PHYSICAL REVIEW E **66**, 055101(R) (2002)

Infinite-order percolation and giant fluctuations in a protein interaction network

J. Kim,[1] P. L. Krapivsky,[2] B. Kahng,[1] and S. Redner[2]
[1]*School of Physics and Center for Theoretical Physics, Seoul National University, Seoul 151-747, Korea*
[2]*Center for BioDynamics, Center for Polymer Studies, and Department of Physics, Boston University, Boston, Massachusetts 02215*
(Received 12 March 2002; published 11 November 2002)

We investigate a model protein interaction network whose links represent interactions between individual proteins. This network evolves by the functional duplication of proteins, supplemented by random link addition to account for mutations. When link addition is dominant, an infinite-order percolation transition arises as a function of the addition rate. In the opposite limit of high duplication rate, the network exhibits giant structural fluctuations in different realizations. For biologically relevant growth rates, the node degree distribution has an algebraic tail with a peculiar rate dependence for the associated exponent.

DOI: 10.1103/PhysRevE.66.055101 PACS number(s): 87.18.Sn, 02.50.Cw, 05.40.−a, 05.50.+q

Interprotein interactions underlie the performance of vital biological functions. Organisms with sequenced genomes, such as the yeast *S. cerevisiae* [1], provide important test beds for analyzing protein interaction networks [2]. The number of interactions per protein of *S. cerevisiae* follows a power law [3–5], a feature common to many complex networks, such as the Internet, the World Wide Web, and metabolic networks [6]. Similar behavior is exhibited by protein interaction networks of various bacteria [7]. Based on the observational data, simple proteome growth models have recently been formulated to account for the evolution of this interaction network [8–11], where proteins are viewed as the nodes of a graph and links connect functionally related proteins.

In this work, we determine the structure of a minimal protein interaction network model that evolves by the biologically inspired processes of protein duplication and subsequent mutation. That is, the functionality of a duplicate protein is similar, but not identical, to the original and can gradually evolve with time due to mutations [4]. Within a rate equation approach [12,13], 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 an algebraic tail with a peculiar rate-dependent exponent when the duplication and mutation rates have biologically realistic values. Some aspects of this last result were recently seen in Refs. [10,11].

In the model, nodes are added sequentially and the new node duplicates a randomly chosen pre-existing "target" node, viz., the new node links to each of the neighbors of the target with probability $1-\delta$; each new node also links to any previous node with probability β/N, where N is the current total number of nodes (Fig. 1). Thus an arbitrary number of clusters can merge when a single node is introduced. As we now discuss, this unusual dynamics appears to be responsible for the unconventional percolation properties of this network in the limit of zero duplication rate but finite mutation rate ($\delta=0$, $\beta>0$).

Let $C_s(N)$ be the expected number of clusters of size $s \geqslant 1$. This cluster-size distribution obeys the rate equation

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

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

Solving for the first few $C_s(N)$, we see that they are all proportional to N. Thus writing $C_s(N) = Nc_s$, and introducing the generating function $g(z) = \Sigma_{s\geqslant 1} s c_s e^{sz}$, Eq. (1) becomes

$$g = -\beta g' + (1+\beta g')e^{z+\beta(g-1)}, \qquad (2)$$

where $g' = dg/dz$. To detect the percolation transition, we use the fact that $g(0) = \Sigma s c_s$ is the fraction of nodes within finite clusters. Thus the size of the infinite cluster (the giant component) is $NG = N[1-g(0)]$. Suppose that we are in the nonpercolating phase; this means that $g(0)=1$. In this regime, the average cluster size equals $\langle s\rangle = \Sigma s^2 c_s = g'(0)$. To determine $g'(0)$, we substitute the expansion $g(z) = 1 + zg'(0) + \cdots$ into Eq. (2) and take the $z\to 0$ limit. This yields a quadratic equation for $g'(0)$ with solution

$$g'(0) = \langle s\rangle = \frac{1-2\beta-\sqrt{1-4\beta}}{2\beta^2}. \qquad (3)$$

