Rate Equation Approach for Growing Networks

P. L. Krapivsky[1] and S. Redner[1]

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

Abstract. The rate equations are applied to investigate the structure of growing networks. Within this framework, the degree distribution of a network in which nodes are introduced sequentially and attach to an earlier node of degree k with rate $A_k \sim k^\gamma$ is computed. Very different behaviors arise for $\gamma < 1$, $\gamma = 1$, and $\gamma > 1$. The rate equation approach is extended to determine the joint order-degree distribution, the degree correlations of neighboring nodes, as well as basic global properties. The complete solution for the degree distribution of a finite-size network is outlined. Some unusual properties associated with the most popular node are discussed; these follow simply from the order-degree distribution. Finally, a toy protein interaction network model is investigated, where the network grows by the processes of node duplication and particular form of random mutations. This system exhibits an infinite-order percolation transition, giant sample-specific fluctuations, and a non-universal degree distribution.

1 Introduction

In this contribution, we apply tools from statistical physics, in particular, the rate equations, to quantify geometrical properties of evolving networks [1]. The utility of the rate equations have been amply demonstrated for diverse non-equilibrium phenomena, such as aggregation [2], coarsening [3], and epitaxial surface growth [4]. We will argue that the rate equations are a similarly powerful yet simple tool to analyze growing network systems. In addition to providing comprehensive information about the node degree distribution, the rate equations can be readily adapted to treat the joint order-degree distribution, correlations between node degrees, global properties, and a variety of intriguing fluctuation effects.

We will focus on two classes of models. In the first, which we simply term the growing network, nodes are added sequentially and a single link is established between the new node and a pre-existing node according to an attachment rate A_k that depends only on the degree of the "target" node (Fig. 1). Here node degree is the number of links that impinge on the node. This appealing model, first introduced by Simon [5] and rediscovered by Barabási and Albert [6], has become extremely fashionable because of its rich phenomenology and timely applications. Examples include the distribution of biological genera, word frequencies, publications, urban populations, income [5,7], and the link distribution of the world-wide web [8–10].

The second class of models are inspired by protein interaction networks, where the nodes are individual proteins and the links represent a functional

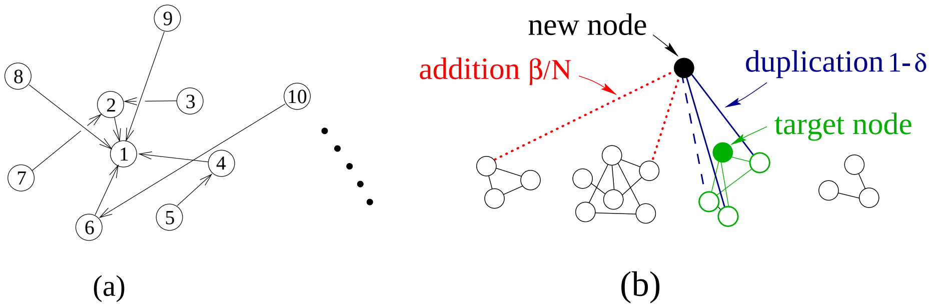

Fig. 1. (a) Growing network. Nodes are added sequentially and a single link joins a new node to an earlier node. Node 1 has degree 5, node 2 has degree 3, nodes 4 and 6 have degree 2, and the remaining nodes have degree 1. (b) Protein interaction network. The new node duplicates 2 out of the 3 links between the target (shaded green) and its neighbors. Each successful duplication occurs with probability $1-\delta$ (blue solid lines). The new node also attaches to any other node with probability β/N (red dotted lines). Thus three previously disconnected clusters are joined by the complete event

relationship between two proteins in an organism. Much effort has been devoted to infer the structure of such networks [11–13] and to formulate models that account for their evolution [14–19]. In the model discussed here [17,18], nodes are added sequentially and the new node may "duplicate" a randomly chosen target, and the new node can link to any other node with with a small probability (Fig. 1). In the duplication step, the new node links to each of the neighbors of the target with probability $1-\delta$. Thus the duplicate protein is functionally similar to the original [14]. The second process can be viewed as mutation in which a protein can becomes functionally linked to a random subset of other proteins. By this latter process, an arbitrary number of clusters can merge when a single node is introduced. As we shall discuss, this many-body merging leads to an infinite-order percolation transition as a function of the mutation rate. While the applicability of this model to describe real protein networks is still not settled [14], it is a useful starting point for theoretical analysis.

Our basic goal is to quantify the structure of these two basic networks by the rate equation approach.

2 Structure of the Growing Network

2.1 The Degree Distribution

A fundamental characteristic of any random network is the *node degree distribution* $N_k(N)$, defined as the number of nodes with k links in a network that contains N total nodes. To determine this distribution, we write the rate equations that account for its evolution after each node is introduced. For the growth process in Fig. 1(a), these rate equations are [20–22]

$$\frac{dN_k}{dN} = \frac{A_{k-1}N_{k-1} - A_k N_k}{A} + \delta_{k1}. \tag{1}$$

The first term on the right, $A_{k-1}N_{k-1}/A$, accounts for processes in which a node with $k-1$ links is connected to the new node, thus increasing N_k by one. Since there are N_{k-1} nodes of degree $k-1$, such processes occur at a rate proportional to $A_{k-1}N_{k-1}$, while the factor $A(N) = \sum_{j\geq 1} A_j N_j(N)$ converts this rate into a normalized probability. A corresponding role is played by the second (loss) term on the right-hand side. Here $A_k N_k/A$ is the probability that a node with k links is connected to the new node, thus leading to a loss in N_k. The last term accounts for the introduction of a new node with degree one.

Let us first determine the moments of the degree distribution, $M_n(N) = \sum_{j\geq 1} j^n N_j(N)$. Summing Eqs. (1) over all k, gives $\dot{M}_0(N) = 1$. This accords with the definition that $M_0(N) = \sum_k N_k$ is just the total number of nodes N in the network. Similarly, the first moment obeys $\dot{M}_1(N) = 2$, or $M_1(N) = M_1(0) + 2N$. Clearly this quantity must grow as $2N$, since introducing a single node creates two link endpoints. Thus the first two moments are *independent* of the attachment kernel A_k and grow linearly in N. On the other hand, higher moments and the degree distribution itself depend in an essential way on A_k.

For general attachment kernels that do not grow faster than linearly with k, it can be easily verified that the asymptotic degree distribution and $A(N)$ both grow linearly with N. Thus substituting $N_k(N) = N\,n_k$ and $A(N) = \mu N$ into Eqs. (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 the formal solution

$$n_k = \frac{\mu}{A_k} \prod_{j=1}^{k} \left(1 + \frac{\mu}{A_j}\right)^{-1}. \tag{2}$$

To complete this solution, we need the amplitude μ. Using the definition $\mu = \sum_{j\geq 1} A_j n_j$ in (2), we obtain the implicit relation

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

which shows that the amplitude μ depends on the entire attachment kernel.

