Characterization of degree frequency distribution in protein interaction networks

Sebastián A. Romano and M. C. Eguia
Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes R. S. Peña 180 Bernal, B1876BXD Buenos Aires, Argentina

(Received 7 September 2004; revised manuscript received 12 November 2004; published 3 March 2005)

In this work, we analyze the degree frequency distribution in the yeast protein interaction network by studying a previously proposed duplication network model. This model correctly predicts the observed degree distribution (a power law for large degree values and a departure from this behavior for small degree). We numerically and analytically characterize this distribution as a mixture of random and power-law behavior, and make a comparative study of the robustness of the network model against realistic perturbations. We conclude that the particular distribution observed in both the model and the experimental data has many advantages in terms of dynamical and topological robustness and could have emerged in the evolutionary history as a sort of compromise between purely deterministic and random underlying mechanisms of network growth.

DOI: 10.1103/PhysRevE.71.031901 PACS number(s): 87.10.+e, 89.75.Hc, 89.75.Da, 87.23.Kg

The study of complex networks has seen an enormous rise in interest in the last few years, particularly since the discovery that a large variety of social, biological, and communicational networks share some common topological properties that deviate from those of random networks [1]. Among the most studied and celebrated of these topological properties are the small-world and scale-free features. Scale-free networks are highly heterogeneous: very few highly connected nodes (or "hubs") organize the wiring, and the frequency distribution of connectivity of nodes follows a power law, which is interpreted as a signature of nontrivial behavior.

The emergence of scale-free networks in the biological context (metabolic and protein interaction networks) has captured a great deal of attention because of the possible evolutionary advantages of a scale invariance at the level of cellular organization [2]. It was argued that the high resilience of some organisms against gene removal in gene-knockout experiments has its counterpart in the robustness against random node removal of the underlying scale-free network. Moreover, a topological property of the protein network (node connectivity) apparently correlates with protein dispensability [3]. However, there remain several controversial issues. The apparent robustness against gene removal is strongly conditioned by the nutrient-rich environment in the experiments [4], and it is not clear at all if the connectivity of a protein in the network is related to its evolutionary importance [5]. Also, there is some controversy regarding whether the apparent scale-free behavior is a result of selection or a side effect of the dynamics of network growth [6]. Furthermore, as we will show in this work, the scale-free nature of the protein interaction network can be hardly demonstrated from the experimental data available.

In a protein interaction network, two proteins are neighbors if they physically interact *in vivo*. These networks are obtained from a wealth of binary protein-protein interaction data from a variety of experiments: high-throughput methods, such as the two-hybrid assay [7], and mass-spectrometric complex identification [8]. The interactions derived from different datasets only match partially, and even though these datasets are cured in diverse databases, a host of false positives and/or negatives still remains [9]. Surprisingly, the only topological feature that is conserved between different datasets and remained qualitatively similar while new interactions were added is the distribution of connectivity between nodes, which has been claimed to be scale-free in previous works. Other topological magnitudes (such as the clustering coefficient) appear to be more sensitive to dataset variation.

In this work, we characterize the distribution of connectivity (or degree) between nodes in the yeast protein interaction network. The baker's yeast (*S. cerevisiae*) is the model organism of eukaryotes, and its interaction network dataset is the most cured among the species studied. We compare this network to a simple yet successful model of network growth via partial duplication of nodes introduced by Solé *et al.* [10]. This model mimics the main process of proteome development: gene duplication followed by a functional divergence of the cloned pairs. Unlike previous studies on duplication models that emphasized the scale-free behavior of the degree distribution for large connectivity values [11], we study the behavior of the distribution over the entire range of connectivity values. We perform a maximum likelihood analysis of the model and refine a previously derived analytical distribution. We also show that this particular shape is robust in parameter space and arises from the simultaneous action of an implicit preferential attachment and a random rewiring.

Finally, we compare the robustness of the yeast protein interaction network to the networks obtained from three different models (duplication-divergence, Barabási-Albert, and Erdös-Renyi) and discuss the results obtained in the light of more recent studies in protein dispensability and topological robustness.