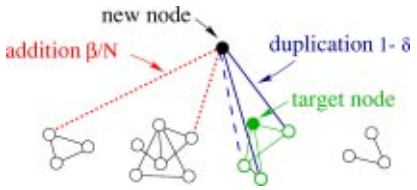

FIG. 1. Growth steps of the protein interaction network: The new node duplicates two out of the three links between the target node (shaded) and its neighbors. Each successful duplication occurs with probability $1-\delta$ (solid lines). The new node also attaches to any other network node with probability β/N (dotted lines). Thus three previously disconnected clusters are joined by the complete event.

This has a real solution only for $\beta \leqslant 1/4$, 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 Eq. (2) to give

$$g'(0) = \frac{e^{-\beta G} + G - 1}{\beta(1 - e^{-\beta G})}. \tag{4}$$

When $\beta \to \beta_c$, we use $G \to 0$ to simplify Eq. (4) and find $\langle s \rangle \to (1-\beta_c)\beta_c^{-2} = 12$. On the other hand, Eq. (3) shows that $\langle s \rangle \to 4$ when $\beta \to \beta_c$ from below. Thus the average size of the *finite* clusters jumps discontinuously from 4 to 12 as β passes through $\beta_c = \frac{1}{4}$.

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 behavior of the generating function as $z \to 0$. If c_s has the power-law behavior

$$c_s \sim B\, s^{-\tau} \quad \text{as} \quad s \to \infty, \tag{5}$$

then the corresponding generating function $g(z)$ has the following small-z expansion

$$g(z) = 1 + g'(0)z + B\Gamma(2-\tau)(-z)^{\tau-2} + \cdots. \tag{6}$$

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 $(-z)^{\tau-2}$. Higher-order regular terms are asymptotically irrelevant. Substituting this expansion into Eq. (2) we find that the dominant terms are of the order of $(-z)^{\tau-3}$. Balancing all contributions of this order gives

$$\tau = 1 + \frac{2}{1 - \sqrt{1-4\beta}}. \tag{7}$$

Intriguingly, a power-law cluster-size distribution with a nonuniversal exponent arises for *all* $\beta < \beta_c$. In contrast to ordinary critical phenomena, the entire range $\beta < \beta_c$ is critical.

The power-law tail implies that the size of the largest cluster $s_{\max}$ grows as a power law of the system size. From the extreme statistics criterion $\Sigma_{s \geqslant s_{\max}} N c_s = 1$ and the asymptotics of Eq. (5), we find $s_{\max} \propto N^{1/(\tau-1)}$, or $s_{\max} \propto N^{1/2 - \sqrt{\beta_c - \beta}}$. In contrast, for conventional percolation below threshold, the largest cluster has size $s_{\max} \propto \ln N$, reflecting the exponential tail of the cluster-size distribution [14].

At the transition, Eq. (7) gives $\tau = 3$. However, the naive asymptotics $c_s \propto s^{-3}$ cannot be correct as it implies that $g'(0)$ diverges. Similarly, we cannot expand the generating function as in Eq. (6) with $\tau = 3$, since the singular term $\Gamma(-1) \times (-z)$ has an infinite prefactor. As in other situations where the order of a singular term coincides with a regular term, we anticipate a logarithmic correction. Thus consider the modified expansion $g(z) = 1 + 4z + z\,u(z) + \cdots$, where $u(z)$ vanishes slower than any power of z, as $z \to 0$. Substituting this into Eq. (2), setting $\beta = \beta_c$, and equating singular terms yields $(8+u)z\,u' + u^2 = 0$. Solving this differential equation asymptotically we obtain the leading behavior $u \approx 8/\ln(-z)$; this indeed vanishes slower than any power of z for $z \to 0$. Substituting this form for $u(z)$ in the modified expansion for $g(z)$ and inverting yields

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

Thus exactly at the transition, the cluster-size distribution acquires a logarithmic correction. This result also implies that the size of the largest component scales as $s_{\max} \propto N^{1/2}/\ln N$.

