Connectivity of Growing Random Networks

P. L. Krapivsky,[1,2] S. Redner,[1] and F. Leyvraz[3]

[1]*Center for BioDynamics, Center for Polymer Studies, and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*CNRS, IRSAMC, Laboratoire de Physique Quantique, Université Paul Sabatier, 31062 Toulouse, France*
[3]*Centro Internacional de Ciencias, Cuernavaca, Morelos, Mexico*

(Received 8 May 2000)

A solution for the time- and age-dependent connectivity distribution of a growing random network is presented. The network is built by adding sites that link to earlier sites with a probability A_k which depends on the number of preexisting links k to that site. For homogeneous connection kernels, $A_k \sim k^\gamma$, different behaviors arise for $\gamma < 1$, $\gamma > 1$, and $\gamma = 1$. For $\gamma < 1$, the number of sites with k links, N_k, varies as a stretched exponential. For $\gamma > 1$, a single site connects to nearly all other sites. In the borderline case $A_k \sim k$, the power law $N_k \sim k^{-\nu}$ is found, where the exponent ν can be tuned to any value in the range $2 < \nu < \infty$.

PACS numbers: 84.35.+i, 05.40.–a, 05.50.+q, 87.18.Sn

Random networks play an important role in epidemiology, ecology (food webs), and many other fields. The geometry of such fixed topology networks have been extensively investigated [1–7]. However, networks based on human interactions, such as transportation systems, electrical distribution systems, biological systems, and the Internet, are open and continuously growing, and new approaches are rapidly developing to understand their structure and time evolution [8–12].

In this Letter, we apply a rate equation approach to solve the growing random network (GN) model, a special case of which was introduced in [13] to account for the distribution of citations and other growing networks [13–18]. Our approach is ideally suited for the GN and is much simpler than the standard probabilistic [1] or generating function [2] techniques. The rate equation formulation can be adapted to study more general evolving graph systems, such as networks with site deletion and link rearrangement.

The GN model is defined as follows. At each time step, a new site is added and a directed link to one of the earlier sites is created. In terms of citations, we may interpret the sites in Fig. 1 as publications, and the directed link from one paper to another as a citation to the earlier publication. This growing network has a directed tree graph topology where the basic elements are sites which are connected by directed links. The structure of this graph is determined by the connection kernel A_k, which is the probability that a newly introduced site links to an existing site with k links ($k - 1$ incoming and 1 outgoing). We will solve for the connectivity distribution $N_k(t)$, defined as the average number of sites with k links as a function of the connection kernel A_k.

We focus on a class of homogeneous connection kernels, $A_k = k^\gamma$, with $\gamma \geq 0$ reflecting the tendency of preferential linking to popular sites. As we shall show, the connectivity distribution crucially depends on whether γ is smaller than, larger than, or equal to unity. For $\gamma < 1$, the connectivity distribution decreases as a stretched exponential in k. The case $\gamma > 1$ leads to phenomenon akin to gelation [19] in which a single "gel" site connects to nearly every other site of the graph. For $\gamma > 2$, this phenomenon is so extreme that the number of connections between other sites is finite in an infinite graph. A power law distribution $N_k \sim k^{-\nu}$ arises *only* for $\gamma = 1$. In this case, finer details of the dependence of the connection kernel on k affect the exponent ν. Hence we consider a more general class of *asymptotically* linear connection kernels, $A_k \sim k$ as $k \to \infty$. We show that ν is tunable to any value in the range $2 < \nu < \infty$. In particular, we can naturally generate values of ν between 2 and 3, as observed in the web graph [10–12] and in movie actor collaboration networks [13].

The rate equations for the time evolution of the connectivity distribution $N_k(t)$ are

$$\frac{dN_k}{dt} = \frac{1}{M_\gamma}[(k-1)^\gamma N_{k-1} - k^\gamma N_k] + \delta_{k1}. \qquad (1)$$

The first term accounts for the process in which a site with $k - 1$ links is connected to the new site, leading to a gain in the number of sites with k links. This happens with probability $(k-1)^\gamma/M_\gamma$, where $M_\gamma(t) = \sum j^\gamma N_j(t)$ provides the proper normalization. A corresponding role is played by the second (loss) term on the right-hand side of Eq. (1). The last term accounts for the continuous introduction of new sites with no incoming links.

