Organization of growing random networks

P. L. Krapivsky and S. Redner
Center for BioDynamics, Center for Polymer Studies, and Department of Physics, Boston University, Boston, Massachusetts 02215
(Received 7 November 2000; published 24 May 2001)

The organizational development of growing random networks is investigated. These growing networks are built by adding nodes successively, and linking each to an earlier node of degree k with an attachment probability A_k. When A_k grows more slowly than linearly with k, the number of nodes with k links, $N_k(t)$, decays faster than a power law in k, while for A_k growing faster than linearly in k, a single node emerges which connects to nearly all other nodes. When A_k is asymptotically linear, $N_k(t) \sim t k^{-\nu}$, with ν dependent on details of the attachment probability, but in the range $2 < \nu < \infty$. The combined age and degree distribution of nodes shows that old nodes typically have a large degree. There is also a significant correlation in the degrees of neighboring nodes, so that nodes of similar degree are more likely to be connected. The size distributions of the in and out components of the network with respect to a given node—namely, its ''descendants'' and ''ancestors''—are also determined. The in component exhibits a robust s^{-2} power-law tail, where s is the component size. The out component has a typical size of order $\ln t$, and it provides basic insights into the genealogy of the network.

DOI: 10.1103/PhysRevE.63.066123 PACS number(s): 02.50.Cw, 05.40.−a, 05.50.+q, 87.18.Sn

I. INTRODUCTION

Networks of many interacting units play an important role in epidemiology, ecology, gene regulation, neural networks, and many other fields [1–3]. In many studies of these networks, the number of nodes is considered to be fixed, and the presence of a link between two nodes is treated as a random event independent of the other links. These assumptions lead naturally to random graph models [4,5]. While these models have a rich behavior and considerable utility, they are not necessarily appropriate for describing *growing* networks, where the addition of nodes and links may depend on local features of the network where the growth event is taking place.

Typical examples of such growing networks include transportation or electrical distribution systems, where growth occurs in response to population-driven demands. Two currently appealing examples are the distribution of scientific citations and the structure of the worldwide web. For both these examples there are now considerable data available, in spite of the very rapid growth of these systems. In the former case, one may consider papers to be nodes of a graph and citations to be links. The structure of the resulting ''citation graph'' was originally studied by Lotka in 1926 [6], and then by many others [7–13]. The basic feature of this citation distribution is that it appears to have a relatively steep power-law tail; thus most papers are minimally cited while highly cited papers are rare.

Similarly, in the web graph, much structural data were recently obtained [14–21] which suggest that the number of nodes with k links has a power-law tail, with an exponent that is somewhat larger than 2. This power-law tail again corresponds to the basic fact that most nodes of the web graph are unimportant, while a relatively small number of nodes garner a large fraction of ''hits.'' Due to the qualitative similarities between the citation and web graphs, insights developed in the field of bibliometrics [9] have been applied to help understand the structure of the web [22].

Because of the dynamic nature of the citation and web graphs, it is not surprising that their topologies at any fixed time are very different from classical random graphs. In distinction to the power-law degree distributions of the citation and web graphs, random graphs have a Poisson node degree distribution. Here ''node degree'' is defined as the number of links at a node. To overcome the shortcomings of random graphs in describing the dynamic natures of these systems, both ''small-world'' networks [23,24] and growing random network models [20,25–28] were recently introduced. The former are aimed at understanding the relatively small diameter of large graphs of socially interacting units, while the latter seek to understand the growth dynamics.

In this paper, we provide a comprehensive quantitative description of a simple *growing network* (GN) model. Our results are based on an analysis of the rate equations for the densities of nodes of a given degree. This approach bears many similarities to the rate equations for the kinetics of aggregation. The rate equations for the evolution of growing networks are relatively simple, and the results that emerge are comprehensive. Thus it appears that the rate equation method is better suited for probing the structure of growing random networks compared to the classical approaches for analyzing random graphs, such as probabilistic [4] or generating function [5] techniques. The rate equation approach also has the advantage that it can be adapted to other evolving graph systems, including networks with the addition and deletion of nodes and links, as well as networks with link rewiring.

We will specifically investigate two types of models: (a) a GN in which nodes are added one at a time, and a link is established with a pre-existing node according to an attachment probability A_k which depends only on the degree of the target node (Fig. 1); and (b) a GN with redirection (GNR), in which the newly created link can be redirected to the ''ancestor'' node of the original target node. An important feature of these models is that the links are *directed*, and the resulting graphs have a simple treelike topology. The moti-

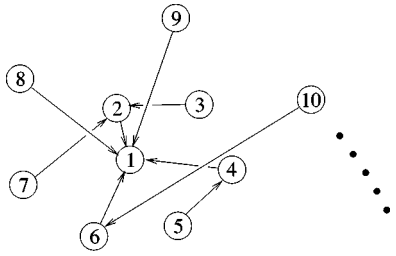

FIG. 1. Schematic illustration of the evolution of the growing random network. Nodes are added sequentially, and a single link joins the new node to an earlier node. In this example, node 1 has degree 5, node 2 has degree 3, nodes 4 and 6 have degree 2, and all the remaining nodes have degree 1. Also note that node 1 is the ''ancestor'' of node 6, while node 10 is the descendant of node 6.

vation for the GNR model is that this redirection process roughly mimics how we might (lazily) construct the references to this paper. In addition to papers that we peruse and cite directly, we are also likely to incorporate some of the references within these papers as part of our reference list. A related ''copying'' process has also been invoked to describe the organization of the web [15].

One of our primary results is that for asymptotically linear attachment kernels, $A_k \sim k$ as $k \to \infty$, the degree distribution of the GN has a power-law form $N_k(t) \sim t k^{-\nu}$, with ν tunable in the range $2 < \nu < \infty$. By choosing the control parameters of our model in a plausible manner, it is then easy to reproduce quantitative observations about the node degree distribution of the web graph.

In Sec. II, we define the GN and GNR models precisely, and then determine their node degree distributions in Sec. III by the rate equation approach. Different distributions arise in the GN model which depend on the asymptotic behavior of the attachment probability as a function of node degree. In Sec. IV, we investigate the joint age-degree distribution, and find (not surprisingly) that ''old'' nodes are typically more highly connected. In Sec. V, we study the correlations which develop between the degrees of connected nodes as the network grows. In Sec. VI, we study a more global measure of the network, namely, the size distributions of the in component and the out component. With respect to a given node **x**, the in component is the set of nodes which can reach node **x** via a directed path of links. Conversely, the out component is the set nodes which can be reached from node **x** via a directed path. The former exhibits a robust power-law size distribution which appears to be independent of the attachment probability. The latter distribution predicts a network ''diameter'' which grows as ln*t*, and also provides basic insights into the genealogy of the network. We conclude in Sec. VII.

II. MODELS

A. Growing network

In the GN, we introduce a new node at each time step, and link it to one of the earlier nodes in the network (Fig. 1). This leads to a network which has a topology of a (directed) tree graph. In terms of citations, we may interpret the nodes as publications, and the directed link from one paper to another as a citation to the earlier publication. In terms of the web graph, nodes are web pages and directed links are hyperlinks. We will refer to the node to which the link is directed as the *ancestor* of the current node.

As the network grows, a degree distribution $N_k(t)$, defined as the average number of nodes with k links ($k-1$ incoming and 1 outgoing), builds up. The initial node is unique, as it does not have an outgoing link. The basic ingredient which determines the structure of the network is the *attachment kernel* A_k, defined as the probability that the newly introduced node links to an existing node which already has k links. On general grounds, this attachment kernel should be a nondecreasing function of k, and natural scenarios are attachment kernels with a power-law dependence on k. For the linear kernel, the GN reduces to the scale free model introduced by Barabási and Albert [20] and further investigated in Refs. [25–27].

The general homogeneous model $A_k = k^\gamma$, with $\gamma \geqslant 0$, was investigated in Ref. [28], where it was found that the degree distribution $N_k(t)$ crucially depends on the value of γ. For $\gamma < 1$, the linking probability grows weakly with the node popularity, and $N_k(t)$ decreases as a stretched exponential in k for any t. The complementary case of $\gamma > 1$ leads to a phenomenon akin to gelation [29] in which a single ''gel'' node links to nearly every other node. For $\gamma > 2$, this phenomenon is so extreme that the number of links between other nodes is finite in an infinite graph. We shall show that these results also apply for the more general situation where $A_k \sim k^\gamma$ as $k \to \infty$, in addition to the strictly homogeneous situation where $A_k = k^\gamma$.

The borderline case of an asymptotically linear attachment kernel $A_k \sim k$, is particularly intriguing as it leads to $N_k \sim k^{-\nu}$, with the exponent ν tunable to *any* value larger than 2 depending on finer details of the attachment kernel. In particular, the strictly linear kernel, $A_k = k$ leads to $\nu = 3$. However, by changing the value of a single attachment probability, for example $A_1 = \alpha$ and $A_k = k$ for $k > 2$, any value of $\nu > 2$ is possible. This sensitivity of the asymptotic behavior to microscopic details indicates that the case of the attachment index $\gamma = 1$ is marginal. A related phenomenon occurs in constant-kernel aggregation, where the asymptotic kinetics is sensitively dependent on the actual values of the reaction rate [30,31].