For the generic case $A_k \sim k^\gamma$, we rewrite the product in (2) as the exponential of a sum of logarithms. In the continuum limit, we convert this sum to an integral, expand the logarithm to lowest order, and evaluate the integral to yield:

$$n_k \sim \begin{cases} k^{-\gamma} \exp\left[-\mu\left(\frac{k^{1-\gamma} - 2^{1-\gamma}}{1-\gamma}\right)\right], & 0 \leq \gamma < 1; \\ k^{-\nu}, \quad \nu = 1 + \mu > 2, & \gamma = 1; \\ \text{singular} & \gamma > 1. \end{cases} \tag{4}$$

That is, for all $0 < \gamma < 1$, the degree distribution is a robust stretched exponential (and pure exponential for $\gamma = 0$). Conversely, for $\gamma > 1$ a phenomenon analogous to gelation occurs in which a single node has almost all of the network links [20,22]. The regime $\gamma > 1$ actually has an infinite sequence of transitions. For $\gamma > 2$ all but a *finite* number of nodes (in an infinite network) are linked to the "gel" node which has the rest of the links of the network. For $3/2 < \gamma < 2$, the number of nodes with two links grows as $N^{2-\gamma}$, 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 $N^{3-2\gamma}$ 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 $N_k \sim N^{k-(k-1)\gamma}$ for $k \leq m$.

The linear kernel ($\gamma = 1$) is on the boundary between these two generic behaviors and leads to a degree distribution that depends on details of the attachment rate. In fact, the exponent $\nu = 1 + \mu$ can be tuned to *any* value larger than 2 [22]. In the special case of the strictly linear kernel, $A_k = k$, the degree distribution has the simple form

$$n_k = \frac{4}{k(k+1)(k+2)} \propto k^{-3}. \tag{5}$$

To illustrate the vagaries of asymptotically linear kernels, consider the shifted linear kernel $A_k = k + \lambda$. For this case, note that $A(N) = \sum_j A_j N_j(N)$ gives $A(N) = M_1(N) + \lambda M_0(N)$. Using $A = \mu N$, $M_0 = N$ and $M_1 = 2N$, we get $\mu = 2 + \lambda$. Hence $\nu = 1 + \mu = 3 + \lambda$. Thus an additive shift in the attachment kernel profoundly affects the asymptotic degree distribution. From (2), the degree distribution is

$$n_k = (2+\lambda)\,\frac{\Gamma(3+2\lambda)}{\Gamma(1+\lambda)}\,\frac{\Gamma(k+\lambda)}{\Gamma(k+3+2\lambda)} \propto k^{-(3+\lambda)}. \tag{6}$$

Finally, we discuss a simple extension in which a newly-introduced node links to exactly p earlier nodes [6]. For the linear attachment kernel, the degree distribution $N_k(N)$ (now defined only for $k \geq p$) obeys the rate equation

$$\frac{dN_k}{dN} = \frac{p}{M_1}\left[(k-1)N_{k-1} - kN_k\right] + \delta_{k,p}. \tag{7}$$

Following the basic approach outlined after (3), we find that the asymptotic degree distribution, $n_k = N_k/N$, is [22]

$$n_k = \frac{2p(p+1)}{k(k+1)(k+2)} \qquad \text{for} \quad k \geq p. \tag{8}$$

Thus for the strictly linear attachment kernel, the number p of links introduced at each node creation event does not affect the exponent of the degree distribution. Generally, however, this multiple link construction affects the degree distribution. For example, for the shifted linear kernel, we find

$$\begin{aligned} n_k &= \text{const.} \times \frac{\Gamma(k+\lambda)}{\Gamma(k+3+\lambda+\lambda/p)} \quad \text{for } k \geq p, \\ n_p &= \left(1 + p\,\frac{p+\lambda}{2p+\lambda}\right)^{-1}, \end{aligned} \tag{9}$$

whose asymptotic behavior is $n_k \sim k^{-(3+\lambda/p)}$. Thus the degree distribution exponent depends strongly on p. This result again shows that fine details of the growth process can be vitally important when the attachment rate is asymptotically linear.

2.2 Order Distribution

In addition to node degree, we further characterize a node according to its order of introduction by associating an order index J to the J^{th} node that was introduced into the network [22,23]. Let $\mathcal{N}_k(N,J)$ be the probability that the J^{th} node has degree k when the network has N total nodes. The original degree distribution may be recovered from this joint order-degree distribution through $N_k(N) = \sum_{J=1}^{N} \mathcal{N}_k(N,J)$. The joint distribution evolves according to the rate equation

$$\left(\frac{\partial}{\partial N} - \frac{\partial}{\partial J}\right)\mathcal{N}_k = \frac{A_{k-1}\mathcal{N}_{k-1} - A_k\mathcal{N}_k}{A} + \delta_{k1}\delta(N-J). \tag{10}$$

The second term on the left account for the order index evolution. We assume that the probability of linking to a given node depends only on its degree and not on its order.

The homogeneous form of this equation suggests that the solution should depend on the *single* variable $x \equiv J/N$. Writing $\mathcal{N}_k(N,J) = f_k(x)$, converts (10) into an ordinary, and readily soluble, differential equation [22]. For the two generic cases of $A_k = k$ and $A_k = 1$, the order-degree distributions are:

$$\mathcal{N}_k(N,J) = \begin{cases} \sqrt{\frac{J}{N}}\left(1-\sqrt{\frac{J}{N}}\right)^{k-1} & A_k = k, \\ \frac{J}{N}\frac{[\ln(N/J)]^{k-1}}{(k-1)!} & A_k = 1. \end{cases} \tag{11}$$

For the average order index $\langle J_k\rangle = \sum_k J\, N_k(N,J)/N_k(N)$ of a node of degree k, we find

$$\frac{\langle J_k\rangle}{N} = \begin{cases} \frac{12}{(k+3)(k+4)} & A_k = k, \\ (2/3)^k & A_k = 1. \end{cases} \tag{12}$$

Similarly, the average degree $\langle k_J\rangle = \sum_k k\, N_k(N,J)$ of a node of order index J is

$$\langle k_J\rangle = \begin{cases} (J/N)^{-1/2} & A_k = k, \\ \ln(N/J)+1 & A_k = 1. \end{cases} \tag{13}$$

The main messages from these results are that for $A_k = k$, high degree nodes must have been introduced early in the network development. Conversely, for the case of random attachment, $A_k = 1$, high degree nodes could also have been introduced relatively late in the network history. This difference plays a crucial role in determining the properties of the node with the highest degree (Section 3.2).