Despite the enormous variety of proteins, they can be classified into families, according to similarities in structure and function. These families are explained by the hypothesis that their members have evolved from a common ancestor. It is thought that this evolution took place mostly through single or multiple gene duplication. After the random duplication of a gene there will be two cloned genes expressing the same redundant function. As a consequence, one or both duplicates should experience relaxed selective constraints and be more prone to mutations, becoming nonfunctional or, in some cases, gaining novel and beneficial functions. In this

1539-3755/2005/71(3)/031901(5)/$23.00

way, new genes that code for new proteins are created within a genome.

This mechanism can be translated into a few simple growing rules for a network model, as proposed by Solé *et al.* [10]. A description of the rules of the model follows. Starting from a small and arbitrary initial network, the graph consisting of N_0 nodes and E_0 edges suffers the following modifications in each discrete time step: (i) a node is randomly selected and copied, including its edges; (ii) the edges of the new node are deleted with probability δ; and (iii) new edges are randomly added between the copied node and any other node in the graph with probability α.

Step (i) mimics a gene duplication process, where the cloned nodes are linked to the same neighbors (i.e., have the same functions), and steps (ii) and (iii) represent functional divergence. From a geometrical point of view, the main attribute of a node is its degree k (number of edges).

This scheme is intended to capture only global topological properties of the proteome, since no protein functionality is included. However, it is remarkable that such a simple, biologically inspired model displays a degree frequency distribution $p(k)$ nearly identical to the real proteome. Previous works [10,11] have studied the contraints and possible values of the two free parameters of the model. By focusing on the average degree of the network $\langle k\rangle$, it is straightforward to see that: (i) the parameter α must be normalized by the total number of nodes. We consequently define $\alpha=\beta/N$ as the probability (per node) of adding new edges to the cloned node and use β as our control parameter. (ii) In order to obtain a stationary distribution $p(k)$, the deletion parameter ought to be $\delta>0.5$.

In order to gain some insight into the mechanisms that shape the network generated by this model, we start by writing down the variation over one time step (i.e., the master equation) of the average number of nodes with k edges N_k. Note that, since the growing of the network is uniform, the size of the network N plays the role of the discrete time $t=N-N_0$. The master equation for N_k was studied by Kim *et al.* [11]. However, while they focused in the large degree limit, we will focus on the prediction of the model for low and intermediate connectivity values. This master equation can be written as

$$N_k(N+1)-N_k(N)=\left[\alpha+\frac{(k-1)}{N}(1-\delta)(1-\alpha)\right]N_{k-1}(N) - \left[\alpha+\frac{k}{N}(1-\delta)(1-\alpha)\right]N_k(N)+G_k(N), \tag{1}$$

where

$$G_k(N)=\sum_{k'=0}^{N}\frac{N_{k'}}{N}\sum_{i=0}^{\min(k,k')}\binom{k'}{k'-i}\binom{N-i}{k-i} \times\delta^{k'-i}(1-\delta)^i\alpha^{k-i}(1-\alpha)^{N-k}. \tag{2}$$

A connection with the probability density for the degree of the network can be made for large network sizes $p(k)\approx N_k/N$. The first two terms on the right-hand side of Eq. (1) correspond to the contribution of nodes that are not the duplicated one. Factors of the form of k/N stand for the probability of a node of degree k to have a neighbor duplicated, increasing its degree by unity. These terms act only conveying the probability $p(k)$ from lower to higher degree values. On the other hand, the last term corresponds to the degree of the new node and contributes to all degree values. We will refer to this last term as the source term $G(k)$. The degree of the duplicated node is selected from the degree distribution $p(k')=N_{k'}/N$, as expressed in the first sum, and it undergoes a series of possible combinations of δ and β, condensed in the second sum, leaving as a result a distribution of new nodes with degree k.