B. Growing network with redirection (GNR)

The GN is built by simultaneous node and link addition, and disregards other elemental processes which can occur in the development of large networks. In the context of the web, these include node and link deletion (for out-of-date websites), link rewiring, the tendency of a new node to connect to nearby nodes, and the copying of links from existing nodes to new nodes. The GNR model incorporates a simple form of link rewiring into the GN model. At each time step a new node **n** is added, and an earlier node **x** is selected *uniformly* as a possible ''target'' for attachment. With a probability $1-r$, the link from **n** to **x** is created; in this case, the

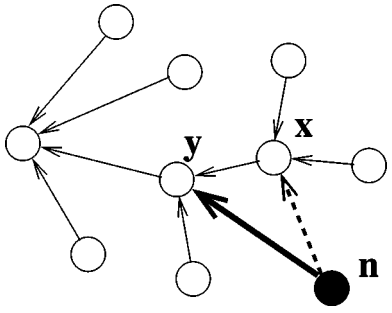

FIG. 2. Illustration of the basic processes in the GNR model. The new node (solid) selects a target node **x**. With probability $1-r$ a link is established to this target node (dashed arrow), while with probability r the link is established with the ancestor of **x** (thick solid arrow).

evolution is the same as in the GN. However, with probability r, the link is *redirected* to the ancestor node **y** of node **x** (Fig. 2).

A model of this spirit was recently mentioned in the context of web development [15]. A related model was also proposed long ago by Simon [32,33] to describe the word frequencies of English text. The Simon model gives a power-law frequency distribution whose exponent is tunable in manner which closely mirrors the behavior in the GNR model. The Simon model was also recently applied to explain power-law distributions in the frequency of family names [34].

While at first sight the GNR model appears complicated, we shall see that its characteristics can be obtained in a simple fashion. Another very helpful and surprising property of the GNR model with a uniform initial attachment probability is that it is equivalent to a GN with a shifted linear attachment kernel and *no* redirection. We shall exploit this equivalence extensively in the following. Nevertheless, we consider the GNR model separately, as in many cases the rate equations for the GNR model with a *uniform* attachment kernel is simpler to appreciate than the rate equations for a GN with a shifted linear attachment kernel.

III. DEGREE DISTRIBUTION

A. GN model

We now study the evolution of the degree distribution of the GN model. The rate equations for $N_k(t)$ are

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

The first term on the right-hand side of Eq. (1) accounts for the process in which a node with $k-1$ links is connected to the new node, leading to a gain in the number of nodes with k links. This occurs with a probability A_{k-1}/A, where $A(t)=\Sigma_{j\geqslant 1}A_jN_j(t)$ is the appropriate normalization factor. A corresponding role is played by the second (loss) term on the right-hand side of Eq. (1). Note that the overall amplitude in A_k is irrelevant, since it appears in both the numerator and denominator of Eq. (1), and can be chosen arbitrarily. The last term on the right-hand side of Eq. (1) accounts for the continuous introduction of new nodes with no incoming links. We also set $N_0\equiv 0$, so that Eq. (1) applies for all $k\geqslant 1$.

At a fundamental level, it is worth noting that Eq. (1) describes the symbolic reaction $[k]\to[k+1]$. Many other reactions, such as the Becker-Döring theory of nucleation [35], additive polymerization [36], hydrolysis [37], catalysis, and submonolayer epitaxial growth [38], fit into this scheme. However, there is one important difference in that we consider strictly a single connected cluster (the growing network), while in the context of aggregation-like processes, one generally deals with a collection of clusters. The effect of having more than one cluster in the framework of growing networks is currently under investigation [39].

We start by solving the equations for the low-order moments of the degree distribution, which are defined by $M_n(t)=\Sigma_{j\geqslant 1}j^nN_j(t)$. Summing Eqs. (1) over all k gives the rate equation for the total number of nodes, $\dot{M}_0=1$, whose solution is $M_0(t)=M_0(0)+t$. The first moment (the total number of link endpoints) obeys $\dot{M}_1=2$, which gives $M_1(t)=M_1(0)+2t$. The first two moments are therefore *independent* of the attachment kernel A_k, while higher moments and the degree distribution itself do depend on the kernel A_k.

To develop an appreciation for the types of behavior that can occur, consider the linear kernel $A_k=k$, for which $A(t)$ coincides with $M_1(t)$. In this case, we can solve Eq. (1) for an arbitrary initial condition. However, since the long-time behavior is most interesting, we limit ourselves to the asymptotic regime ($t\to\infty$) where the initial condition is irrelevant. Using therefore $M_1=2t$, we solve Eq. (1) and obtain $N_1=2t/3$, $N_2=t/6$, etc., which implies that each N_k grows linearly with time. Accordingly, we substitute $N_k(t)=tn_k$ in Eq. (1) to yield the simple recursion relation $n_k=n_{k-1}(k-1)/(k+2)$. Solving for n_k gives

$$n_k=\frac{4}{k(k+1)(k+2)}. \tag{2}$$

In the context of discrete functions defined on the positive integers, this distribution is algebraic over the entire range of k. Indeed, as explained in Ref. [40], the proper analog of the *continuous* power-law function $f(x)=x^{-\lambda}$ is the *discrete* function $f_k=\Gamma(k)/\Gamma(k+\lambda)$, where Γ is the Euler gamma function. Rewriting Eq. (2) as $n_k=4\Gamma(k)/\Gamma(k+3)$, we see that n_k is indeed algebraic over the entire range $k\geqslant 1$.

Returning to more general attachment kernels, let us assume that the degree distribution and $A(t)$ both grow linearly with time. We anticipate that this hypothesis will hold for attachment kernels which do not grow faster than linearly with k. Substituting $N_k(t)=tn_k$ and $A(t)=\mu t$ into Eq. (1), we obtain the recursion relation $n_k=n_{k-1}A_{k-1}/(\mu+A_k)$ and $n_1=\mu/(\mu+A_1)$. Solving for n_k, we obtain

$$n_k=\frac{\mu}{A_k}\prod_{j=1}^{k}\left(1+\frac{\mu}{A_j}\right)^{-1}. \tag{3}$$

To complete the solution, we need to find the amplitude μ. Combining the definition $\mu=\Sigma_{j\geqslant 1}A_j n_j$ and Eq. (3), we obtain the implicit relation

$$\sum_{k=1}^{\infty}\prod_{j=1}^{k}\left(1+\frac{\mu}{A_j}\right)^{-1}=1. \tag{4}$$

Thus the amplitude μ always depends on the entire attachment kernel. On the other hand, we shall show that the degree distribution exhibits a robust behavior which depends only on gross features of the attachment kernel, as long as A_k grows *more slowly* than linearly. The case where A_k is asymptotically linear is perhaps the most intriguing, as the degree distribution has a power-law behavior whose exponent depends on microscopic details of the dependence of A_k on k. When A_k grows *more quickly* than linearly, a drastically different gelationlike behavior arises. It is again worth noting that these three regimes of kinetic behavior also arise in the solutions to the rate equations for additive polymerization processes, with the different regimes arising when the attachment exponent γ is smaller than, larger than, or equal to 1 [41]. We now describe these three cases separately.

1. Sublinear kernels

Consider sublinear kernels which are *asymptotically homogeneous*, that is, $A_k\sim k^{\gamma}$, with $0<\gamma<1$. Substituting this asymptotics into Eq. (3), writing the product as the exponential of a sum, converting the sum to an integral, and performing this integral, we obtain

$$n_k\sim\begin{cases} k^{-\gamma}\exp\left[-\mu\left(\dfrac{k^{1-\gamma}-2^{1-\gamma}}{1-\gamma}\right)\right], & \dfrac{1}{2}<\gamma<1,\\[2ex] k^{(\mu^2-1/2)}\exp[-2\mu\sqrt{k}], & \gamma=\dfrac{1}{2},\\[2ex] k^{-\gamma}\exp\left[-\mu\dfrac{k^{1-\gamma}}{1-\gamma}+\dfrac{\mu^2}{2}\dfrac{k^{1-2\gamma}}{1-2\gamma}\right], & \dfrac{1}{3}<\gamma<\dfrac{1}{2}, \end{cases} \tag{5}$$

etc. The pattern given in Eq. (5) continues *ad infinitum*: Whenever γ decreases below $1/m$, with m a positive integer, an additional term in the exponential arises from the now relevant contribution of the next-higher-order term in the expansion of the product in Eq. (3).

To complete the solution, we require the amplitude μ. We have been unable to find an explicit expression for μ, even if the attachment kernel is strictly homogeneous, $A_k=k^{\gamma}$, as this requires solving Eq. (4). However, this relation can be easily evaluated numerically, and it shows that $\mu(\gamma)$ varies smoothly between 1 and 2 as γ increases from 0 to 1 (Fig. 3). These two limits correspond to the known limiting behaviors for M_0 and M_1.