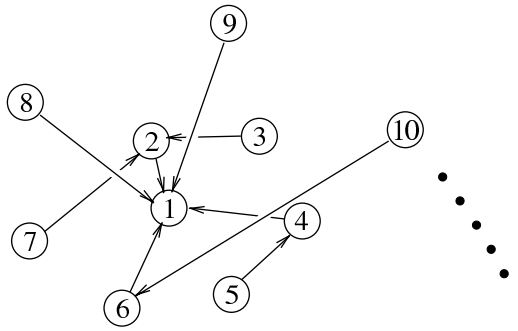

FIG. 1. Schematic illustration of the evolution of the growing random network. Sites are added sequentially and a single link joins the new site to an earlier site.

 © 2000 The American Physical Society

We start by finding the low-order moments $M_n(t)$ of the connectivity distribution. Summing Eqs. (1) over all k gives the rate equation for the total number of sites, $\dot{M}_0 = 1$, whose solution is $M_0(t) = M_0(0) + t$. The first moment (the total number of bond end points) obeys $\dot{M}_1 = 2$, which gives $M_1(t) = M_1(0) + 2t$. The first two moments are therefore *independent* of γ, while higher moments and the connectivity distribution itself do depend on γ.

For the linear connection kernel, Eqs. (1) can be solved for an arbitrary initial condition. We limit ourselves to the most interesting asymptotic regime ($t \to \infty$) where the initial condition is irrelevant. Using $M_1 = 2t$, we solve the first few of Eqs. (1) and obtain $N_1 = 2t/3$, $N_2 = t/6$, etc., which implies that the N_k grow linearly with time. Accordingly, we substitute $N_k(t) = tn_k$ in Eqs. (1) to yield the recursion relation $n_k = n_{k-1}(k-1)/(k+2)$. Solving for n_k then gives

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

To solve the model with a sublinear connection kernel, $0 < \gamma < 1$, notice that M_γ satisfies the obvious inequalities $M_0 \le M_\gamma \le M_1$. Consequently, in the long-time limit

$$M_\gamma = \mu t, \qquad 1 \le \mu \le 2, \tag{3}$$

with a yet undetermined prefactor $\mu = \mu(\gamma)$. Now substituting $N_k(t) = tn_k$ and $M_\gamma = \mu t$ into Eqs. (1) and again solving for n_k we obtain

$$n_k = \frac{\mu}{k^\gamma} \prod_{j=1}^{k} \left(1 + \frac{\mu}{j^\gamma}\right)^{-1}, \tag{4}$$

whose asymptotic behavior is

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

etc. This pattern in (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. (4).

To complete the solution for the n_k, we need to establish the dependence of the amplitude μ on γ. Using the defining relation $M_\gamma/t = \mu = \sum_{k\ge 1} k^\gamma n_k$, together with Eq. (4), we obtain the implicit relation for $\mu(\gamma)$,

$$\mu = \sum_{k=2}^{\infty} \prod_{j=2}^{k} \left(1 + \frac{\mu}{j^\gamma}\right)^{-1}. \tag{6}$$

Despite the simplicity of this exact expression, it is not easy to extract explicit information except for the limiting cases $\gamma = 0$ and $\gamma = 1$, where $\mu = 1$ and $\mu = 2$, respectively, and the corresponding connectivity distributions are given by $n_k = 2^{-k}$ and by Eq. (2). However, numerical evaluation shows that μ varies smoothly between 1 and 2 as γ increases from 0 to 1 (Fig. 2). This result, together with Eq. (4), provides a comprehensive description of the connectivity distribution in the regime $0 \le \gamma \le 1$. It is worth emphasizing that for $0.8 \lesssim \gamma \lesssim 1$, n_k depends weakly on γ for $1 \le k \le 1000$. Thus, it is difficult to discriminate between different γ's and even to distinguish a power law from a stretched exponential in the GN model. This subtlety was already encountered in the analysis of the citation distribution [15,16].