It is interesting to analyze two limiting cases of the model: a case without creation of new edges ($\beta=0$) and the case where all edges are new ($\delta=1$ and $\beta>0$). In the first case, the number of isolated nodes $N_0(N)$ can only vary when a duplicated node with degree k' loses k' edges. Then, the variation in the population of the zero state (isolated nodes) is

$$N_0(N+1)-N_0(N)=\sum_{k'=0}^{N}\frac{N_{k'}(N)}{N}\delta^{k'}. \tag{3}$$

It is straightforward to analyze the stationary limit $N_k(N)=p(k)N$ (time dependence is through N only), where $p(k)=0$ for $k\geqslant 1$ and $p(0)=1$, so the network becomes the trivial one. In this limiting case, there is a decoupling between the dynamics of the isolated nodes and the rest of the network. When a node becomes disconnected it cannot be reconnected and, in the $N\gg 1$ limit, the fraction of connected nodes is negligible. If we are not interested in the stationary limit, then it can be demonstrated that the connected nodes have a power-law distribution [12]. In fact, this is the only case when one can obtain a pure scale-free degree distribution.

The other extreme case is when no edges are inherited by the cloned node. This is, actually, a purely random growing network, therefore, the distribution of connectivity between nodes is Poissonian.

Between these two limiting cases we obtain a mixture of power-law and random behavior. For $1>\delta>0.5$ and $\beta>0$ the degree distribution of the model displays a distinctive shape that is qualitatively similar to those of the experimental networks. This shape has a clear power-law tail only for the most connected nodes (approximately for $k>20$) and displays a severe departure of one or two orders of magnitude for low connectivity values.

We now characterize this particular distribution by deriving an approximate analytic solution of the master equation valid for all connectivities in the stationary limit. We follow a similar approach to that used by Krapivsky *et al.* in [13]. Assuming an asymptotic solution for the degree distribution $N_k(N)=p(k)N$, the source term (2) can be rewritten as

$$G_k^\infty \equiv \lim_{N\to\infty} G_k = \sum_{k'=0}^{\infty} p(k') \sum_{i=0}^{\min(k,k')} \binom{k'}{k'-i} \times \frac{\delta^{k'-i}}{(k-i)!}(1-\delta)^i \beta^{k-i} e^{-\beta} \tag{4}$$

and the rate equation (1) becomes

$$p(k) = [\beta + (k-1)(1-\delta)]p(k-1) - [\beta + k(1-\delta)]p(k) + G_k^\infty. \tag{5}$$

This equation reveals an implicit preferential attachment rule for the duplication model, since factors linear with k appear in terms corresponding to nodes linked to the duplicated one [14]. This preferential attachment rule is controlled by parameter δ and guarantees a power-law degree distribution [1]. On the other hand, terms with β contribute to the random character of the network growth. The roles played by the parameters of the model were apparent when we previously analyzed the two limiting cases. In order to obtain a more explicit expression for the stationary degree distribution we can derive, by iterating the preceding formula (5), the following recurrence relation:

$$p(k) = p(0)\frac{\left(k-1+\dfrac{\beta}{1-\delta}\right)!\left(\dfrac{1+\beta}{1-\delta}\right)!}{\left(\dfrac{\beta}{1-\delta}-1\right)!\left(k+\dfrac{1+\beta}{1-\delta}\right)!} + \frac{1}{(1-\delta)}H(k), \tag{6}$$

where

$$H(k) = \sum_{j=0}^{k-1} \frac{\left(k-1-j+\dfrac{1+\beta}{1-\delta}\right)!\left(k-1+\dfrac{\beta}{1-\delta}\right)!}{\left(k+\dfrac{1+\beta}{1-\delta}\right)!\left(k-1-j+\dfrac{\beta}{1-\delta}\right)!} G_{k-j}^\infty. \tag{7}$$

The coefficient $p(0)$ is calculated solving (1) for $k=0$.