More detailed results can be obtained for the limiting solvable cases of $A_k=\text{const}$ and $A_k=k$. In these limits, $\mu=1$ and 2, respectively, and the corresponding degree distributions are given by $n_k=2^{-k}$ and by Eq. (2). The former can be easily obtained by exactly following the same steps as those used to solve the network with the linear kernel. We can then apply perturbation theory to find the respective limiting behaviors of $\mu(\gamma)$ for γ close to 0 or 1,

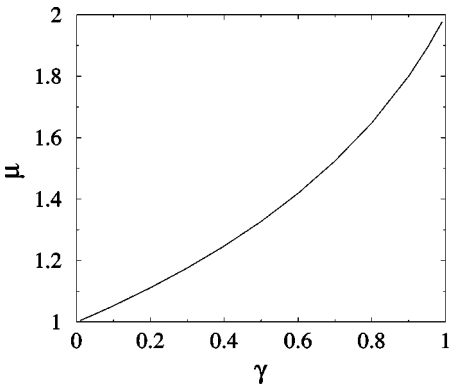

FIG. 3. The amplitude, μ in $M_{\gamma}(t)=\mu t$, vs γ.

$$\mu=1+B_0\gamma+\mathcal{O}(\gamma^2),$$

$$\mu=2-B_1(1-\gamma)+\mathcal{O}((1-\gamma)^2),$$

with

$$B_0=\sum_{j=1}^{\infty}\frac{\ln j}{2^j}=0.5078\ldots,$$

$$B_1=4\sum_{j=1}^{\infty}\frac{\ln j}{(j+1)(j+2)}=2.407\ldots.$$

2. Linear kernels

Now consider *asymptotically* linear attachment kernels, $A_k\sim k$ as $k\to\infty$. As already mentioned, we can always choose the amplitude in the asymptotic relation to be equal to 1, as attachment kernels which differ by a multiplicative factor give identical behaviors. For an asymptotically linear kernel, expanding the product in Eq. (3) and following step by step the approach that led to Eq. (5) now gives the power-law asymptotic behavior

$$n_k\sim k^{-\nu}\quad\text{with}\quad\nu=1+\mu. \tag{6}$$

An important feature of this result is that the exponent ν can be tuned to *any* value larger than 2. This lower bound immediately follows from the fact that the sum $\mu=\Sigma_j A_j n_j\sim\Sigma_j j n_j$ must converge, and this, in turn, requires that ν must be larger than 2.

As an explicit example, consider the attachment kernel $A_k=k$ for $k\geqslant 2$, while $A_1\equiv\alpha$ is an arbitrary positive number. Now it is convenient to treat separately A_1 and A_k for $k\geqslant 2$ in Eq. (4) to recast it as

$$\mu=A_1\sum_{k=2}^{\infty}\prod_{j=2}^{k}\left(1+\frac{\mu}{A_j}\right)^{-1}. \tag{7}$$

The right-hand side of Eq. (7) can be simply expressed as the ratio of Euler gamma functions, to yield

$$\mu=\alpha\sum_{k=2}^{\infty}\Gamma(2+\mu)\frac{\Gamma(1+k)}{\Gamma(1+\mu+k)}. \tag{8}$$

This sum can be evaluated by employing the identity [40]

$$\sum_{k=2}^{\infty}\frac{\Gamma(a+k)}{\Gamma(b+k)}=\frac{\Gamma(a+2)}{(b-a-1)\Gamma(b+1)}, \tag{9}$$

so that Eq. (8) reduces to $\mu(\mu-1)=2\alpha$, with solution $\mu=(1+\sqrt{1+8\alpha})/2$. Thus the exponent $\nu=1+\mu$ is

$$\nu=\frac{3+\sqrt{1+8\alpha}}{2}. \tag{10}$$

Furthermore, following the steps that lead to Eq. (3), the degree distribution for the GN with the attachment kernel $A_1=\alpha$ and $A_k=k$ for $k\geqslant 2$ is

$$n_1=\frac{\mu}{\mu+\alpha},\quad n_k=\frac{\mu\alpha}{\mu+\alpha}\,\frac{\Gamma(2+\mu)\Gamma(k)}{\Gamma(1+\mu+k)}. \tag{11}$$

Note that for $0<\alpha<1$, the exponent lies in the range $2<\nu<3$; in particular, $\nu=2+2\alpha-4\alpha^2+\cdots$ as $\alpha\to 0$. When $\alpha=1$, we recover the connectively distribution of Eq. (2). For $\alpha>1$, we have $\nu>3$; in particular, $\nu\to\sqrt{2\alpha}$ as $\alpha\to\infty$.

The GN is also solvable when $A_k=k+w$. This shifted linear kernel can be motivated naturally by explicitly keeping track of the directionality of the links. In particular, the node degree for an undirected graph generalizes to the in-degree and out-degree for a directed graph. These are just the number of incoming and outgoing links at a node, respectively. Thus the node degree k in a directed graph is the sum of the in-degree i and out-degree j. The most general linear attachment kernel for a directed graph is therefore of the form $A_{ij}=ai+bj$. The GN corresponds to the case where the out-degree of any node is equal to 1; thus $j=1$ and $k=i+1$. Hence the general linear attachment kernel reduces to $A_k=a(k-1)+b$. Since, as mentioned above, the overall scale factor in the kernel is irrelevant, we can rewrite A_k as the shifted linear kernel $A_k=k+w$, with $w=-1+b/a$, so that it can vary over the range $-1<w<\infty$.

We can now easily determine the degree distribution for the shifted linear attachment kernel. First we note that $A(t)=\Sigma_j A_j N_j=M_1(t)+wM_0(t)$. Then using the basic results $A=\mu t$, $M_0=t$ and $M_1=2t$, we have $\mu=2+w$ and thus $\nu=3+w$, according to Eq. (6). Furthermore, from Eq. (3) we easily determine the entire degree distribution to be

$$n_k=(2+w)\,\frac{\Gamma(3+2w)}{\Gamma(1+w)}\,\frac{\Gamma(k+w)}{\Gamma(k+3+2w)}. \tag{12}$$

In a similar vein, we can solve the GN with an arbitrary piecewise linear attachment kernel. In all these cases, the exponent ν can be tuned to any value larger than 2, and for sufficiently large degree n_k can be expressed as the ratio of gamma functions, i.e., the degree distribution is a purely (discrete) algebraic function.

3. Superlinear kernels

For superlinear homogeneous attachment kernels $A_k=k^\gamma$, with $\gamma>1$, we now show that a ''winner take all'' phenomenon arises, namely, there emerges a single dominant ''gel'' node which is linked to almost every other node. A particularly singular behavior occurs for $\gamma>2$, where there is a nonzero probability that the initial node is connected to *every* other node of the network.

Let us first determine the probability that the initial node connects to all other nodes. It is convenient to consider a discrete time version of the GN in which one node is introduced at each elemental step, which always links to the initial node. After N steps, the probability that the new node will link to the initial node is $N^\gamma/(N+N^\gamma)$. The probability that this connectivity pattern continues indefinitely is

$$\mathcal{P}=\prod_{N=1}^{\infty}\frac{1}{1+N^{1-\gamma}}. \tag{13}$$

Clearly, $\mathcal{P}=0$ when $\gamma\leqslant 2$ but $\mathcal{P}>0$ when $\gamma>2$. Thus, for $\gamma>2$ there is a nonzero probability that the initial node connects to all other nodes.

To determine the behavior for general $\gamma>1$, we first need the asymptotic time dependence of M_γ. To this end, it is useful to consider the discretized version of the master equations Eq. (1), where the time t is limited to integer values. Then $N_k(t)=0$ whenever $k>t$, and the rate equation for $N_k(k)$ immediately leads to

$$N_k(k)=\frac{(k-1)^\gamma N_{k-1}(k-1)}{M_\gamma(k-1)}=N_2(2)\prod_{j=2}^{k-1}\frac{j^\gamma}{M_\gamma(j)}. \tag{14}$$

From this, and the obvious fact that $N_k(k)$ must be less than unity, it follows that $M_\gamma(t)$ cannot grow more slowly than t^γ. On the other hand, $M_\gamma(t)$ cannot grow faster than t^γ, as follows from the estimate

$$M_\gamma(t)=\sum_{k=1}^{t}k^\gamma N_k(t)\leqslant t^{\gamma-1}\sum_{k=1}^{t}kN_k(t)=t^{\gamma-1}M_1(t). \tag{15}$$

Thus $M_\gamma\propto t^\gamma$. In fact, the amplitude of t^γ is unity as we will derive self-consistently after solving for the N_k's.

We now use $M_\gamma\sim t^\gamma$, with $\gamma>1$, in the rate equations to solve recursively for each N_k. Starting with the equation $\dot N_1=1-N_1/M_\gamma$, we see that the second term on the right-hand side is subdominant. Thus, by neglecting this term, we obtain $N_1=t$. Similarly, $\dot N_2=(N_1-2^\gamma N_2)/M_\gamma\sim N_1/M_\gamma$ gives $N_2\sim t^{2-\gamma}/(2-\gamma)$. Continuing this same line of reasoning for each successive rate equation gives the leading behavior of N_k,

$$N_k(t)=J_k t^{k-(k-1)\gamma}\quad\text{for}\quad k\geqslant 1, \tag{16}$$

with $J_k=\Pi_{j=1}^{k-1}j^\gamma/[1+j(1-\gamma)]$. This pattern of behavior for N_k continues as long as its exponent $k-(k-1)\gamma$ remains positive, or $k<\gamma/(\gamma-1)$. The full temporal behavior of

$N_k(t)$ may be determined straightforwardly by keeping the next correction terms in the rate equations. For example, $N_1(t)=t-t^{2-\gamma}/(2-\gamma)+\cdots$.

For $k>\gamma/(\gamma-1)$, each N_k has a finite limiting value in the long-time limit. Since the total number of connections is equal to $2t$, and t of them are associated with N_1, the remaining t links must all connect to a single node which has t connections (up to corrections which grow no faster than sublinearly with time). Consequently the amplitude of M_γ is equal to unity, as argued above.