A striking feature of the GN model is that we can "tune" the exponent ν by augmenting the linear connection kernel to the asymptotically linear connection kernel, with $A_k \to a_\infty k$ as $k \to \infty$, but otherwise *arbitrary*. For this asymptotically linear kernel, by repeating the steps leading to Eq. (4) we find

$$n_k = \frac{\mu}{A_k} \prod_{j=1}^{k} \left(1 + \frac{\mu}{A_j}\right)^{-1}. \tag{7}$$

Expanding the product in Eq. (7) leads to $n_k \sim k^{-\nu}$ with $\nu = 1 + \mu/a_\infty$, while the amplitude μ is found from

$$\mu = A_1 \sum_{k=2}^{\infty} \prod_{j=2}^{k} \left(1 + \frac{\mu}{A_j}\right)^{-1}. \tag{8}$$

As an explicit example, consider the connection kernel $A_1 = 1$ and $A_k = a_\infty k$ for $k \ge 2$. In this case, we can reduce Eq. (8) to a quadratic equation from which we obtain $\nu = (3 + \sqrt{1 + 8/a_\infty})/2$ which can indeed be tuned to *any* value larger than 2.

The GN model with superlinear connection kernels, $\gamma > 1$, exhibits a "winner takes all" phenomenon, namely, the emergence of a single dominant gel site which is linked to almost every other site. A particularly singular behavior occurs for $\gamma > 2$, where there is a nonzero probability that the initial site is connected to every other site of the graph. To determine this probability, it is convenient to consider a discrete time version process where one site is introduced at each step which always links to the initial site. After N steps, the probability that the new site will link to the

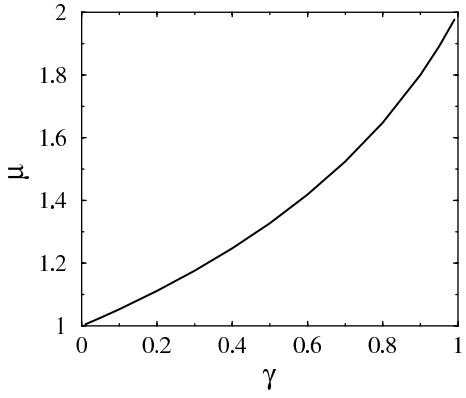

FIG. 2. The amplitude μ in $M_\gamma(t) = \mu t$ vs γ.

initial site is $N^\gamma/(N+N^\gamma)$. This pattern continues indefinitely with probability

$$\mathcal{P} = \prod_{N=1}^{\infty} \frac{1}{1+N^{1-\gamma}}. \tag{9}$$

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

To determine the behavior for general $\gamma>1$, we 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

$$\begin{aligned} N_k(k) &= \frac{(k-1)^\gamma N_{k-1}(k-1)}{M_\gamma(k-1)} \\ &= N_2(2)\prod_{j=2}^{k-1}\frac{j^\gamma}{M_\gamma(j)}. \end{aligned} \tag{10}$$

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:

$$\begin{aligned} M_\gamma(t) &= \sum_{k=1}^{t} k^\gamma N_k(t) \\ &\le t^{\gamma-1}\sum_{k=1}^{t} kN_k(t) = t^{\gamma-1}M_1(t). \end{aligned} \tag{11}$$

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

We now use $M_\gamma \sim t^\gamma$ in the rate equations to solve recursively for each N_k. Starting with the equation $\dot{N}_1 = 1 - N_1/M_\gamma$, the second term on the right-hand side is subdominant; neglecting this term gives $N_1 = t$. Continuing this same line of reasoning for each successive rate equation gives the leading behavior of N_k,

$$N_k = J_k t^{k-(k-1)\gamma} \quad \text{for } k \ge 1, \tag{12}$$

with $J_k = \prod_{j=1}^{k-1} j^\gamma/[1+j(1-\gamma)]$. 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 behavior of the N_k may be determined straightforwardly by keeping the next correction terms in the rate equations. For example, $N_1 = t - t^{2-\gamma}/(2-\gamma)+\ldots$.