As it was previously addressed, in the $k \gg 1$ limit, the expression (6) becomes a power-law with a δ dependent exponent [11]. Simulations confirm a slow convergence of the tail of $p(k)$ to this behavior. The power-law distribution is valid for large k because this limit endows the implicit preferential attachment with enough power to shadow the purely stochastic process of the model, controlled by β. The requirement for the onset of this regime is $k \gg (1+\beta)/(1-\delta)$, since this is the most relevant parameter repeatedly neglected when we approximated the factorials in (6) using the Stirling formula. In order to obtain an analytical approximation for the full degree distribution, we replace $p(k')$ in the source term Eq. (4) by a reasonable zeroth-order approximation and obtain a first-order approximation using (6) and (7). To improve the accuracy of the approximation, one could repeat this process by inserting the first-order approximation in the source term and obtaining a second-order approximation and so on. Therefore it is possible to calculate an approximate distribution with a reasonable degree of accuracy much more efficiently than numerically solving Eq. (1).

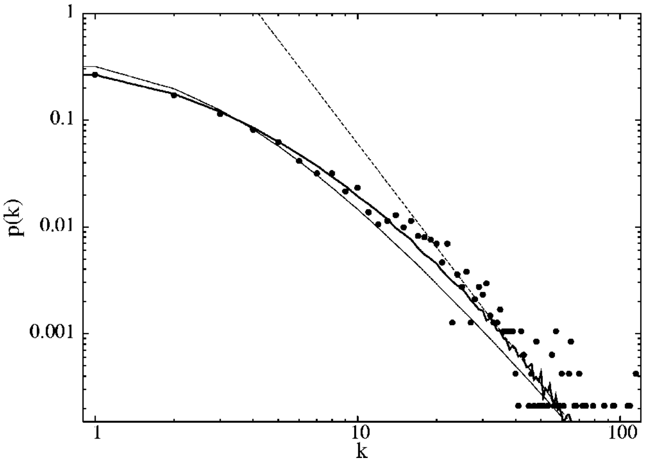

FIG. 1. The degree distribution obtained from averaged simulations of the duplication model evolved $N=6000$ time steps with parameters $\beta=0.22$, $\delta=0.55$ (thick line) compared to the distribution derived from the yeast proteome (black circles), the analytic approximation expressed by Eqs. (6) and (7) (thin line) and the asymptotic power-law distribution for high k values (dotted line).

Let us now compare the degree distribution of the model to the degree distribution obtained from a real proteome. As we already pointed out, the distributions are qualitatively similar for all parameter values that bring a stationary distribution (excluding some extreme cases). Furthermore, for some parameter values of the model we can also obtain a quantitative agreement between the two distributions. We used cured data from the DIP database [15] for the baker's yeast (*Saccharomyces cerevisiae*) proteome, the model organism of eucaryotes.

We performed a maximum likelihood analysis of the model, assuming a Gaussian dispersion of the experimental $p(k)$ values and obtained both a pair of optimal parameters ($\beta=0.22, \delta=0.55$) and a goodness of fit estimation. A maximum likelihood test against a power-law distribution gave us a goodness of fit estimator two orders of magnitude below of that of the model. Even when we take into account only $k > 20$ values (a range where the power law is well established), the goodness of the fit estimator is still greater for the model. Note that the power-law range includes only the 5% of the population. A noteworthy fact is that if we disregard the $k=0$ value, a lognormal distribution fit the experimental distribution quite well.

It should be pointed out that there are no consistent estimation measures of the divergence parameters. Hence, the only purpose of fitting the model to the experimental data is to discern if these mechanisms of evolution could be responsible for qualitatively shaping the proteome.

In Fig. 1 we show the degree distributions obtained from the yeast proteome, simulations of the model with parameter values obtained from the maximum likelihood analysis, and the approximate analytic solution (fourth order) obtained from Eqs. (6) and (7). As can be seen in the figure, the model correctly predicts the departure of the power-law behavior for low connectivity values.

As far as we know, previous works have not focused on the departure from the scale-free behavior for low k values. The work of Jeong *et al.* [3] proposed a phenomenological curve: a generalized power law with an exponential cutoff for large k values, and subsequent studies focused on the scale invariance as the relevant feature. Now two questions naturally arise: (a) Is this departure from the power-law behavior significant or is it a product of experimental biases? (b) Does it reflect some topological property of the network that is relevant to the protein interaction network?