Therefore for superlinear kernels, the GN undergoes an infinite sequence of connectivity transitions as a function of γ. For $\gamma>2$ all but a finite number of nodes are linked to the ''gel'' node, which has the rest of the links of the network. This is the ''winner take all'' situation. For $3/2<\gamma<2$, the number of nodes with two links grows as $t^{2-\gamma}$, while the number of nodes with more than two links is again finite. For $4/3<\gamma<3/2$, the number of nodes with three links grows as $t^{3-2\gamma}$ and the number with more than three is finite. Generally for $(m+1)/m<\gamma<m/(m-1)$, the number of nodes with more than m links is finite, while $N_k\sim t^{k-(k-1)\gamma}$ for $k\leqslant m$. Logarithmic corrections also arise at the transition points.

B. Relation to citation data

Let us now attempt to relate some of our predictions from the GN model to the distribution of citations in recent scientific publications [11,12]. The GN model represents an extreme idealization of the citation process in which each publication cites only a single paper and the probability of citing a paper depends only on its current number of citations, and not on its intrinsic quality or any other realistic features. Thus we anticipate that the connection between the model and the data will be, at best, tenuous.

The data that we discuss is based on (a) 783 339 papers with 6 716 198 citations [provided by the Institute of Scientific Information (ISI)], and (b) 24 296 papers with 351 872 citations from all issues of Physical Review D (PRD) from 1975 to 1994 (provided by the SPIRES database) [42]. A cursory visual inspection of this data suggests that the number of publications with k citations decays as a stretched exponential function of k (see, e.g., Fig. 1 of Ref. [12]). However, an analysis based on presenting the data in a Zipf plot, in conjunction with scaling, is suggestive of a power-law form for the citation distribution, $k^{-\nu}$, with $\nu\approx 3$ (Fig. 2 of Ref. [12]). This ambiguity between a stretched exponential and power-law form for the citation distribution corresponds to the situation where the predictions of the GN itself are difficult to discern numerically.

If we consider a GN with an attachment kernel $A_k\sim k^\gamma$ for $\gamma\lesssim 1$, then a plot of n_k in Eq. (5) versus k, for $1\leqslant k\leqslant 1000$, changes relatively slowly as γ varies in the range $(0.9,1)$. If one attempts to fit this data to a power law, then an exponent value somewhat larger than 3 gives a reasonable fit to the data. It is only as $\gamma\to 1$ from below, however, that the factors in the exponential of Eq. (5) conspire to give a pure power-law form for n_k. Because of the relatively small change in n_k as γ varies, the relatively incomplete data on the distribution of citations are insufficient to provide a clear test for the existence of a power law. Further, for a GN model with a linear attachment kernel, the degree distribution depends on additional details of this kernel, and can achieve *any* value greater than 2. In short, it is difficult to relate the GN model to citation data based on the form of the distribution alone.

Another interesting aspect of the citation distribution which can be compared with the GN model is the nature of highly cited publications. Within the GN model, the degree of the most popular node, $k_{\max}$, may be determined by the extreme statistics criterion $\Sigma_{k>k_{\max}} N_k=1$, which states that there is one node in the network whose degree lies in the range $(k_{\max},\infty)$. This criterion gives

$$k_{\max}\sim\begin{cases}(\ln t)^{1/(1-\gamma)}, & 0\leqslant\gamma<1\\ t^{1/(\nu-1)} & \text{asymptotically linear}\\ t, & \gamma>1.\end{cases} \tag{17}$$

We now compare this prediction with the data about the most-cited paper. To make a correspondence between citations and Eq. (17), we identify the total number of publications in each dataset with t. The most cited paper had 8904 citations in the ISI data set and 2026 citations in the PRD data set. These results are consistent with the first line of Eq. (17) when $\gamma\approx 0.86$ and 0.7 respectively, and also with the second line for $\nu\approx 2.5$ and 2.3, respectively. Thus an analysis of the most-cited paper does not cleanly indicate whether the citation distribution is a power law or a stretched exponential. These ambiguities indicate that some of the issues that should be clarified to provide a clear description of citations in terms of a growing network model.

C. GNR model

We now solve the GNR model within the rate equation framework. According to the basic processes in the model (Fig. 2), the degree distribution $N_k(t)$ evolves by the rate equations

$$\frac{dN_k}{dt}=\delta_{k1}+\frac{1-r}{M_0}[N_{k-1}-N_k]+\frac{r}{M_0}[(k-2)N_{k-1}-(k-1)N_k]. \tag{18}$$

For the redirection probability $r=0$, the first three terms on the right-hand side of Eq. (18) are the same as in the GN. The last two terms account for the change in N_k due to the redirection process. To understand their origin, consider the gain term due to redirection. Since the initial node is chosen uniformly, if redirection does occur, the probability that a node with $k-1$ pre-existing links receives the new ''redirected'' link is proportional to $k-2$, the number of preexisting incoming links. A similar argument applies for the redirection-driven loss term. Since $N_0\equiv 0$ is tacitly assumed, Eq. (18) applies for all $k\geqslant 1$.

By combining the terms in Eq. (18), the rate equation reduces to that of the original GN with $A_k=(k-1)r+1-r=r[k-1+(1-r)/r]$. By scaling out the factor r, we then

reduce A_k to the form of the shifted linear kernel $k+w$, with $w=[(1-r)/r]-1=(1/r)-2$. Thus we can merely transcribe our results about the GN with the shifted linear kernel to determine the degree distribution for the GNR model. Amusingly, for $r=1/2$, the GNR model is identical to the GN with the purely linear kernel. In general, the degree distribution in the R model is a power law with exponent $\nu=1+1/r$, which can be tuned to any value larger than 2. This exponent value was first obtained in Simon's original paper [32], but in a rather different context, by employing an approach which is similar to ours.

IV. AGE DISTRIBUTION

In addition to the distribution of degree, we study *when* connections occur in the GN. This provides a deeper understanding of the overall development of growing networks. Naively, we expect that older nodes will be better connected, and this can be quantified by categorizing nodes both by their degree and age. It should be emphasized that the GN does *not* have explicit aging, in which the connection probability depends on the age of the target node; this feature is treated in Ref. [26]. Instead, we merely extend the categorization of the nodes to include their age as well as their degree.

A. Linear connection kernel

Let $c_k(t,a)$ be the average number of nodes of age a which have $k-1$ incoming links at time t. Here age a means that the node was introduced at time $t-a$. That is, we are now resolving each node both by its degree and its age. The resulting joint age-degree distribution is simply related to the degree distribution through $N_k(t)=\int_0^t da c_k(t,a)$. The joint distribution evolves according to

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)c_k=\frac{A_{k-1}c_{k-1}-A_k c_k}{A(t)}+\delta_{k1}\delta(a). \qquad (19)$$

The second term on the left accounts for the aging of nodes, and the probability of connecting to a given node again depends only on its degree and not on its age.

We start by considering the linear attachment kernel $A_k=k$, and focus on the long-time asymptotic behavior. Then we can disregard the initial condition and write $A(t)\equiv M_1(t)=2t$. This transforms Eqs. (19) into

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)c_k=\frac{(k-1)c_{k-1}-kc_k}{2t}+\delta_{k1}\delta(a). \qquad (20)$$

The homogeneous form of this equation implies that a solution should be self-similar. Thus we seek a solution as a function of the *single* variable a/t rather than two separate variables. Thus we write

$$c_k(t,a)=f_k(x) \quad \text{with} \quad x=1-\frac{a}{t}. \qquad (21)$$

This turns the partial differential equation (20) into the ordinary differential equation

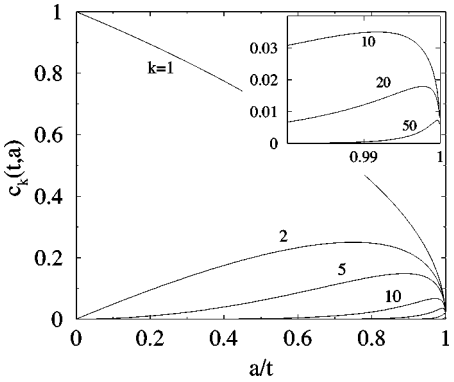

FIG. 4. Age-dependent degree distribution for the GN for the linear attachment kernel. Low-degree nodes tend to be relatively young, while high-degree nodes are old. The inset shows detail for $a/t\geqslant 0.98$.

$$-2x\frac{df_k}{dx}=(k-1)f_{k-1}-kf_k. \qquad (22)$$

We have omitted the delta function term, since it merely provides the boundary condition $c_k(t,a=0)=\delta_{k1}$, or

$$f_k(1)=\delta_{k1}. \qquad (23)$$

The solution to this boundary-value problem may be simplified by assuming an exponential solution $f_k=\Phi\varphi^{k-1}$; this is consistent with the boundary condition, provided that $\Phi(1)=1$ and $\varphi(1)=0$. The above ansatz reduces the infinite set of rate equations [Eq. (22)] into two elementary differential equations for $\varphi(x)$ and $\Phi(x)$, whose solutions are $\varphi(x)=1-\sqrt{x}$ and $\Phi(x)=\sqrt{x}$. In terms of the original variables of a and t, the joint age-degree distribution is then

$$c_k(t,a)=\sqrt{1-\frac{a}{t}}\left\{1-\sqrt{1-\frac{a}{t}}\right\}^{k-1}. \qquad (24)$$

Thus the degree distribution for nodes of fixed age decays exponentially with degree, with a characteristic degree which diverges as $\langle k\rangle\sim(1-a/t)^{-1/2}$ for $a\to t$. As expected, young nodes (those with $a/t\to 0$) typically have a small degree, while old nodes have a large degree (Fig. 4). It is the slow decay of the degree distribution for old nodes which ultimately leads to a power-law degree distribution when this joint age-degree distribution is integrated over all ages to give $N_k(t)$.