2.3 Degree Correlations

The rate equation approach also allows us to obtain degree correlations between connected nodes [22]. These develop because a node with large degree is likely to be old [22,24–26]. Thus its ancestor is also old and hence has a large degree. To quantify these degree correlations, define $C_{kl}(N)$ as the number of nodes of degree k that attach to an ancestor node of degree l (Fig. 2(a)). For example, in the network of Fig. 1, there are $N_1 = 6$ nodes of degree 1, with $C_{12} = C_{13} = C_{15} = 2$. There are also $N_2 = 2$ nodes of degree 2, with $C_{25} = 2$, and $N_3 = 1$ nodes of degree 3, with $C_{35} = 1$.

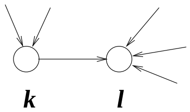

Fig. 2. Definition of the node degree correlation C_{kl} for $k = 3$ and $l = 4$

For simplicity, we consider the linear attachment kernel for which the degree correlation $C_{kl}(N)$ evolves according to

$$M_1 \frac{dC_{kl}}{dN} = (k-1)C_{k-1,l} - kC_{kl} + (l-1)C_{k,l-1} - lC_{kl} + (l-1)C_{l-1}\,\delta_{k1}. \quad (14)$$

The processes that gives rise to each term in this equation are illustrated in Fig. 3. The first two terms on the right account for the change in C_{kl} due to the addition of a link onto a node of degree $k-1$ (gain) or k (loss) respectively, while the second set of terms gives the change in C_{kl} due to the addition of a link onto the ancestor node. Finally, the last term accounts for the gain in C_{1l} due to the addition of a new node. A crucial feature of this equation is that it is closed; the 2-particle correlation function does not depend on 3-particle quantities.

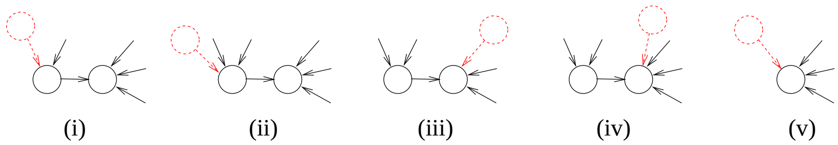

Fig. 3. Processes that contribute ((i)–(v) in order) to the terms in the rate equation (14) for the case $k = 3$ and $l = 4$ ((i)–(iv)). The newly-introduced node and link are shown dashed. The last case (v) arises only for $k = 1$

As in the case of the node degree, the N dependence is simply $C_{kl} = Nc_{kl}$. This reduces (14) to an N-independent recursion relation. While the details of the solution are unwieldy [22], the asymptotic solution is relatively simple in the scaling regime, $k \to \infty$ and $l \to \infty$ with $y = l/k$ finite:

$$c_{kl} = k^{-4}\,\frac{4y(y+4)}{(1+y)^4}. \tag{15}$$

For fixed large k, the distribution c_{kl} has a maximum at $y^* = (\sqrt{33}-5)/2 \cong 0.372$. Thus a node of degree k is typically attached to an ancestor node whose degree is 37% that of the daughter node. In general, when k and l are both large and their ratio is different from one, the limiting behaviors of c_{kl} are

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

Here we explicitly see the absence of factorization in the degree correlation: $c_{kl} \neq n_k n_l \propto (k\,l)^{-3}$.

2.4 Global Properties

The rate equations can be adapted to determine the *in-component* and *out-component* of the network with respect to a given node $\mathbf{x}$ [22]. The former is just the set of nodes that point to the node, plus all nodes that refer these daughter nodes, *etc.* The latter are the set of nodes that can be reached by following directed links that emanate from $\mathbf{x}$ (Fig. 4). We study the distribution of these component sizes for the constant attachment kernel, $A_k = 1$, because many results about components are independent of the form of the kernel and thus it suffices to consider the simplest situation.

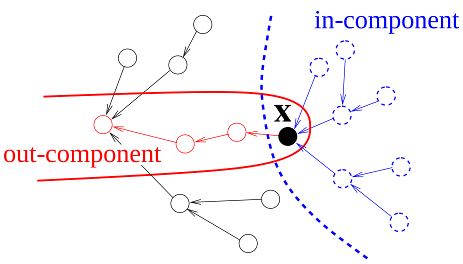

Fig. 4. In-component and out-components of node $\mathbf{x}$

The In-Component The number of in-components with s nodes, $I_s(N)$, satisfies the rate equation

$$\frac{dI_s}{dN} = \frac{(s-1)I_{s-1} - sI_s}{A} + \delta_{s1}. \tag{17}$$

Here the loss term accounts for processes in which the attachment of a new node to an in-component of size s increases its size by one. This gives a loss rate 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 $I_s(N)$ is simply $sI_s(N)$. A similar argument applies for the gain term. Dividing by $A(N) = \sum_j A_j N_j(N)$ converts these rates to probabilities, where $A(N) = M_0(N) \sim N$ for the constant attachment kernel.

It is again easy to verify that each I_s grows linearly in N. Thus we substitute $I_s(N) = N\, i_s$ into Eqs. (17) to obtain $i_s = i_{s-1}(s-1)/(s+1)$ and $i_1 = 1/2$. This immediately gives

$$i_s = \frac{1}{s(s+1)}. \tag{18}$$

The s^{-2} tail for the in-component distribution is independent of the form of the attachment kernel [22]. The exponent value also agrees with recent measurements of the web [10].

The Out-Component The complementary out-component (Fig. 4) from each node can be determined by mapping the out-component to an underlying network "genealogy". We build a genealogical tree for the growing network by taking generation $g = 0$ to be the initial node. Nodes that attach to those in generation g are defined to form generation $g + 1$; the node index does not matter in this characterization. For example, in the network of Fig. 1(a), node 1 is the ancestor of 6, while 10 is the descendant of 6; there are 5 nodes in generation $g = 1$ and 4 in $g = 2$ (Fig. 5).

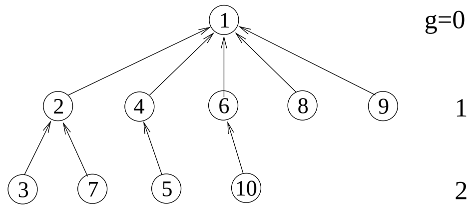

Fig. 5. Genealogy of the network in Fig. 1(a). The nodes indices indicate when each is introduced. The nodes are also arranged according to generation number

The genealogical tree is convenient because the number O_s of out-components with s nodes equals L_{s-1}, the number of nodes in generation $s-1$ in the tree (Fig. 5). We therefore compute $L_g(N)$, the size of generation g when the network has N total nodes. We again treat the constant attachment kernel; more general cases are treated in [22]. We determine $L_g(N)$ by noting that $L_g(N)$ increases when a new node attaches to a node in generation $g-1$. This occurs with rate L_{g-1}/M_0, where $M_0(N) = 1 + N$ is the number of nodes. Thus $\dot{L}_g(N) = L_{g-1}/(1+N)$, with solution $L_g(\tau) = \tau^g/g!$, where $\tau = \ln(1+N)$. Thus

$$O_s(\tau) = \tau^{s-1}/(s-1)!. \tag{19}$$

The generation size $L_g(N)$ rapidly grows with g for $g < \tau$, and then decreases and becomes of order 1 when $g = e\,\tau$. To accommodate a network of N nodes, the genealogical tree uses approximately $e\tau$ generations. Therefore the network diameter is $2e\tau \approx 2e\ln N$, since the maximum distance between any pair of nodes is twice the distance from the root to the last generation.