Above the percolation transition, both $g(0) = 1 - G$ and $g'(0)$ [Eq. (4)] are finite, so that the expansion for $g(z)$ has the form $g(z) = 1 - G + g'(0)z + \cdots$. Substituting this into Eq. (2) one can show that: (i) the full expansion of $g(z)$ is regular in z, and (ii) the generating function diverges at $z_* = 1/s_*$. This latter fact implies that $c_s \propto e^{-s/s_*}$ as $s \to \infty$. The location of the singularity is determined by the condition $e^{z + \beta(g-1)} = 1$. This gives $s_* \to 16/G$ as $\beta \to \beta_c$. Realistic protein interaction networks are always above the percolation transition, e.g., for yeast the giant component includes 54% of all nodes and 68% of the links of the system [3]; thus a giant component always exists and the cluster-size distribution has an exponential tail.

The size of the giant component $G(\beta)$ is obtained by solving Eq. (2) near $z=0$. A lengthy analysis [15] shows that near the percolation threshold

$$G(\beta) \propto \exp\left(-\frac{\pi}{\sqrt{4\beta - 1}} \right), \tag{9}$$

so that all derivatives of $G(\beta)$ vanish as $\beta \to \beta_c$. Thus the transition is of infinite order. Similar behavior has been recently observed [16–18,13] for several growing network models where single nodes and links were introduced independently. This generic growth mechanism seems to give rise to fundamentally new percolation phenomena.

We now examine the complementary limit of no mutations ($\beta = 0$) and show that individual realizations of the evolution lead to widely differing results. Consider first the limit of deterministic duplication of $\delta = 0$, where all the links of the duplicated protein are completed. There is still a stochastic element in this growth, as the node to be duplicated is chosen randomly. When $\delta = 0$, the rate equation approach [Eqs. (14) and (15) below] predicts that the degree distribution N_k (defined as the number of nodes that are linked to k other nodes) is given by $N_k = 2(1-2/N)^{k-1}$.

However, this ''solution'' does not correspond to the outcome of any single realization of the duplication process. To appreciate this, consider the simple and 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 [19] that $K_{n,m}$ denotes a 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, one finds that at every stage the network always remains a complete bipartite graph, say $K_{k,N-k}$, and

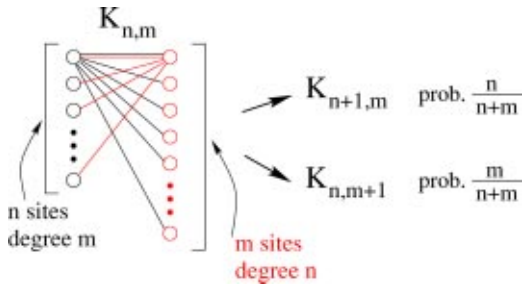

FIG. 2. Evolution of the complete bipartite graph $K_{m,n}$ after one deterministic duplication event. Only the links emanating from the top nodes of each component are shown.

that every value of $k=1,\ldots,N-1$ occurs with equal probability (Fig. 2). Thus the degree distribution remains singular—it is always the sum of two delta functions!

For fixed N, we average over all realizations of the evolution to obtain the *average* degree distribution

$$\langle N_k\rangle = 2\left(1-\frac{k-1}{N-1}\right). \tag{10}$$

Computing $\langle N_k\rangle$ for other generic initial conditions, e.g., complete m-partite graphs and ring graphs [15], we find that the initial condition dependence persists throughout the evolution. More importantly, self-averaging breaks down: different realizations of the growth lead to statistically distinguishable networks. Similar giant fluctuations arise in the general case of imperfect duplication where $\beta=0$ and $\delta>0$ [15]. To illustrate the origin of these macroscopic fluctuations, consider the network growth in the limit $\delta\ll 1$. The probability that the first few duplication steps are complete (all eligible links are created) is close to one. For this initial development, the degrees of each node increase and the probability to create isolated nodes becomes very small as the network grows. On the other hand, if the first duplication event was totally incomplete, an isolated node would be created. The creation of isolated nodes necessarily leads to more isolated nodes in subsequent duplication events. Thus the number of isolated nodes is a non-self-averaging quantity. In a similar fashion, the number of nodes of degree k for any finite $k>0$ is also non-self-averaging.