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

Thus 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 sites are linked to the gel site which has the rest of the links of the network. This is the "winner takes all" situation. For $3/2<\gamma<2$, the number of sites with two links grows as $t^{2-\gamma}$, while the number of sites with more than two links is again finite. For $4/3<\gamma<3/2$, the number of sites with three links grows as $t^{3-2\gamma}$ and the number with more than three is finite. Generally, for $\frac{m+1}{m}<\gamma<\frac{m}{m-1}$, the number of sites with more than m links is finite, while $N_k \sim t^{k-(k-1)\gamma}$ for $k\le m$. Logarithmic corrections also arise at the transition points.

The connectivity distribution leads to an amusing consequence for the most popular site. Its connectivity $k_{\max}$ is determined by $\sum_{k>k_{\max}} N_k = 1$; that is, there is one site whose connectivity lies in the range $(k_{\max},\infty)$. This criterion gives

$$k_{\max} \sim \begin{cases} (\ln t)^{1/(1-\gamma)} & 0\le\gamma<1; \\ t^{1/(\nu-1)} & \text{asymptotically linear;} \\ t & \text{superlinear.} \end{cases} \tag{13}$$

Since t also equals the total number of sites, we can compare this prediction about the most popular site with available data from the Institute of Scientific Information based on 783 339 papers with 6 716 198 total citations (details in Ref. [16]). Here the most cited paper had 8904 citations. This accords with the first line of Eq. (13) for $\gamma\approx 0.86$, and also with the second when $\nu\approx 2.5$.

In addition to the connectivity of a site, we also may ask about its *age*. Within the GN model, older sites should clearly be more highly connected. We quantify this feature and also determine how the connection kernel affects the combined age and connectivity distribution. Note that our model does *not* have explicit aging where the connection kernel depends on the age of each site; this feature is treated in Ref. [17].

Let $c_k(t,a)$ be the average number of sites of age a which have $k-1$ incoming links at time t. Here age a means that the site was introduced at time $t-a$. The quantity $c_k(t,a)$ evolves according to

$$\frac{\partial c_k}{\partial t} + \frac{\partial c_k}{\partial a} = \frac{1}{M_\gamma}[(k-1)^\gamma c_{k-1} - k^\gamma c_k] + \delta_{k1}\delta(a). \tag{14}$$

The second term on the left-hand side accounts for the aging of sites, while the right-hand side accounts for the (age independent) connection changing processes. Consider first the linear kernel, $A_k = k$. Let us focus again on the most interesting limit, namely, asymptotic behavior. Then we can disregard the initial condition and write $M_1(t) = 2t$. This transforms Eq. (14) into

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial a}\right)c_k = \frac{(k-1)c_{k-1} - kc_k}{2t} + \delta_{k1}\delta(a). \tag{15}$$

The homogeneous form of this equation suggests that solution should be self-similar. Specifically, one can seek a solution as a function of the *single* variable a/t rather than two separate variables, $c_k(t,a) = f_k(a/t)$. This simplifies

the partial differential equation (15) into an ordinary differential equation for $f_k(x)$ which can easily be solved. In terms of the original variables of a and t, we find

$$c_k(t,a) = \sqrt{1-\frac{a}{t}}\left[1-\sqrt{1-\frac{a}{t}}\right]^{k-1}. \qquad (16)$$

Notice that this age distribution satisfies the normalization requirement, $N_k(t) = \int_0^t da\, c_k(t,a)$. As expected, young sites (those with $a/t \to 0$) typically have a small connectivity while old sites have large connectivity. Further, old sites have a broad distribution of connectivities up to a characteristic number which asymptotically grows as $\langle k\rangle \sim (1-a/t)^{-1/2}$ as $a \to t$. These properties and related issues may be worthwhile to investigate in citation and other information networks.

