qquad \nu_{\rm in} = 2 + p\lambda\,, \tag{14}$$

$$O_j \sim j^{-\nu_{\rm out}}, \qquad \nu_{\rm out} = 1 + q^{-1} + \mu p q^{-1} \tag{15}$$

with $\nu_{\rm in}$ and $\nu_{\rm out}$ necessarily > 2. These can be tuned to the observed values for the web, $\nu_{\rm in} \approx 2.1$, $\nu_{\rm out} \approx 2.7$ [8], by using the fact that p is fixed by the constraint that $p^{-1} = \mathbb{D}_{\rm in} = \mathbb{D}_{\rm out} \approx 7.5$, and then choosing $\lambda = 0.75$ and $\mu = 3.55$. The fact that these adjustable parameters are of order 1 indicates that the linear–bilinear rate is a viable working hypothesis.

4. Summary

In this presentation, I have tried to highlight how the rate equations of aggregation give a powerful and appealing way to obtain many geometrical properties of growing networks. For the degree distribution, we find a stretched exponential, power law, or a "winner take all" situation, depending on whether the exponent in the attachment rate $A_k \sim k^\gamma$ is $\gamma < 1$, $=1$ or > 1. More general properties can be obtained by natural extensions of the basic approach. There are many other applications of the rate equation approach to growing network phenomena that can be envisioned.

Acknowledgements

It is a pleasure to thank Eli Ben-Naim, Paul Krapivsky, Francois Leyvraz, and Geoff Rodgers for the pleasant collaborations that led to the work reported here. I am also grateful to NSF Grants INT9600232 and DMR9978902 for financial support.

References

[1] S.K. Friedlander, Smoke, Dust and Haze: Fundamentals of Aerosol Behavior, Wiley, New York, 1977.; F. Family, D.P. Landau (Eds.), Kinetics of Aggregation and Gelation, North-Holland, Amsterdam, 1984.
[2] R.L. Drake, in: G.M. Hidy, J.R. Brock (Eds.), Topics in Current Aerosol Research, Vol. III, Part 2, Pergamon, Oxford, UK, 1972, p. 201;
M.H. Ernst, in: E.G.D. Cohen (Ed.), Fundamental Problems in Statistical Physics VI, Elsevier, New York, 1985.
[3] P.G.J. van Dongen, M.H. Ernst, Phys. Rev. Lett. 54 (1985) 1396.
[4] E. Ben-Naim, P.L. Krapivsky, S. Redner, Phys. Rev. E 50 (1994) 822.
[5] A.L. Barabási, R. Albert, Science 286 (1999) 509.
[6] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, R. Lukose, Science 280 (1998) 95;
B.A. Huberman, L.A. Adamic, Nature 401 (1999) 131.
[7] G. Caldarelli, R. Marchetti, L. Pietronero, Europhys. Lett. 52 (2000) 386.
[8] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, in: Proceedings of the International Conference on Combinatorics and Computing, Lecture Notes in Computer Science, Vol. 1627, Springer, Berlin, 1999;

M. Faloutsos, P. Faloutsos, C. Faloutsos, Comp. Commun. Rev. 29 (4) (1999) 251;
A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Comput. Networks 33 (2000) 309.
[9] Science Citation Index Journal Citation Reports, Institute for Scientific Information, Philadelphia, Web site: http://www.isinet.com/welcome.html.
[10] S. Redner, Eur. Phys. J. B 4 (1998) 131.
[11] J. Laherrere, D. Sornette, Eur. Phys. J. B 2 (1998) 525.
[12] P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. 85 (2000) 4629.
[13] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Phys. Rev. Lett. 85 (2000) 4633.
[14] P.L. Krapivsky, S. Redner, Phys. Rev. E 63 (2001) 066123.
[15] P.L. Krapivsky, G.J. Rodgers, S. Redner, Phys. Rev. Lett. 86 (2001) 5401.
[16] H.A. Simon, Biometrica 42 (1955) 425, to describe word frequency.
[17] R. Albert, A.-L. Barabási, Phys. Rev. Lett. 85 (2000) 5234;
S.N. Dorogovtsev, J.F.F. Mendes, Europhys. Lett. 52 (2000) 33.