B. General connection kernels

Let us now consider a GN with a connection kernel which grows either linearly or more slowly with k. Ansatz (21) still is valid, so that the distribution f_k evolves according to

$$-\mu x\frac{df_k}{dx}=A_{k-1}f_{k-1}-A_k f_k. \qquad (25)$$

We now solve Eq. (25), subject to the boundary condition (23), and with μ determined from Eq. (4). Let us first replace x by $X=-\mu^{-1}\ln x$, which reduces the left-hand side of Eq.

(25) to df_k/dX. Applying a Laplace transform, $\hat{f}_k(s) = \int_0^\infty dX e^{-sX} f_k(X)$, $\hat{f}_k(s)$ obeys a simple algebraic recursion formula whose solution is

$$\hat{f}_k(s) = \frac{1}{A_k} \prod_{j=1}^{k} \left(1 + \frac{s}{A_j}\right)^{-1}. \tag{26}$$

Apart from the notation, this is identical to Eq. (3), and can be analyzed accordingly. In particular, we can determine $\hat{f}_k(s)$ for various asymptotically linear attachment kernels. For example, for the shifted linear attachment kernel $A_k = k + w$, we find

$$\hat{f}_k(s) = \frac{\Gamma(1+w+s)}{\Gamma(1+w)} \frac{\Gamma(k+w)}{\Gamma(k+1+w+s)}. \tag{27}$$

To invert this Laplace transform, it is useful to rewrite this expression as a sum of rational functions $\hat{f}_k(s) = \Sigma_{1 \leqslant j \leqslant k} F_j^k (j+w+s)^{-1}$. This then gives $f_k(X) = \Sigma_{1 \leqslant j \leqslant k} F_j^k e^{-(j+w)X}$, with

$$F_j^k = \frac{(-1)^{j-1}\Gamma(k+w)}{\Gamma(j)\Gamma(k-j+1)\Gamma(1+w)}. \tag{28}$$

We then re-express this in terms of the original variable $x = e^{-(2+w)X}$. Hence $f_k(x)$ can be rewritten as the sum of k power laws, $f_k(x) = \Sigma_{1 \leqslant j \leqslant k} F_j^k x^{(j+w)/(2+w)}$. Substituting the explicit expression (28) into this sum reduces the joint age-degree distribution to

$$f_k(x) = \frac{\Gamma(k+w)}{\Gamma(k)\Gamma(1+w)} x^{(1+w)/(2+w)} [1 - x^{1/(2+w)}]^{k-1}. \tag{29}$$

This expression shows that old nodes have a broad distribution of degrees up to a characteristic degree $\langle k \rangle = (1 - a/t)^{-1/(2+w)}$. One can also verify that the average age a_k of nodes of degree k, defined as $a_k = N_k^{-1} \int_0^t a c_k(t,a) da = t n_k^{-1} \int_0^1 (1-x) f_k(x) dx$, is

$$\frac{a_k}{t} = 1 - \frac{\Gamma(5+3w)\Gamma(k+3+2w)}{\Gamma(3+2w)\Gamma(k+5+3w)} \sim 1 - \frac{\text{const.}}{k^{2+w}}. \tag{30}$$

Thus nodes with a very large degree necessarily have an age which approaches that of the entire network.

Finally, the joint age-degree distribution simplifies in the limit $k \to \infty$ and $x \to 0$, with the scaling variable $\xi = k x^{1/(2+w)}$ kept finite. In this case, we can rewrite Eq. (29) in the scaling form

$$f_k(x) = k^{-1} F(\xi), \quad F(\xi) = \frac{\xi^{1+w}}{\Gamma(1+w)} \exp(-\xi). \tag{31}$$

The scaling variable can also be written as $\xi = k/\langle k \rangle$, and thus Eq. (31) clearly shows that old nodes have a broad distribution of degrees: $1 \leqslant k \lesssim \langle k \rangle$.

We can derive explicit age-degree distributions for other attachment kernels. For example, for the constant attachment kernel, $A_k = 1$, the joint age-degree distribution is the Poisson distribution,

$$f_k(X) = \frac{X^{k-1}}{(k-1)!} e^{-X}, \tag{32}$$

or, in terms of the original variables a and t,

$$c_k(t,a) = \left(1 - \frac{a}{t}\right) \frac{|\ln(1-a/t)|^{k-1}}{(k-1)!}. \tag{33}$$

The characteristic degree now diverges relatively slowly, *viz.* $\langle k \rangle \sim -\ln(1-a/t)$ as $a \to t$, than for asymptotically linear attachment kernels. On the other hand, the average age approaches the maximal age t at a much faster rate, as $a_k = t[1 - (2/3)^k]$, as k approaches its maximal value.

For cases where we have been unable to obtain an explicit solution, the Laplace transform method still allows us to extract the asymptotics. For example, for asymptotically homogeneous attachment kernels, $A_k \to k^\gamma$ as $k \to \infty$, Eq. (26) gives the large-k asymptotics $\hat{f}_k(s) \sim k^{-\gamma} \exp[-s k^{1-\gamma}/(1-\gamma)]$ [see Eq. (5)]. (For concreteness, here we consider the range $1/2 < \gamma < 1$.) Inverting this Laplace transform yields

$$f_k(X) \sim k^{-\gamma} \delta\left(X - \frac{k^{1-\gamma}}{1-\gamma}\right). \tag{34}$$

In particular, the age of nodes with k links is peaked about the value a_k which satisfies

$$\frac{a_k}{t} \simeq 1 - \exp\left(-\mu \frac{k^{1-\gamma}}{1-\gamma}\right). \tag{35}$$

This again shows that old nodes are much better connected.

V. NODE DEGREE CORRELATIONS

We now demonstrate that correlations between the degrees of connected nodes spontaneously develop as the network grows. One motivation for focusing on these correlations is that recently random graph models with arbitrary degree distributions have been investigated [43–45]. While the degree distribution can be chosen arbitrarily in these models, the degrees of connected nodes are *uncorrelated*. This lack of correlation suggests that such random graphs may have limited applicability to growing network systems.

For the GN, a useful characterization of node degree correlations is $N_{kl}(t)$, the number of nodes of total degree k which attach to an ancestor node of total degree l. For example, in the network of Fig. 1, there are $N_1 = 6$ nodes of degree 1, with $N_{12} = N_{13} = N_{15} = 2$. There are also $N_2 = 2$ nodes of degree 2, with $N_{25} = 2$, and $N_3 = 1$ node of degree 3, with $N_{35} = 1$. The correlation function is not defined for the initial node. Generally, N_{kl} is defined for $k \geqslant 1$ and $l \geqslant 2$, and obeys the sum rule $N_k = \Sigma_l N_{kl}$. A gratifying feature of the rate equation approach is that the correlation function N_{kl} can be understood in a natural and simple fashion.

A. Linear connection kernel

For the GN with the linear attachment kernel $A_k=k$, the joint distribution $N_{kl}(t)$ evolves according to

$$M_1\frac{dN_{kl}}{dt}=[(k-1)N_{k-1,l}-kN_{kl}]+[(l-1)N_{k,l-1}-lN_{kl}] +(l-1)N_{l-1}\delta_{k1}. \tag{36}$$

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

Asymptotically, $M_1\to 2t$ and $N_{kl}\to tn_{kl}$, and we use these hypotheses to reduce Eqs. (36) to the time-independent recursion relations

$$(k+l+2)n_{kl}=(k-1)n_{k-1,l}+(l-1)n_{k,l-1} +(l-1)n_{l-1}\delta_{k1}. \tag{37}$$

This can be reduced to a constant-coefficient inhomogeneous recursion relation by the substitution

$$n_{kl}=\frac{\Gamma(k)\Gamma(l)}{\Gamma(k+l+3)}m_{kl} \tag{38}$$

to yield

$$m_{kl}=m_{k-1,l}+m_{k,l-1}+4(l+2)\delta_{k1}. \tag{39}$$

By solving Eq. (39) for the first few k, one can grasp the pattern of dependence on k and l and thereby infer the general solution

$$m_{kl}=4\frac{\Gamma(k+l)}{\Gamma(k+2)\Gamma(l-1)}+12\frac{\Gamma(k+l-1)}{\Gamma(k+1)\Gamma(l-1)}. \tag{40}$$

This solution can also be obtained in a more systematic manner by the generating function method (see below for the shifted linear kernel). Combining Eqs. (38) and (40), we finally obtain

$$n_{kl}=\frac{4(l-1)}{k(k+1)(k+l)(k+l+1)(k+l+2)} +\frac{12(l-1)}{k(k+l-1)(k+l)(k+l+1)(k+l+2)}. \tag{41}$$

The important feature of this result is that the joint distribution does not factorize, that is, $n_{kl}\neq n_k n_l$. This confirms our earlier assertion that correlations between the degrees of connected nodes form spontaneously. This is arguably the most important distinction between classical random graphs — where node degrees are uncorrelated—and the GN.