Finally, we investigate the evolution of the network when both incomplete duplication and mutation occur ($\delta<1$ and $\beta>0$). Let us first determine the average node degree of the network, $\mathcal{D}$, for such general rates. In each growth step, the average number of links L increases by $\beta+(1-\delta)\mathcal{D}$. Therefore, $L=[\beta+(1-\delta)\mathcal{D}]N$. Combining this with $\mathcal{D}=2L/N$ gives [9,10]

$$\mathcal{D}=\frac{2\beta}{2\delta-1}, \tag{11}$$

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

$$\frac{dL}{dN}=\beta+2(1-\delta)\frac{L}{N}, \tag{12}$$

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

$$\mathcal{D}(N)=\begin{cases}\text{finite}, & \delta>1/2\\ \beta\ln N, & \delta=1/2\\ \text{const}\times N^{1-2\delta}, & \delta<1/2.\end{cases} \tag{13}$$

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 consider this case of $\delta>1/2$ and $\beta>0$ and apply the rate equation approach [12,13] to study the degree distribution $N_k(N)$. 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

$$\frac{dN_k}{dN}=\frac{A_{k-1}N_{k-1}-A_kN_k}{N}+G_k. \tag{14}$$

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 $\Sigma_{s\geqslant a}n_s\binom{s}{a}(1-\delta)^a\delta^{s-a}$, 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 $\beta^be^{-\beta}/b!$. Since duplication and random attachment are independent processes, the source term is

$$G_k=\sum_{a+b=k}\sum_{s=a}^{\infty}n_s\binom{s}{a}(1-\delta)^a\delta^{s-a}\frac{\beta^b}{b!}e^{-\beta}. \tag{15}$$

From Eq. (14), the N_k grows linearly with N. Substituting $N_k(N)=N\,n_k$ in the rate equations yields

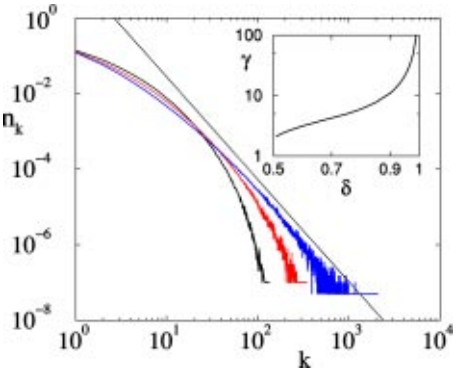

FIG. 3. Degree distribution n_k versus k for the protein interaction network with $\delta=0.53$ and $\beta=0.06$. Shown is the distribution for $N=10^3$, 10^4, and 10^6 (bottom to top), with 10^4, 10^3, and 20 realizations, respectively. A straight line (dotted) of the predicted slope of -2.37 is shown for visual reference. The inset shows the degree distribution exponent γ as a function of δ from the numerical solution of Eq. (18).

$$\left(k+\frac{\beta+1}{1-\delta}\right)n_k=\left(k-1+\frac{\beta}{1-\delta}\right)n_{k-1}+\frac{G_k}{1-\delta}. \quad (16)$$