3 Finiteness, Fluctuations, and Extremes

3.1 Role of Finiteness

Thus far, we have focused on asymptotic properties when the number of nodes is sufficiently large that the ansatz $N_k = N\,n_k$ is valid. We now consider the role of finiteness on growing networks with attachment rate $A_k = k+\lambda$ ($\lambda > -1$) [27,28]. This interpolates between linear attachment (for $\lambda = 0$) to random attachment, $A_k = 1$ (for $\lambda \to \infty$).

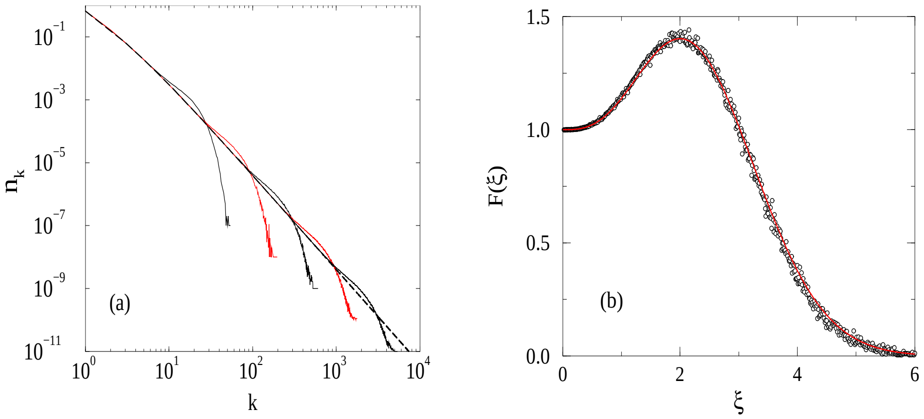

Fig. 6. (a) Normalized degree distribution for networks of $10^2, 10^3, \ldots, 10^6$ nodes (upper left to lower right), with 10^5 realizations for each N, for $A_k = k$ for a "triangle" initial condition. The dashed line is the asymptotic result $n_k = 4/[k(k+1)(k+2)]$; the last three data sets were averaged over 3, 9, and 27 points, respectively. (b) The corresponding scaling function as defined in $F(\xi)$ in (20) from simulation data of 10^6 realizations of a network with $N = 10^4$ nodes for the "dimer" initial condition (circles). The solid curve (red) is the analytical result of (25)

As quoted in (6), the degree distribution of a network with $N \gg 1$ nodes is $N_k(N) \propto N k^{-(3+w)}$ for attachment rate $A_k = k + \lambda$. However, for finite N the degree distribution must eventually deviate from this prediction because the maximal degree cannot exceed N. To establish the range of applicability of Eqs. (6), we estimate the largest degree in the network, $k_{\max}$ by the extreme statistics criterion $\sum_{k\geq k_{\max}} N_k(N) \approx 1$ [29]. This yields $k_{\max} \propto N^{1/(2+\lambda)}$. The degree distribution should therefore deviate from (6) when k becomes of the order of $k_{\max}$. The existence of a maximal degree suggests that the degree distribution should have the finite-size scaling form (see also [27,28,30–32])

$$N_k(N) \simeq N n_k F(\xi), \qquad \xi = k/k_{\max}. \tag{20}$$

To determine the finite-N behavior of the network, we start by writing the exact recursion relation for the degree distribution after a single node is added:

$$N_k(N+1) = N_k(N) + \frac{(k-1)N_{k-1}(N) - kN_k(N)}{2N}. \tag{21}$$

To solve this recursion we introduce the two-variable generating function [28]

$$\mathcal{N}(w,z) = \sum_{N=1}^{\infty}\sum_{k=1}^{\infty} N_k(N)\, w^{N-1}\, z^k\,, \tag{22}$$

to transform (21) into

$$\left(2(1-w)\,\frac{\partial}{\partial w} + z(1-z)\,\frac{\partial}{\partial z} - 2\right)\mathcal{N} = \frac{2z}{(1-w)^2}\,. \tag{23}$$

The exact solution to this equation can be obtained by standard methods and has the unwieldy form [28],

$$\begin{aligned}\mathcal{N}(w,z) = \frac{(3-2z^{-1})}{(1-w)^2} - \frac{1}{1-w} + \frac{2(z^{-1}-1)}{(1-w)^{3/2}} + \frac{2(1-w)^{-1/2}}{(z^{-1}-1)+(1-w)^{1/2}}\\ -\frac{2(z^{-1}-1)^2}{(1-w)^2}\,\ln\left[1-z+z(1-w)^{1/2}\right].\end{aligned} \tag{24}$$

By expanding $\mathcal{N}(w,z)$, we can determine all the $N_k(N)$. By this approach, we find that the scaling function defined in (20) is

$$F(\xi) = \mathrm{erfc}\left(\frac{\xi}{2}\right) + \frac{2\xi+\xi^3}{\sqrt{4\pi}}\; e^{-\xi^2/4}\,, \tag{25}$$

where $\mathrm{erfc}(x)$ is the complementary error function. A related result was found previously in [27]. This scaling function quantitatively accounts for the large-degree tail of the degree distribution (Fig. 6(b)).

3.2 Extremes and Lead Changes

We now investigate properties associated with the statistics of the node with the largest degree – the most popular node [33]; see also [34]. The degree of this node can be determined by a simple extreme statistics argument [29,33,34]. Here we discuss related, socially-motivated questions of the identity of the most popular node (the leader). These include the dependence of the leader identity on network size, the rate at which lead changes occur, and the probability that a leader retains the lead as a function of network size.

Leader Identity We first determine the order index of the leader node. To start with an unambiguous leader, we initialize the system with 3 nodes, with the initial leader having degree 2 (and index 1) and the other two nodes having degree 1. A new leader arises when its degree exceeds that of the current leader. For the linear attachment rate, $A_k = k$, the average order index of the leader $J_{\rm lead}(N)$ saturates to a finite value of approximately 3.4 as $N \to \infty$ (Fig. 7(a)). With probability ≈ 0.9, the leader is one of the 10 earliest nodes, while the probability that the leader is not among the 30 earliest nodes is less than 0.01. Thus only the earliest nodes have appreciable probabilities to be the leader; the rich really do get richer. In the case of $A_k = k + \lambda$, the average index of the leader also saturates to a finite value that is an increasing function of λ.