While the solution of Eq. (41) is unwieldy, it greatly simplifies in the scaling regime, $k\to\infty$ and $l\to\infty$, with $y=l/k$ kept finite. The scaled form of the solution is

$$n_{kl}=k^{-4}\frac{4y(y+4)}{(1+y)^4}. \tag{42}$$

For fixed large k, the distribution n_{kl} has a single maximum at $y_*=(\sqrt{33}-5)/2\cong 0.372$. Thus a node whose degree k is large is typically linked to another node whose degree is also large; the typical degree of the ancestor is 37% of the degree of the daughter node. In the complementary case of a fixed degree l for the ancestor node, the distribution n_{kl} reaches a maximum when $k=1$, i.e., the daughter node is usually dangling. From Eq. (41), we find that this configuration occurs with a probability

$$n_{1l}=\frac{2(l-1)(l+6)}{l(l+1)(l+2)(l+3)}. \tag{43}$$

Finally, when both k and l are large and their ratio is also very different from 1, the limiting behaviors of n_{kl} are

$$n_{kl}\to\begin{cases}16(l/k^5) & \text{when } l\ll k\\ 4/(k^2l^2) & \text{when } l\gg k.\end{cases} \tag{44}$$

This last result demonstrates the correlations in the network most cleanly. If there were no correlations, then $n_k n_l$ would be proportional to $(kl)^{-3}$.

B. General connection kernels

In general, correlations between the degrees of neighboring connected nodes exist for any attachment kernel. The analysis of these correlations for an arbitrary kernel is tedious, and we merely outline some of the primary results in the relatively simple cases of the shifted linear and constant attachment kernels.

In the former case, we follow the same approach as the linear kernel, to reduce the rate equation for the correlation function to recursion relations of a similar form to Eq. (37), *viz.*

$$(k+l+2+3w)n_{kl}=(k+w-1)n_{k-1,l}+(l+w-1) \times[n_{k,l-1}+n_{l-1}\delta_{k1}]. \tag{45}$$

Here n_l is determined from Eq. (12). In analogy with Eq. (38), the substitution

$$n_{kl}=\frac{\Gamma(k+w)\Gamma(l+w)}{\Gamma(k+l+3+3w)}m_{kl} \tag{46}$$

reduces Eqs. (45) to

$$m_{kl}=m_{k-1,l}+m_{k,l-1}+\delta_{k1}W\frac{\Gamma(l+3+3w)}{\Gamma(l+2+2w)}, \tag{47}$$

where $W=(2+w)\Gamma(3+2w)/[\Gamma(1+w)]^2$. We solve recursion (47) by the generating function method [40]. Multiplying Eq. (47) by $x^k y^l$, and summing over all $k\geq 1, l\geq 2$, we find that the generating function

$$\mathcal{M}(x,y)=\sum_{k=1}^{\infty}\sum_{l=2}^{\infty} m_{kl}x^k y^l \tag{48}$$

is given by

$$\mathcal{M}(x,y)=\frac{Wxy^2}{1-x-y}\sum_{j=0}^{\infty}\frac{\Gamma(j+5+3w)}{\Gamma(j+4+2w)}y^j. \tag{49}$$

Expanding $\mathcal{M}(x,y)$, we obtain

$$m_{kl}=W\sum_{j=0}^{l-2}\frac{\Gamma(k+l-2-j)\Gamma(j+5+3w)}{\Gamma(k)\Gamma(l-1-j)\Gamma(j+4+2w)}. \tag{50}$$

Equations (46) and (50) constitute an exact solution for the correlation function of the GN with the shifted linear attachment kernel.

When the parameter w is an integer, we can reduce n_{kl} to a rational function. In the general case, the exact solution also simplifies in several extreme limits. When $k\gg l$, the dominant contribution to n_{kl} is provided by the first term in the sum in Eq. (50). Assuming, additionally, that $l\gg 1$ and repeatedly using the asymptotic relation $\Gamma(N+n)/\Gamma(N)\to N^n$ as $N\to\infty$, we ultimately find

$$n_{kl}\simeq W\frac{\Gamma(5+3w)}{\Gamma(4+2w)}l^{1+w}k^{-5-2w},\quad k\gg l\gg 1. \tag{51}$$

In the complementary case of $l\gg k\gg 1$, all the terms in the sum of Eq. (50) are important. However, we can simplify this sum by employing the above asymptotics for the ratio of gamma functions, and then replacing the sum by an easily computable integral. We find

$$n_{kl}\simeq W\Gamma(2+w)k^{-2}l^{-2-w}. \tag{52}$$

When the attachment kernel is uniform, correlations between the degrees of a node and its ancestor still develop. To see how this comes about quantitatively, we again follow the same steps as those which led to Eq. (37), and find that the joint distribution n_{kl} now satisfies the recursion relation

$$3n_{kl}=n_{k-1,l}+n_{k,l-1}+2^{-(l-1)}\delta_{k1}. \tag{53}$$

This recursion relation can again be solved by the generating function technique, to give

$$n_{kl}=\frac{1}{2^{l-1}}-\frac{1}{3^{l-1}}\sum_{i=0}^{k-1}\frac{\Gamma(l-1+i)}{\Gamma(l-1)\Gamma(i+1)}\frac{1}{3^i}. \tag{54}$$

To appreciate the qualitative behavior of the joint distribution n_{kl}, it is again useful to fix one variable and vary the other. For fixed l, Eq. (54) shows that n_{kl} has a maximum at $k=1$. The magnitude of this maximum is $n_{1l}=2^{-(l-1)}-3^{-(l-1)}$. To analyze the behavior when k is fixed, it is convenient to transform Eq. (54) into

$$n_{kl}=\frac{1}{3^{l-1}}\sum_{i=k}^{\infty}\frac{\Gamma(l-1+i)}{\Gamma(l-1)\Gamma(i+1)}\frac{1}{3^i}. \tag{55}$$

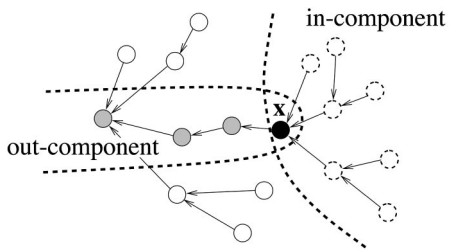

FIG. 5. In component and out component of node **x**.

Now a straightforward analysis shows that for large k, the maximum is attained at $l=k/2$.

The form of the joint distribution n_{kl} remains relatively complex even in the scaling regime $k,l\to\infty$, with the scaling variable $y=l/k$ kept finite. We determine the scaled form of the solution (55) by applying Stirling's formula and the identity $\Gamma(x+\lambda)/\Gamma(x)\to x^{\lambda}$ as $x\to\infty$. For $y<2$, we find

$$n_{kl}\simeq\frac{1}{\sqrt{2\pi k}}\frac{\sqrt{1+y^{-1}}}{2-y}e^{-kY}, \tag{56}$$

where $Y=y\ln y-(y+1)\ln[(y+1)/3]$. For $y>2$, it is preferable to use the solution in the form of Eq. (54). After some algebra, we can verify that the dominant contribution equals $2^{-(l-1)}$, that is, *independent* of k [46].

Finally, the limiting behavior of the correlation function is

$$n_{kl}\to 2^{-1}\times\begin{cases}3^{-(k+l-2)}[k^{l-2}/(l-2)!] & \text{when } l\ll k\\ 2^{-(l-2)} & \text{when } l\gg k.\end{cases} \tag{57}$$

Thus correlations are strong even for the random attachment kernel, and the qualitative behavior is similar to that of the linear attachment kernel.

VI. LARGE-SCALE PROPERTIES

The degree of a node is an important but local network characteristic, and we now seek to quantify more global features of the network. One such characteristic is the partitioning of the network into an *in component* and an *out component* with respect to any node (Fig. 5).

The in component to node **x** is the set of all nodes from which node **x** can be reached by following a path of directed links. Similarly, the out component of node **x** is the set of nodes which can be reached by following the path directed links which emanate from node **x**. For the GN model, the out component is just a single path, while in more realistic networks both the in and out components will be branched. In the context of citations, the in component is the set of all publications which refer to **x**, either directly or through intermediate reference lists until **x** is reached. The out component is the set of cited publications generated by iteratively following the reference list(s) of **x** and its ancestors.

A. In-component size distribution

The size distribution of the in component can be easily obtained by the rate equation formalism for the GN with a

uniform attachment kernel and also for the GNR. Given the equivalence between the latter and the GN with a shifted linear kernel, the latter case is also soluble. We start by considering the GN with the uniform attachment kernel. In this case, the number $I_s(t)$ of in components with s nodes satisfies the rate equation

$$\frac{dI_s}{dt} = \frac{(s-1)I_{s-1} - sI_s}{t} + \delta_{s1}. \tag{58}$$

To understand this equation, first consider the loss term. For an in component of size s there are s nodes in which the attachment of a new node causes this component to increase in size by 1. This gives a loss rate for I_s which is proportional to s. If there is more than one in component of size s they must be disjoint, so that the total loss rate for I_s is simply sI_s. A similar argument applies for the gain term. Finally, the overall factor of t^{-1} converts these rates to normalized probabilities. Curiously, Eq. (58) is almost identical to the rate equations for the degree distribution of a GN with a linear attachment kernel, except that the prefactor is equal to t^{-1} rather than to $(2t)^{-1}$.

From Eq. (58) we can determine all moments of the in-component size distribution, $\mathcal{I}_n(t) = \Sigma_{s\geq 1} s^n I_s(t)$. The zeroth moment obeys $\dot{\mathcal{I}}_0 = 1$, whose solution is $\mathcal{I}_0(t) = \mathcal{I}_0(0) + t$. This is obvious, since the total number of in components is equal to the total number of nodes. The first moment obeys $\dot{\mathcal{I}}_1 = 1 + \mathcal{I}_1/\mathcal{I}_0$, whose asymptotic solution is $\mathcal{I}_1(t) \sim t \ln t$. We shall see that this logarithmic factor is an outcome of the asymptotic power law for I_s with the tail decaying as s^{-2}.