In order to answer the first question we checked that the departure is not a product of experimental biases of false negatives and/or positives of the assessing methods. We studied three different datasets for the yeast proteome: noncured data from double-hybrid experiments (with many false positives), core dataset from the DIP database (with presumably many false negatives), and the more confident cured data from DIP (displayed in Fig. 1). In the three cases we observed the low connectivity departure and the model fitted the dataset for some parameter values. Even for protein interaction networks datasets obtained from more recent high-throughput experiments from other species (*D. melanogaster* [17] and *C. elegans* [16]) with less reliable (or not cured) data, we observed a qualitatively curved distribution and a good agreement with the duplication model (data not shown).

Our second question can be investigated through the more relevant property derived from scale-free networks: its topological robustness. Previous works that highlighted the power-law character of the distributions claimed that being scale-free, protein networks could be more robust to random mutations and, therefore, in a more favorable position (from an evolutionary point of view). In fact, it is well known that scale-free networks are more robust than random networks under accidental node removal [18]. However, some of the "worst" node attacks (such as hubs) could be much more harmful in scale-free than in purely random networks. From this last observation it is clear that scale-invariant networks could not be the more robust topology in the long term (although rare, hubs removal could happen).

In order to investigate this last hypothesis we performed a comparative analysis for the topological robustness for three different models with three characteristic degree distributions: (a) a random network or duplication model with $\delta=1$ (peaked distribution), (b) a duplication model with parameters that fit the curved experimental distribution, and (c) pure scale-free distribution (obtained with the Barabási-Albert model [1]).

The property of topological robustness takes account of the insensitivity of some global measure of the network to specific changes in its structure. In a biological context we have to define a good global equivalent of "fitness" or viability for a protein interaction network and a realistic perturbation in its internal structure. Because there are no flows inside a topological representation of the protein interaction network, the more adequate measure of viablity is the topological closeness [19] (also called efficiency in [20]). This quantity grows with both the degree of compactness of the network and the closeness between nodes (measured by the inverse of the path length) and decreases as the network breaks up into components and as the length of the paths between nodes increases (two things that clearly deteriorate the performance of the protein interaction network).

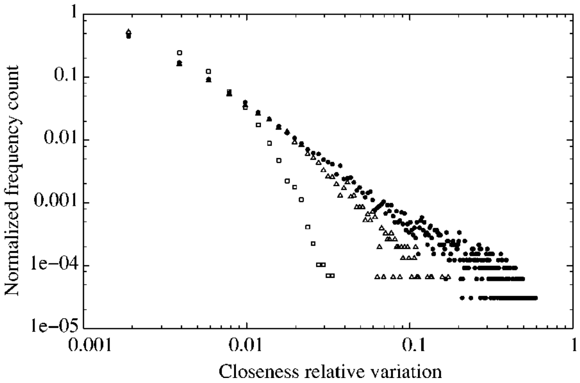

FIG. 2. Histograms of the relative variation of the global closeness of the networks, after removing a single node. The networks were created by a purely random model (squares), the duplication-divergence model with parameter values $\beta=0.2$ and $\delta=0.65$ (triangles), and the Barabási-Albert model (circles). Computations are made over 100 realizations for each model, every realization consisting of the same number of nodes ($N\approx 1000$) and links ($\langle k\rangle \approx 2$).

The usual perturbation for the internal structure of a network is to sequentially remove nodes until the network breaks up. However, this is not a biologically reasonable perturbation because a protein interaction network with half of the nodes removed is not viable [4]. We make use of a more realistic failure simulation: a random deletion of one or very few genes (nodes). Also, as we are interested in all possible cases, we take account of all closeness variations over the whole network against all possible node removals. In this way we can measure the network robustness in terms of a topological analog of the viability of a protein network against realistic perturbations.

In Fig. 2 we display the histograms obtained for the relative closeness variation against all possible single-node removals for the three network models. In all cases we started with networks of $N\approx 1000$ and $\langle k\rangle \approx 2$. Even when the three networks have an average closeness variation of the same order of magnitude, the three distributions are clearly different. In the case of the random network there are many node removals that cause drops in the closeness of 1%, but there are no cases of catastrophic node removals (none of the perturbations cause drops greater than 5%). On the contrary, most of the perturbations of the pure scale-free network produce a very small effect in the global closeness, while there is a small but significative fraction of node removals (5% of the cases) that causes great closeness variations (drops of more than 5%) and a few that cause a catastrophic drop-off in global closeness (note the long tail of the distribution). In this context, the duplication model represents an intermediate case. This suggests that the mixture of being scale-free and random could have emerged as a sort of compromise between two behaviors that have their own benefits under different circumstances.