Similarly, we can obtain $c_k(t,a)$ for the GN model with an arbitrary homogeneous connection kernel [20] which grows slower than linearly in k. Assuming a self-similar solution $c_k(t,a) = f_k(a/t)$, applying a Laplace transform, we find a recursion relation for $\hat{f}_k$ whose solution is identical in structure to Eq. (4). Although it appears impossible to perform the inverse Laplace transform in explicit form for arbitrary k, we can compute $c_k(t,a)$ for small k; for example, we find $c_1 = (1-a/t)^{1/\mu}$. The behavior also simplifies in the large-k limit. Here we find that the age of sites with k links is peaked about the value a_k which satisfies

$$\frac{a_k}{t} \simeq \begin{cases} 1-\exp(-\mu\frac{k^{1-\gamma}}{1-\gamma}) & \gamma < 1\,; \\ 1-\frac{12}{(k+3)(k+4)} & \gamma = 1\,. \end{cases} \qquad (17)$$

This shows how old sites are better connected.

In summary, we solved for both the connectivity distribution and the age-dependent structure of the growing random network. The most interesting connectivity arises in a network with an asymptotically linear connection kernel. Here the number of sites with k connections has the power-law form $N_k \sim k^{-\nu}$, with ν tunable to any value in the range $2 < \nu < \infty$. This accords with the connectivity distributions observed in various contemporary examples of growing networks.

We are grateful for Grants No. NSF INT9600232, No. NSF DMR9978902, and No. DGAPA IN112998 for financial support.

Note added.—While writing this manuscript we learned of Ref. [21] which overlaps some of our results. We thank J. Mendes for informing us of this work.

[1] B. Bollobás, *Random Graphs* (Academic Press, London, 1985).
[2] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs* (Wiley, New York, 2000).
[3] S. A. Kauffman, *The Origin of Order: Self-Organization and Selection in Evolution* (Oxford University Press, London, 1993).
[4] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[5] B. Derrida and H. Flyvbjerg, J. Phys. (Paris) **48**, 971 (1987).
[6] H. Flyvbjerg and N. J. Kjaer, J. Phys. A **21**, 1695 (1988).
[7] U. Bastolla and G. Parisi, Physica (Amsterdam) **98D**, 1 (1996).
[8] R. V. Solé and S. C. Manrubia, Phys. Rev. E **54**, R42 (1996).
[9] S. Jain and S. Krishna, Phys. Rev. Lett. **81**, 5684 (1998).
[10] J. Kleinberg, R. Kumar, P. Raphavan, S. Rajagopalan, and A. Tomkins, in *Proceedings of the International Conference on Combinatorics and Computing,* Lecture Notes in Computer Science Vol. 1627 (Springer-Verlag, Berlin, 1999).
[11] S. R. Kumar, P. Raphavan, S. Rajagopalan, and A. Tomkins, in *Proceedings of the 25th Very Large Databases Conference, Edinburgh, Scotland, 1999* (Morgan Kaufmann, Orlando, FL, 1999).
[12] A. Broder, R. Kumar, F. Maghoul, P. Raphavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Comput. Netw. **33**, 309 (2000).
[13] A. L. Barabási and R. Albert, Science **286**, 509 (1999).
[14] A. J. Lotka, J. Wash. Acad. Sci. **16**, 317 (1926); W. Shockley, Proc. IRE **45**, 279 (1957).
[15] J. Lahererre and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).
[16] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[17] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E **62**, 1842 (2000).
[18] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. Lukose, Science **280**, 95 (1998); S. M. Maurer and B. A. Huberman, nlin.CD/0003041.
[19] M. H. Ernst, in *Fundamental Problems in Stat. Physics VI,* edited by E. G. D. Cohen (Elsevier, New York, 1985).
[20] P. L. Krapivsky and S. Redner (to be published).
[21] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, cond-mat/0004434.