To solve for $I_s(t)$, we note that it again grows linearly in time. Thus we substitute the ansatz $I_s(t) = t i_s$ into Eq. (58) to obtain $i_1 = 1/2$ and $i_s = i_{s-1}(s-1)/(s+1)$, which immediately leads to

$$i_s = \frac{1}{s(s+1)}. \tag{59}$$

For this s^{-2} decay, the moments $\mathcal{I}_n$ diverge when $n \geq 1$. However, the size of the largest in component, $s_{\max} = t$, provides an upper threshold in the computation of the moments. For example, $\mathcal{I}_1 \sim \Sigma_{s\leq t} s I_s(t) = t \ln t$. It is intriguing that the algebraic in-component distribution coexists with an exponential in-degree distribution, $n_k = 2^{-k}$.

Similarly, we can determine $I_s(t)$ for the GNR model. In this case, the number $I_s(t)$ of in-components with s nodes satisfies

$$\frac{dI_s}{dt} = \frac{[s-2+(1-r)]I_{s-1} - [s-1+(1-r)]I_s}{t} \tag{60}$$

for $s \geq 2$, and $\dot{I}_1 = 1 - (1-r)I_1/t$. This rate equation can be understood in a similar manner as Eq. (58). Consider the loss term for an in component of size s. There are two possibilities to consider: (i) If the apex of the in component is initially chosen, then the new node will attach to this apex with probability $1-r$ (i.e., attach with no redirection). (ii) If any other of the $s-1$ nodes of the in component is chosen, the new node will surely attach to the in component even if redirection occurs. These two processes give a loss rate for I_s which is proportional to $[s-1+(1-r)]I_s$. Solving for the in-component distribution in this process now yields $I_s(t) = t i_s$, with

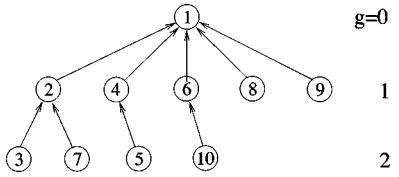

FIG. 6. Genealogy of the growing random network of Fig. 1. The indices indicate when a node is introduced, while the ancestor determines where a new node is positioned.

$$i_s = \frac{1-r}{(s-r)(s+1-r)}. \tag{61}$$

Remarkably, the asymptotic power law $I_s \propto s^{-2}$ holds for any r. It is striking that this apparently universal behavior has also recently been observed in measurements of the Internet [47].

Since the GNR model is identical to the GN with the shifted linear attachment kernel $A_k = k + [(1/r) - 2]$, Eq. (61) also applies to the in-component distribution for the GN with shifted linear attachment kernels. For example, the in-component distribution for the linear kernel is simply $i_s = 2/(4s^2 - 1)$. Since the same $I_s \propto s^{-2}$ decay holds for a GN with both constant and linear attachment kernels, we conjecture that the in-component distribution exhibits a universal s^{-2} decay for an *arbitrary* attachment kernel, as long as it does not grow faster than linearly with node degree.

B. Out-component size distribution

The out component from each node reveals basic insights into the ''genealogy'' of the growing network in an extremely simple fashion. For example, it allows us to estimate the diameter of the network, an important characteristic which has been measured for the web graph [48,49] and for social networks [23].

For this characterization, we begin by reorganizing the GN into a genealogical tree according to a procedure which is suggested by the growth process itself. Generation $g = 0$ contains a single ''seed'' node. The nodes which attach to the seed node form generation $g = 1$, and generally the nodes which attach to nodes in generation g form generation $g+1$, *independent* of when the attachment actually occurs. Thus the position of a node in the genealogical tree depends only on the position of the ancestor node and *not* on when the node is introduced. In this respect, the GN genealogical tree differs from typical genealogies, where each new generation is born into a progressively later position in the genealogical tree. For example, the network of Fig. 1 has five nodes in the first generation and four nodes in the second generation, leading to the genealogical tree of Fig. 6. The sizes of all generations grow continuously, except for generation $g = 0$ which always consists of a single node.

Once we understand the genealogical structure of the GN, we simultaneously establish the out-component distribution. Indeed, the number O_s of out components with s nodes is equal to L_{s-1}, the number of nodes in generation $s-1$ in the genealogical tree. We therefore compute $L_g(t)$, the size of generation g at time t. We start with the simplest situation when the attachment rate is uniform. In this case, $L_g(t)$ increases when a new node attaches to a node in the previous generation. This occurs with rate L_{g-1}/M_0, where $M_0(t)=1+t$ is the total number of nodes. Because of the simplicity of the corresponding rate equations, we use an exact expression for M_0 rather than the asymptotic expression $M_0 \sim t$, as was done in solving for the in component. Thus we write

$$\frac{dL_g}{dt}=\frac{L_{g-1}}{1+t}. \tag{62}$$

Solving these equations gives

$$L_g(\tau)=\frac{\tau^g}{g!} \quad \text{where} \quad \tau=\ln(1+t). \tag{63}$$

We therefore conclude that for a fixed (large) time, the generation size grows with g when $g<\tau$, reaches a maximum size which is equal to

$$L_{\max}\simeq\frac{t}{\sqrt{2\pi \ln t}} \tag{64}$$

when $g=\tau$, and then decreases and eventually becomes of order 1 when $g=e\tau$. The distribution L_g quickly decays when g exceeds the cutoff value $e\tau$. At time t, the genealogical tree therefore contains approximately $e\tau$ generations. Hence the diameter D of the network is approximately $2e\tau$, or

$$D\approx 2e\ln N, \tag{65}$$

where $N=1+t$ is the total number of nodes. Thus the diameter of an *evolving* GN exhibits the same N dependence as a *static* random graph [4].

We can also find the generation size distribution for shifted linear attachment kernels. It is again simpler to derive the rate equations in the framework of the GNR model, and then transcribe the results to the shifted linear kernel. For the GNR model, the rate equation for the generation size distribution is

$$\frac{dL_g}{dt}=\frac{(1-r)L_{g-1}+rL_g}{1+t} \tag{66}$$

for $g>1$, and $\dot{L}_1=(1+t)^{-1}[1+rL_1]$. The first term in Eq. (66) has the same origin as in a GN without redirection, and the second term accounts for the change in L_g due to the redirection. In the latter case, the new node provisionally attaches to a node in generation g; this occurs with a relative probability L_g. However, by the redirection process, this new node actually attaches to a node in generation $g-1$, and thereby joins generation g.

To solve Eq. (66), we again use $\tau=\ln(1+t)$, and apply the Laplace transform technique. After some elementary steps, we obtain

$$L_{g+1}(\tau)=\int_0^\tau dx\,\frac{[(1-r)x]^g}{g!}e^{xr}. \tag{67}$$

From this solution, we find that for a fixed (large) time, the generation size grows with g when $g<(1-r)\tau$, reaches a maximum value $L_{\max}\simeq t/\sqrt{2\pi(1-r)\ln t}$ at $g=(1-r)\tau$, and then decreases when $g>(1-r)\tau$. Eventually the generation size becomes of order 1 when $g=G\tau$, where G is the root of equation $G\ln[G/(1-r)]=G+r$. The diameter of the network is then $D\approx 2G\tau$.

These two solvable cases again suggest that the genealogy of the GN is robust, as long as the attachment kernel does not grow faster than linearly with node degree. For superlinear kernels, however, the genealogy changes drastically. When the attachment exponent exceeds 2, there will be only a few generations overall, and one generation g^* will contain all but a finite number of nodes. For such a network, the gel node will reside in generation g^*-1. When the attachment exponent lies in the range $1<\gamma<2$, a single generation will also contain almost all t nodes. However, the number of nodes which reside in other generations is of order $t^{2-\gamma}$ and thus grows as well. Additionally, the number of nonempty generations grows indefinitely with the total number of nodes.

The above results can be reformulated in terms of the out-component distribution. In particular, for a GN with a uniform attachment kernel, the number O_s of out components with s nodes is equal to

$$O_s(\tau)=\frac{\tau^{s-1}}{(s-1)!} \quad \text{where} \quad \tau=\ln(1+t). \tag{68}$$

Similar results apply for the linear attachment kernel, suggesting that the out-component distribution is robust as long as the attachment kernel does not grow faster than linearly with node degree.

VII. DISCUSSION AND CONCLUSIONS

In this paper, we have analyzed the structure of the growing network (GN) model and shown that many of its properties can be easily determined within a rate equation approach. We have found that the GN has a power-law node degree distribution $N_k(t)\sim tk^{-\nu}$ for asymptotically linear attachment kernels, with an exponent ν which is always larger than 2. By tuning parameters of the model in a reasonable way, it is easy to obtain a node degree distribution which is in quantitative agreement with available data for the web graph [14–16,19–21,49].

A remarkable feature of this network is the spontaneous development of correlations between connected nodes. These correlations provide a much more sensitive characterization

of the structure of growing networks than the extensively studied degree distribution. These correlations are also crucial features which distinguish the GN from classical random graphs. Thus testing for the presence of correlations between node degrees in large evolving networks may provide crucial insights to help determine the underlying mechanism of their growth.