For random attachment ($A_k = 1$), the leader index grows as $J_{\rm lead}(N) \sim N^{\psi}$ with $\psi \approx 0.41$ (Fig. 7). The leader is still an early node (since $\psi < 1$), but not necessarily one of the earliest. From our simulations, a node with index greater than 100 has a probability of approximately 10^{-2} of being the leader for a network of 10^5 nodes. Thus, in random attachment, the order of node creation plays a significant, but not deterministic, role in the identity of the leader node.

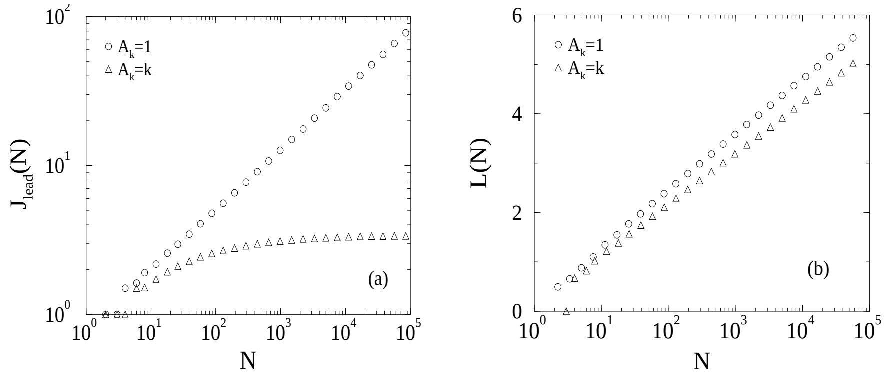

Fig. 7. (a) Average index of the leader $J_{\rm lead}(N)$ as a function of the total number of nodes N for 10^5 realizations of a growing network. Shown are the cases of attachment rates $A_k = 1$ and $A_k = k$. (b) Average number of lead changes $L(N)$ as a function of network size N for 10^5 realizations of the network for $A_k = 1$ and $A_k = k$

For constant attachment rate, the identity of the leader can be simply read off from (13); thus the index of the leader node, $J_{\rm lead}(N) = N(2/3)^{k_{\rm max}}$ [33]. We estimate the maximum degree from the extremal criterion $\sum_{k \geq k_{\rm max}} N_k(N) \approx 1$. Using $N_k(N) = N/2^k$, we find $2^{k_{\rm max}} \approx N$, or $k_{\rm max} \sim \ln N / \ln 2$. Therefore

$$J_{\rm lead}(N) \propto N^{\psi}, \quad \text{with} \quad \psi = 2 - \frac{\ln 3}{\ln 2} \approx 0.415\,037,$$

in excellent agreement with our numerical results.

For the linear attachment rate, (13) now gives $J_k(N) \sim 12N/k^2$. Since $N_k(N) \sim 4N/k^3$, the extremal criterion $\sum_{k \geq k_{\rm max}} N_k(N) \approx 1$ now gives $k_{\rm max} \approx$

$\sqrt{N}$. Therefore $J_{\rm lead}(N) \sim 12N/k_{\rm max}^2 = \mathcal{O}(1)$ indeed saturates to a finite value. A similar result holds in the general case $A_k = k + \lambda$. Thus the leader is one of the first few nodes in the network.

Lead Changes The average number of lead changes $L(N)$ grows logarithmically in N for both $A_k = 1$ and $A_k = k$ (Fig. 7), although the details of the underlying distributions of the number of lead changes, $P(L)$, are quite different. For $A_k = 1$, $P(L)$ has a sharp peak, while for $A_k = k$, $P(L)$ has a significant tail that stems from repeated lead changes among the two leading nodes. We also observe numerically that the number of *distinct* nodes that enjoy the lead grows logarithmically in N.

This logarithmic behavior can be easily understood. For $A_k = 1$, the number of lead changes cannot exceed the maximal degree $k_{\rm max} \sim \ln N/\ln 2$. For the general case $A_k = k + \lambda$, when a new node is added, the lead changes if the leadership is currently shared between two (or more) nodes and the new node attaches to a co-leader. The number of co-leader nodes (with degree $k = k_{\rm max}$) is $N/k_{\rm max}^{3+\lambda}$, while the probability of attaching to a co-leader is $k_{\rm max}/N$. Thus the average number of lead changes satisfies

$$\frac{d}{dN}\,L(N) \propto \frac{k_{\rm max}}{N}\,\frac{N}{k_{\rm max}^{3+\lambda}}. \tag{26}$$

Since $k_{\rm max}$ grows as $N^{1/(2+\lambda)}$, (26) reduces to $dL(N)/dN \propto N^{-1}$ or $L(N) \propto \ln N$. This argument can be adapted to arbitrary attachment rates that do not grow faster than linearly with k.

Fate of the First Leader Finally, we study the survival probability $S(N)$ that a node that was initially in the lead (has the maximum degree) remains in the lead as the network evolves. For $A_k = k + \lambda$ with $\lambda < \infty$, $S(N)$ is non-zero as $N \to \infty$ (Fig. 8). Thus the rich get richer holds in a strong form – the lead never changes with a positive probability.

For constant attachment rate the situation is more interesting, as being rich at birth is not as deterministic an influence as in the case of linear attachment. Numerically, $S(N)$ decays very slowly to zero as $N \to \infty$ (Fig. 8); a power law $S(N) \propto N^{-\phi}$ is a reasonable fit, but the local exponent is still slowly decreasing at $N \approx 10^8$ where it has reached $\phi(N) \approx 0.18$. To understand this behavior, consider the degree distribution of the first node. This quantity satisfies the recursion relation

$$P(k,N) = \frac{1}{N}P(k-1,N-1) + \frac{N-1}{N}P(k,N-1) \tag{27}$$

which reduces to the convection-diffusion equation

$$\left(\frac{\partial}{\partial \ln N} + \frac{\partial}{\partial k}\right) P = \frac{1}{2}\,\frac{\partial^2 P}{\partial k^2} \tag{28}$$

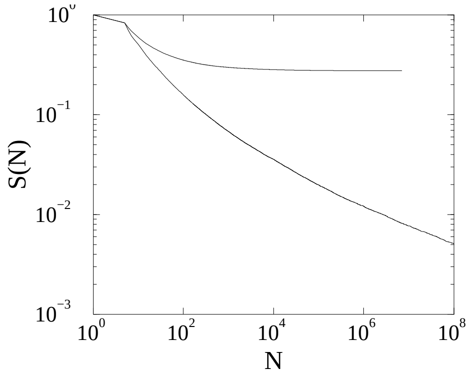

Fig. 8. Probability that the first node leads throughout the evolution for 10^5 realizations for $N \leq 10^7$ for $A_k = k$ (upper), and $N \leq 10^8$ for $A_k = 1$ (lower)

in the continuum limit. The solution is a Gaussian

$$P(k,N) = \frac{1}{\sqrt{2\pi \ln N}} \exp\left\{-\frac{(k-\ln N)^2}{2\ln N}\right\}. \tag{29}$$

Thus the degree of the first node grows as $\ln N$, with fluctuations of the order of $\sqrt{\ln N}$. On the other hand, from the degree distribution $N_k(N) = N/2^k$ the maximal degree grows as $k_{\max} = v \ln N$ with $v = 1/\ln 2 \approx 1.44$, and its fluctuations are negligible.