Since G_k depends on n_s for all $s\geqslant k$, the above equation is not a recursion. However, for large k, we can reduce it to a recursion by simple approximations. As $k\to\infty$, the main contribution to the sum in Eq. (15) arises when b is small, so that a is close to k, and the summand is sharply peaked around $s\approx k/(1-\delta)$. This simplifies the sum, as we may 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

$$G_k\approx(1-\delta)^{\gamma}n_k\sum_{s=k}^{\infty}\binom{s}{k}(1-\delta)^k\delta^{s-k}\sum_{b=0}^{\infty}\frac{\beta^b}{b!}e^{-\beta}$$
$$=(1-\delta)^{\gamma-1}n_k, \quad (17)$$

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

Thus for $k\to\infty$, Eq. (16) reduces to a recursion relation, from which we deduce that n_k has the power-law behavior $\sim k^{-\gamma}$, with γ determined from the relation

$$\gamma(\delta)=1+\frac{1}{1-\delta}-(1-\delta)^{\gamma-2}. \quad (18)$$

Notice that the replacement of n_s by $(1-\delta)^{\gamma}n_k$ is valid only asymptotically. This explains the slow convergence of the degree distribution to the predicted power-law form (Fig. 3). Intriguingly, the exponent $\gamma(\delta)$ is independent of the mutation rate β [20]. 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. If we adopt $\delta=0.53$, as suggested by observations [4], we obtain $\gamma=2.373\cdots$, compared to the numerical simulation result of $\gamma=2.5\pm0.1$ [10].

In summary, network growth by duplication and mutation leads to rich behavior with an infinite-order percolation transition and no self-averaging in the absence of mutations. Without mutation, different realizations of the network lead to drastically different outcomes and each outcome is itself singular. Mutations are needed to form networks that are statistically similar to observed protein interaction networks. Thus mutations seem to play a constructive role in forming robust networks.

We are grateful for NSF Grant No. DMR9978902, KOSEF Grant No. 2002-2-11200-002-3 in the BRP program, and a travel grant from the BK21 project.

[1] A. Goffeau *et al.*, Science **274**, 546 (1996).
[2] P.L. Uetz *et al.*, Nature (London) **403**, 623 (2000); E.M. Marcotte *et al.*, *ibid.* **402**, 83 (1999); A.J. Enright *et al.*, *ibid.* **402**, 86 (1999).
[3] T. Ito *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 1143 (2000); *ibid.* **98**, 4569 (2001).
[4] A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001).
[5] H. Jeong *et al.*, Nature (London) **411**, 41 (2001).
[6] For reviews, see S.H. Strogatz, Nature (London) **410**, 268 (2001); R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002).
[7] J.-C. Rain *et al.*, Nature (London) **409**, 211 (2001).
[8] A similar network growth mechanism was proposed earlier in the context of biological evolution, F. Slanina and M. Kotrla, Phys. Rev. E **62**, 6170 (2000).
[9] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, e-print cond-mat/0108043.
[10] R.V. Solé, R. Pastor-Satorras, E.D. Smith, and T. Kepler, Adv. Complex Syst. **5**, 43 (2002).
[11] R. Pastor-Satorras, E. D. Smith, and R.V. Solé (unpublished).
[12] P.L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000); P.L. Krapivsky and S. Redner, Phys. Rev. E **63**, 066123 (2001).
[13] P.L. Krapivsky and S. Redner, Comput. Netw. **39**, 261 (2002).
[14] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor and Francis, London, 1992).
[15] J. Kim, P.L. Krapivsky, B. Kahng, and S. Redner (unpublished).
[16] D.S. Callaway, J.E. Hopcroft, J.M. Kleinberg, M.E.J. Newman, and S.H. Strogatz, Phys. Rev. E **64**, 041902 (2001).
[17] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, Phys. Rev. E **64**, 066110 (2001).
[18] M. Bauer and D. Bernard, e-print cond-mat/0203232.
[19] B. Bollobás, *Modern Graph Theory* (Springer, New York, 1998).
[20] Ref. [11] used an approximation to the rate equations that apply only as $\delta\to0$ and obtained a different form for the exponent γ than that predicted in Eq. (18).