We have also studied two specific large-scale properties of the network, namely, the size distributions of the in and out components with respect to a given site. The in-component distribution exhibits a robust s^{-2} power-law behavior, where s is the component size, as long as the attachment probability does not grow faster than linearly with node degree. The out-component distribution reveals the basic genealogical feature that the number of ''generations'' in the network grows logarithmically with the total number of nodes, again for attachment kernels which do not grow faster than linearly in node degree.

The qualitative agreement between the degree distributions of real evolving networks, such as the web graph, and the GN is reassuring given that the model ignores many important features of real networks. Nevertheless, a number of characteristics of real growing networks are difficult to treat in the framework of the GN model. One important such example is the out-degree distribution. Within the GN model the out-degree of each node is 1 by construction. In contrast, for real growing networks the out-degree distribution has a power-law form [49]. Additionally, the average in- and out-degrees at each node are generally larger than 1; for the web graph, for example, $\langle i\rangle=\langle j\rangle\approx 7.5$ [49].

There are several natural ways to extend the GN model to generate an average out-degree which is greater than 1. A simple construction is to link every new node to more than one earlier node, as already discussed in Ref. [20]. Let us consider a network which is built by attaching every new node to exactly p earlier nodes. For the linear attachment kernel, the degree distribution $N_k(t)$, which is now defined only for $k\geqslant p$, evolves according to

$$\frac{dN_k}{dt}=\frac{p}{M_1}[(k-1)N_{k-1}-kN_k]+\delta_{kp}. \tag{69}$$

Clearly, the average in-degree $\langle i\rangle$ and out-degree $\langle j\rangle$ of each node in this network are equal to p. By applying the basic approach of Sec. III to this rate equation, we find that the degree distribution again asymptotically approaches a stable distribution $N_k\to tn_k$, with

$$n_k=\frac{2p(p+1)}{k(k+1)(k+2)} \quad \text{for } k\geqslant p. \tag{70}$$

Thus for the linear attachment kernel, the average node degree does not affect the exponent ν of the degree distribution. However, for other solvable examples, the feature of attaching the new node to more than one preexisting node leads to different degree distributions. For example, for the shifted linear kernel we find

$$n_k=\text{const}\times\frac{\Gamma(k+w)}{\Gamma(k+3+w+w/p)} \quad \text{for } k\geqslant p, \tag{71}$$

$$n_p=\left(1+p\frac{p+w}{2p+w}\right)^{-1}. \tag{72}$$

This gives the asymptotic behavior $n_k\sim k^{-(3+w/p)}$. Thus the exponent of the degree distribution *depends* on the average node degree, with $\nu=3+w/p$.

The multiple linking construction also reduces the number of nodes with in-degree zero. For example, for a GN with a shifted linear attachment kernel, the fraction of such nodes is $n_1=(2+w)/(3+2w)$, which is always larger than 1/2. However, for the multiple linking construction, the fraction of nodes with in-degree zero is reduced to the value n_p given in Eq. (72). If we use $p=7$ to reproduce the correct average node degree of the web graph, the fraction of nodes with in-degree zero always exceeds 1/8, which, however, apparently disagrees with web data [49]. Thus, while multiple attachment does reduce the number of poorly connected nodes, this reduction is still insufficient to account for web-graph data. However, it is clear that the multiple linking construction has the potential to provide a better description of citation data.

Another shortcoming of multiple attachment is that it cannot dynamically generate a nontrivial out-degree distribution. However, we can extend the GN model by allowing for creation of links between existing nodes [50]. This simple construction allows us to generate nontrivial out-degree distributions which closely match web-graph data. An even more challenging direction is to describe the global topological structure of growing networks. The GN model leads to a single-component tree graph, while the web graph has numerous disconnected components. A deeper understanding of the web graph may provide valuable insights to help develop algorithms for web crawling, searching, and community discovery.

ACKNOWLEDGMENTS

We are grateful to NSF Grant No. DMR9978902 and ARO Grant No. DAAD19-99-1-0173 for partial financial support of this research.

[1] S. A. Kauffman, *The Origin of Order: Self-Organization and Selection in Evolution* (Oxford University Press, London, 1993).

[2] *Models of Neural Networks I,* 2nd ed., edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer Verlag, New York, 1995).

[3] See, e.g., G. Caldarelli, P. G. Higgs, and A. J. McKane, J. Theor. Biol. **193**, 345 (1998), and references therein; C. J. Camacho, R. Guimera, and L. A. N. Amaral, e-print cond-mat/0103114.

[4] B. Bollobás, *Random Graphs* (Academic Press, London, 1985).

[5] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs* (Wiley, New York, 2000).
[6] A. J. Lotka, J. Wash. Acad. Sci. **16**, 317 (1926).
[7] W. Shockley, Proc. IRE **45**, 279 (1957).
[8] E. Garfield, Science **178**, 471 (1972).
[9] L. Egghe and R. Rousseau, *Introduction to Informetrics* (Elsevier, Amsterdam, 1990).
[10] N. Gilbert, Sociol. Res. **2**, 2 (1997).
[11] J. Lahererre and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).
[12] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[13] M. E. J. Newman, e-print cond-mat/0007214.
[14] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, in *Proceedings of the Eighth World Wide Web Conference, Toronto, Canada, May 1999* (http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html).
[15] J. Kleinberg, R. Kumar, P. Raphavan, S. Rajagopalan, and A. Tomkins, in *Lecture Notes in Computer Science* (Springer-Verlag, Berlin, 1999), Vol. 1627.
[16] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, in *VLDB'99 Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, UK, September, 1999*, edited by Malcolm P. Atkinson, Maria E. Orlowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie (Morgan Kaufmann, San Francisco, 1999), pp. 639–650.
[17] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. Lukose, Science **280**, 95 (1998).
[18] B. A. Huberman and L. A. Adamic, Nature (London) **401**, 131 (1999).
[19] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comp. Commun. Rev. **29**(4), 251 (1999).
[20] A. L. Barabási and R. Albert, Science **286**, 509 (1999).
[21] A. Medina, I. Matta, and J. Byers, Comp. Commun. Rev. **30**(2), 18 (2000).
[22] R. R. Larson, in *Proceedings of the 1996 Annual ASIS Meeting, Baltimore, MD, October 1996* (http://www.asis.org/annual-96/ElectronicProceedings/index.html#5).
[23] S. Milgram, Psychol. Today **2**, 60 (1967).
[24] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998); M. Barthelemy and L. A. N. Amaral, Phys. Rev. Lett. **82**, 3180 (1999).
[25] A. L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[26] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E **62**, 1842 (2000).
[27] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, e-print cond-mat/0004434.
[28] P. L. Krapivsky, S. Redner, and F. Leyvraz, e-print cond-mat/0005139.
[29] M. H. Ernst, in *Fundamental Problems in Statistical Physics VI*, edited by E. G. D. Cohen (Elsevier, New York, 1985).
[30] F. Leyvraz and S. Redner, Phys. Rev. Lett. **57**, 163 (1986); Phys. Rev. A **36**, 4033 (1987).
[31] F. Calogero and F. Leyvraz, J. Phys. A **33**, 5619 (2000).
[32] H. A. Simon, Biometrika **42**, 425 (1955).
[33] H. A. Simon, *Models of Man* (Wiley, New York, 1957).
[34] D. H. Zanette and S. C. Manrubia, e-print nlin.AO/009046.
[35] E. Becker and W. Döring, Ann. Phys. (Leipzig) **24**, 719 (1935); see also J. S. Langer, in *Solids Far From Equilibrium*, edited by C. Godrèche (Cambridge University Press, Cambridge, 1992).
[36] E. M. Hendricks and M. H. Ernst, J. Colloid Interface Sci. **97**, 176 (1984).
[37] T. Matsoukas and E. Gulari, J. Colloid Interface Sci. **132**, 13 (1989).
[38] P. L. Krapivsky, J. F. F. Mendes, and S. Redner, Phys. Rev. B **59**, 15 950 (1999).
[39] P. L. Krapivsky, G. J. Rogers, and S. Redner (unpublished).
[40] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science* (Addison-Wesley, Reading, MA, 1989).
[41] See, e.g., N. V. Brilliantov and P. L. Krapivsky, Fiz. Tverd. Tela. (Leningrad) **31**, 172 (1989) [Sov. Phys. Solid State **31**, 271 (1989)]; J. Phys. A **24**, 4787 (1991); J. A. Blackman and A. Wielding, Europhys. Lett. **16**, 115 (1991); Ph. Laurencot, Nonlinearity **12**, 229 (1999).
[42] All of the citation data discussed here can be obtained from http://physics.bu.edu/~redner.
[43] M. Molloy and B. Reed, Random Struct. Alg. **6**, 161 (1995); Combin. Probab. Comput. **7**, 295 (1998).
[44] W. Aiello, F. Chung, and L. Lu (unpublished).
[45] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, e-print cond-mat/0007235.
[46] The apparently singular behavior at $y=2$ indicates that a scale finer than linear in k is required in the proximity of $y=2$. The proper scale is $\sqrt{k}$, and its use resolves the singularity and matches solutions for $y<2$ and $y>2$.
[47] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).
[48] R. Albert, H. Jeong, and A. L. Barabási, Nature (London) **401**, 130 (1999).
[49] A. Broder, R. Kumar, F. Maghoul, P. Raphavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Computer Networks **33**, 309 (2000).
[50] P. L. Krapivsky, G. J. Rodgers, and S. Redner, Phys. Rev. Lett. (to be published).