We now estimate $S(N)$ as the probability that the degree of the first node exceeds the maximal degree. For large N, this criterion, $S(N) \approx \sum_{k \geq k_{\max}} P(k,N)$, becomes

$$\begin{aligned} S(N) &\propto \int_{v\ln N}^{\infty} \frac{dk}{\sqrt{\ln N}} \exp\left\{-\frac{(k-\ln N)^2}{2\ln N}\right\} \\ &\propto N^{-\phi}\,(\ln N)^{-1/2}\,, \end{aligned} \tag{30}$$

with $\phi = (v-1)^2/2 \approx 0.0979\ldots$. The recursion (27) can, in fact, be solved exactly and gives $P(k,N) = \left[{N \atop k}\right]/N!$, for the dimer initial condition, where $\left[{N \atop k}\right]$ is the Stirling number of the first kind [35]. Using this instead of the Gaussian approximation leads to the exact exponent $\phi = 1 - v + v\ln v \approx 0.08607$. In either case, the logarithmic factor leads to the very slow approach to asymptotic behavior seen in Fig. 8.

4 Protein Networks

Finally, we study a toy protein interaction network model that evolves by the biologically-inspired processes of protein duplication and subsequent mutation, as illustrated in Fig. 1(b) [14,16–18]. By adapting the rate equation to account for these growth steps, we show that: (i) the system undergoes an infinite-order percolation transition as a function of mutation rate, with a rate-dependent power-law cluster-size distribution everywhere below the threshold, (ii) there are giant fluctuations in network structure and no self-averaging for large duplication rate, and (iii) the degree distribution has a power-law tail with a peculiar rate-dependent exponent.

4.1 Infinite-Order Percolation Transition

The protein network has rich percolation properties because the mutation process in Fig. 1(b) can lead to an arbitrary number of clusters being joined in a single step of the evolution. To study these percolation properties, we consider the simpler limit where mutations can occur, but no duplication ($\beta > 0, \delta = 1$). Let $C_s(N)$ be the number of clusters of size $s \geq 1$. This distribution obeys the rate equation

$$\frac{dC_s}{dN} = -\beta\,\frac{sC_s}{N} + \sum_{n=0}^{\infty}\frac{\beta^n}{n!}\,e^{-\beta}\sum_{s_1\cdots s_n}\prod_{j=1}^{n}\frac{s_jC_{s_j}}{N}, \tag{31}$$

where the sum is over all $s_1 \geq 1, \ldots, s_n \geq 1$ such that $s_1 + \cdots + s_n + 1 = s$. The first term on the right accounts for the loss of C_s due to the linking of a cluster of size s with the newly-introduced node. The gain term accounts for all possible merging processes of n initially separated clusters whose total size is $s - 1$.

Employing the now familiar ansatz that $C_s(N) = Nc_s$, and introducing the generating function $g(z) = \sum_{s\geq 1} sc_s\, e^{sz}$, (31) becomes

$$g = -\beta g' + (1 + \beta g')\, e^{z+\beta(g-1)}, \tag{32}$$

where $g' = dg/dz$. To detect the percolation transition, we use the fact that $g(0) = \sum sc_s$ is the fraction of nodes within finite clusters. Thus in the non-percolating phase $g(0) = 1$ and the average cluster size $\langle s\rangle = \sum s^2c_s = g'(0)$, while in the percolating phase the size of the infinite cluster (the giant component) is $NG = N(1 - g(0))$. To determine $g'(0)$, we substitute the expansion $g(z) = 1 + zg'(0) + \ldots$ into (32) and take the $z \to 0$ limit. This yields a quadratic equation for $g'(0)$, with solution

$$g'(0) = \langle s\rangle = \frac{1 - 2\beta - \sqrt{1-4\beta}}{2\beta^2}. \tag{33}$$

This real only for $\beta \leq 1/4$, thus identifying the percolation threshold as $\beta_c = 1/4$. For $\beta > \beta_c$, we express $g'(0)$ in terms of the size of the giant component by setting $z = 0$ in (32) to give

$$g'(0) = \frac{e^{-\beta G} + G - 1}{\beta\,(1 - e^{-\beta G})}. \tag{34}$$

As $\beta \to \beta_c$, we use $G \to 0$ to simplify (34) and find $\langle s\rangle \to (1 - \beta_c)\beta_c^{-2} = 12$. On the other hand, (33) shows that $\langle s\rangle \to 4$ when $\beta \to \beta_c$ from below. Thus the average size of finite clusters jumps discontinuously from 4 to 12 as β passes through $\beta_c = 1/4$.

The cluster size distribution c_s exhibits distinct behaviors below, at, and above the percolation transition. For $\beta < \beta_c$, the asymptotic behavior of c_s can be read off from the generating function as $z \to 0$. If c_s has the power-law behavior $c_s \sim B\,s^{-\tau}$ as $s \to \infty$, then the corresponding generating function $g(z)$ has the small-z expansion $g(z) = 1 + g'(0)\,z + B\Gamma(2-\tau)\,(-z)^{\tau-2} + \ldots$. The regular terms are needed to reproduce the known zeroth and first derivatives of the generating function, while the asymptotic behavior is controlled by the dominant singular term $(-z)^{\tau-2}$. Substituting this expansion into (32) we find that the dominant terms are of the order of $(-z)^{\tau-3}$. Balancing all contributions of this order gives

$$\tau = 1 + \frac{2}{1-\sqrt{1-4\beta}}. \tag{35}$$

Thus a power-law cluster size distribution with a non-universal exponent arises for *all* $\beta < \beta_c$; that is, the entire range $\beta < \beta_c$ is critical.

At the transition, (35) gives $\tau = 3$. However, $c_s \propto s^{-3}$ cannot be correct as it implies that $g'(0)$ diverges. The above expansion of the generating function is also not valid for $\tau = 3$. As in other such situations, we anticipate a logarithmic correction. A detailed analysis of the generating function under this assumption gives [18]

$$c_s \sim \frac{8}{s^3\,(\ln s)^2} \quad \text{as} \quad s \to \infty. \tag{36}$$

The size of the giant component $G(\beta)$ is obtained by solving (32) near $z = 0$. A detailed analysis shows that near β_c

$$G(\beta) \propto \exp\left(-\frac{\pi}{\sqrt{4\beta-1}}\right), \tag{37}$$

so that all derivatives of $G(\beta)$ vanish as $\beta \to \beta_c$. Thus the transition is of infinite order. Similar behaviors were observed [23,36–38] for growing network models where single nodes and links were introduced independently. This generic growth mechanism seems to give rise to fundamentally new percolation phenomena.