From the point of view of a living organism, a scale-free proteome represents an advantage in the short term, but over long evolutionary time scales some randomness wiring could also be beneficial [21]. The curved degree distribution observed in all experimental protein interaction networks derived to the date strongly suggests that a mixture of scale invariance and randomness could be more appropriate than a perfect scale-free network in the long term and a purely random network in the short term. Furthermore, this distribution can be obtained in a very robust manner from a simplest model of duplication and divergence of genes, which has a clear biological basis. In the divergence process (rewiring of nodes) there is some memory of the copied node (that conveys a deterministic preferential attachment rule) and some randomness in the rewiring. The balance between these two processes could have emerged in the evolutionary history as a product of natural selection over the whole protein network.

In this work, we assess the reliability of a simple duplication model able to describe the degree distribution observed in the yeast protein interaction network. We scrutinized the claimed scale-free behavior of this distribution and pointed out a severe departure in both the duplication model and experimental data. The emerging distribution was explained as an interplay between a preferential attachment rule and a purely random process. Even when we cannot draw general conclusions for the protein interaction network from the study of a single species, our prelimany studies of the network robustness under realistic perturbations suggests that the interplay between deterministic evolutionary memory (from the duplication process) and pure randomness make possible a better environmental response for a biological network than a purely random or deterministic preferential attachment underlying mechanism.

ACKNOWLEDGMENTS

The authors thank J. Echave for useful discussions. This work was partially funded by UNQ and ANPCyT, Argentina.

[1] R. Albert and A-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] A. Wagner, Proc. R. Soc. London, Ser. B **270**, 457 (2003).
[3] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).
[4] B. Papp, C. Pal, and L. D. Hurst, Nature (London) **429**, 661 (2004).
[5] M. W. Hahn, G. C. Conant, and A. Wagner, J. Mol. Evol. **58**, 203 (2004); H. B. Fraser, D. P. Wall and A. E. Hirsh, BMC Evol. Biol. **3**, 11 (2003); I. K. Jordan, Y. I. Wolf, and E. V. Koonin, *ibid.* **3**, 5 (2003).
[6] E. Eisenberg and E. Y. Levanon, Phys. Rev. Lett. **91**, 138701 (2003).
[7] P. Uetz *et al.*, Nature (London) **403**, 623 (2000); T. Ito *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 4569 (2001).
[8] Y. Ho *et al.*, Nature (London) **415**, 180 (2002); A. C. Gavin *et al.*, *ibid.* **415**, 141 (2002).
[9] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant, Nat. Biotechnol. **22**, 78 (2004).
[10] R. V. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler, Adv. Complex Syst. **5**, 43 (2002); A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, ComPlexUs **1**, 38 (2003); R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science* (IEEE, New York, 2000), pp. 57–65.
[11] J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner, Phys. Rev. E **66**, 055101(R) (2002).
[12] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas, J. Comput. Biol. **10**, 677 (2003).
[13] 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).
[14] A. Vázquez, Phys. Rev. E **67**, 056104 (2003).
[15] I. Xenarios *et al.*, Nucleic Acids Res. **30**, 303 (2002).
[16] S. Li *et al.*, Science **303**, 540 (2004).
[17] L. Giot *et al.*, Science **302**, 1727 (2003).
[18] R. Albert, H. Jeong and A.-L. Barabási, Nature (London) **406**, 378 (2000).
[19] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[20] V. Latora and M. Marchiori, Phys. Rev. Lett. **87**, 198701 (2001).
[21] V. Venkatasubramanian, S. Katare, P. R. Patkar, and F.-p. Mu, Comput. Chem. Eng. **28**, 1789 (2004).