Giant Fluctuations In the complementary limit of no mutations ($\beta = 0$), individual realizations of the network evolution fluctuate strongly. We can understand the underlying mechanism for these fluctuations most directly by studying the limit of deterministic duplication ($\delta = 0$), where all the links of the duplicated protein are completed [18]. There is still a stochastic element in this growth, as the node to be duplicated is chosen randomly. Consider the generic initial state of two nodes that are joined by a single link. We denote this graph as $K_{1,1}$, following the graph theoretic terminology [39] that $K_{n,m}$ is the complete bipartite graph in which every node in the subgraph of size n is linked to every node in the subgraph of size m. Duplicating one of the nodes in $K_{1,1}$ gives $K_{2,1}$ or $K_{1,2}$, equiprobably. By continuing to duplicate nodes, it is easy to verify that at

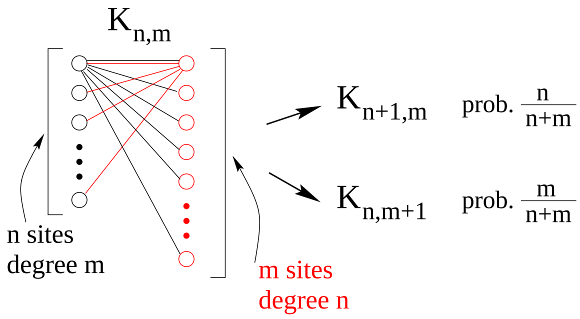

Fig. 9. Evolution of the complete bipartite graph $K_{m,n}$ after one deterministic duplication. Only the links emanating from the top nodes of each component are shown

every stage the network always remains a complete bipartite graph, say $K_{k,N-k}$, and that every value of $k = 1, \ldots, N-1$ occurs with equal probability (Fig. 9). Thus the degree distribution remains singular – it is always the sum of two delta functions! For fixed N, an average over all realizations of the evolution gives the *average* degree distribution

$$\langle N_k \rangle = 2\left(1 - \frac{k-1}{N-1}\right). \tag{38}$$

This loss of self averaging is generic; different realizations of the growth lead to statistically distinguishable networks for any initial condition. Similar giant fluctuations also arise in the general case of imperfect duplication where $\delta > 0$ [18].

4.2 Non-Universal Degree Distribution

Finally, consider the evolution when both incomplete duplication and mutation occur ($\delta < 1, \beta > 0$). In each growth step, the average number of links L increases by $\beta + (1-\delta)\mathcal{D}$ (Fig. 1(b)), where $\mathcal{D}$ is the average node degree of the network. Therefore, $L = [\beta + (1-\delta)\mathcal{D}]N$. Combining this with $\mathcal{D} = 2L/N$ gives [16,17]

$$\mathcal{D} = \frac{2\beta}{2\delta - 1}, \tag{39}$$

a result that applies only when $\delta > \delta_c = 1/2$. Below this threshold, the number of links grows as

$$\frac{dL}{dN} = \beta + 2(1-\delta)\,\frac{L}{N}, \tag{40}$$

and combining with $\mathcal{D}(N) = 2L(N)/N$, we find

$$\mathcal{D}(N) = \begin{cases} \text{finite} & \delta > 1/2, \\ \beta \ln N & \delta = 1/2, \\ \text{const.} \times N^{1-2\delta} & \delta < 1/2. \end{cases} \tag{41}$$

Without mutation ($\beta = 0$) the average node degree always scales as $N^{1-2\delta}$, so that a realistic finite average degree is recovered *only* when $\delta = 1/2$. Thus mutations play a constructive role, as a finite average degree arises for any duplication rate $\delta > 1/2$.

We now apply the rate equations to study the degree distribution $N_k(N)$ for this case of $\delta > 1/2$ and $\beta > 0$. The degree k of a node increases by one at a rate $A_k = (1-\delta)k + \beta$. The first term arises because of the contribution from duplication, while mutation leads to the k-independent contribution. The rate equations for the degree distribution are therefore

$$\frac{dN_k}{dN} = \frac{A_{k-1}N_{k-1} - A_k N_k}{N} + G_k. \tag{42}$$

The first two terms account for processes in which the node degree increases by one. The source term G_k describes the introduction of a new node of k links, with a of these links created by duplication and $b = k - a$ created by mutation. The probability of the former is $\sum_{s\geq a} n_s \binom{s}{a}(1-\delta)^a \delta^{s-a}$, where $n_s = N_s/N$ is the probability that a node of degree s is chosen for duplication, while the probability of the latter is $\beta^b\, e^{-\beta}/b!$. Since duplication and random attachment are independent processes, the source term is

$$G_k = \sum_{a+b=k} \sum_{s=a}^{\infty} n_s \binom{s}{a} (1-\delta)^a \delta^{s-a}\, \frac{\beta^b}{b!}\, e^{-\beta}. \tag{43}$$

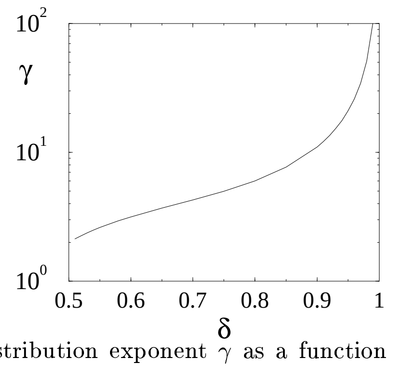

Fig. 10. The degree distribution exponent γ as a function of δ from the numerical solution of (46)

Substituting $N_k(N) = N\, n_k$ into the rate equations yields

$$\left(k + \frac{\beta+1}{1-\delta}\right) n_k = \left(k - 1 + \frac{\beta}{1-\delta}\right) n_{k-1} + \frac{G_k}{1-\delta}. \tag{44}$$

Since G_k depends on n_s for all $s \geq k$, the above equation is not a recursion. However, for large k, we reduce it to a recursion by noting that as $k \to \infty$, the main contribution to the sum in (43) arises when b is small. Thus a is close to k, and the summand is sharply peaked around $s \approx k/(1-\delta)$. We may then replace the lower limit by $s = k$, and n_s by its value at $s = k/(1-\delta)$. Further, if n_k decays as $k^{-\gamma}$, we write $n_s = (1-\delta)^\gamma n_k$ and simplify G_k to

$$\begin{aligned} G_k &\approx (1-\delta)^\gamma\, n_k \sum_{s=k}^{\infty} \binom{s}{k} (1-\delta)^k \delta^{s-k} \sum_{b=0}^{\infty} \frac{\beta^b}{b!}\, e^{-\beta} \\ &= (1-\delta)^{\gamma-1} n_k, \end{aligned} \tag{45}$$

since the former binomial sum equals $(1-\delta)^{-1}$.

These steps reduce (44) to a recursion, from which we deduce that n_k has the power-law behavior $n_k \sim k^{-\gamma}$, with γ determined from [18,40]

$$\gamma(\delta) = 1 + \frac{1}{1-\delta} - (1-\delta)^{\gamma-2}. \tag{46}$$

The exponent γ has a strong dependence on δ (Fig. 10). Further, since the replacement of n_s by $(1-\delta)^\gamma n_k$ is valid only asymptotically, the degree distribution should converge slowly to the predicted power law form. This slow approach to asymptotic behavior is observed in large-scale simulations [18]. The corresponding exponent $\gamma(\delta)$ is independent of the mutation rate β but depends sensitively on the duplication rate. Nevertheless, the presence of mutations ($\beta > 0$) is vital to suppress the non-self-averaging as the network evolves and thus make possible a smooth degree distribution.

5 Outlook

We hope that the reader is persuaded that the rate equations are a powerful, yet readily applicable tool, to investigate the structure of growing networks. For incrementally growing networks, we have obtained rather complete results for the degree distribution and some of the most important ensuing consequences. We also studied a toy protein interaction network model that evolves by duplication and mutation. In the absence of duplication, the network undergoes an infinite-order percolation transition as a function of the mutation rate. In the absence of mutation, the network exhibits giant sample-specific fluctuations. It is only with the inclusion of mutations that robust and statistically similar networks can be generated.

In summary, the rate equation approach is well-suited to treat a wide range phenomenology associated with evolving networks. Its full potential in this field is just starting to be fully exploited.

The work on protein networks was in collaboration with Byungnam Kahng and Jeenu Kim. This research was supported in part by NSF grant DMR9978902.

References

1. Recent reviews include: S. H. Strogatz: Nature **410**, 268 (2001); R. Albert, A.-L. Barabási: Rev. Mod. Phys. **74**, 47 (2002); S. N. Dorogovtsev, J. F. F. Mendes: Adv. Phys. **51**, 1079 (2002).
2. M. H. Ernst: in *Fractals in Physics*, edited by L. Pietronero, E. Tosatti (Elsevier, Amsterdam, 1986), p. 289.
3. A. J. Bray: Adv. Phys. **43**, 357 (1994).
4. A. Pimpinelli, J. Villain: *Physics of Crystal Growth* (Cambridge University Press, Cambridge, 1998).
5. H. A. Simon: Biometrica **42**, 425 (1955); H. A. Simon: *Models of Man* (Wiley, New York, 1957).
6. A.-L. Barabási, R. Albert: Science **286**, 509 (1999); R. Albert, H. Jeong, A.-L. Barabási: Nature **401**, 130 (1999).
7. G. U. Yule: Phil. Trans. Roy. Soc. B **213**, 21 (1924); *The Statistical Study of Literary Vocabulary* (Cambridge University Press, Cambridge, 1944).
8. S. R. Kumar, P. Raphavan, S. Rajagopalan, A. Tomkins: in *Proc. 8th WWW Conf.* (1999); J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: in Lecture Notes in Computer Science, Vol. 1627 (Springer-Verlag, Berlin, 1999).
9. B. A. Huberman, L. A. Adamic: Nature **401**, 131 (1999); G. Caldarelli, R. Marchetti, L. Pietronero: Europhys. Lett. **52**, 386 (2000)
10. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener: Computer Networks **33**, 309 (2000).
11. P. L. Uetz *et al.*: Nature **403**, 623 (2000); E. M. Marcotte *et al.*: Nature **402**, 83 (1999); A. J. Enright *et al.*: Nature **402**, 86 (1999); T. Ito *et al.*: Proc. Natl. Acad. Sci. USA **97**, 1143 (2000); *ibid* **98**, 4569 (2001).
12. J.-C. Rain *et al.*: Nature **409**, 211 (2001).
13. H. Jeong *et al.*: Nature **411**, 41 (2001).
14. A. Wagner: Mol. Biol. Evol. **18**, 1283 (2001).
15. F. Slanina, M. Kotrla: Phys. Rev. E **62**, 6170 (2000).
16. A. Vazquez, A. Flammini, A. Maritan, A. Vespignani: *cond-mat*/0108043.
17. R. V. Solé, R. Pastor-Satorras, E. D. Smith, T. Kepler: Adv. Complex Syst. **5**, 43 (2002); R. Pastor-Satorras, E. D. Smith, R. V. Solé: preprint.
18. J. Kim, P. L. Krapivsky, B. Kahng, S. Redner: Phys. Rev. E **66**, 055101 (2002).
19. J. Berg, M. Lässig, A. Wagner: *cond-mat*/0207711.
20. P. L. Krapivsky, S. Redner, F. Leyvraz: Phys. Rev. Lett. **85**, 4629 (2000).
21. S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin: Phys. Rev. Lett. **85**, 4633 (2000).
22. P. L. Krapivsky, S. Redner: Phys. Rev. E **63**, 066123 (2001).
23. S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin: Phys. Rev. E **64**, 066110 (2001).
24. M. E. J. Newman: Phys. Rev. Lett. **89**, 208701 (2002).
25. P. Bialas, Z. Burda, J. Jurkiewicz, A. Krzywicki: *cond-mat*/0211527.
26. A. Vazquez: *cond-mat*/0211528.
27. S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin: Phys. Rev. E **63**, 062101 (2001).
28. P. L. Krapivsky, S. Redner: J. Phys. A **35** 9517 (2002).
29. J. Galambos: *The Asymptotic Theory of Extreme Order Statistics* (R.E. Krieger Publishing Co., Malabar, 1987).
30. D. H. Zanette, S. C. Manrubia: Physica A **295**, 1 (2001).

31. L. Kullmann, J. Kertész: Phys. Rev. E **63**, 051112 (2001); D. Lancaster: J. Phys. A **35**, 1179 (2002).
32. Z. Burda, J. D. Correia, A. Krzywicki: Phys. Rev. E **64**, 046118 (2001).
33. P. L. Krapivsky, S. Redner: Phys. Rev. Lett. **xx** xxxx (2002).
34. A. A. Moreira, J. S. de Andrade Jr., L. A. N. Amaral: *cond-mat*/0205411.
35. R. L. Graham, D. E. Knuth, O. Patashnik: *Concrete Mathematics: A Foundation for Computer Science*, (Reading, Mass.: Addison-Wesley, 1989).
36. P. L. Krapivsky, S. Redner: Computer Networks **39**, 261 (2002).
37. D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, S. H. Strogatz: Phys. Rev. E **64**, 041902 (2001).
38. M. Bauer, D. Bernard: *cond-mat*/0203232.
39. B. Bollobás: *Modern Graph Theory* (Springer, New York, 1998); S. Janson, T. Luczak, A. Rucinski, *Random Graphs* (Wiley, New York, 2000).
40. F. Chung, L. Lu, T. G. Dewey, D. J. Galas: *cond-mat*/02009008.