# Truncation of Power Law Behavior in "Scale-Free" Network Models due to Information Filtering

Stefano Mossa,[1,2] Marc Barthélémy,[3] H. Eugene Stanley,[1] and Luís A. Nunes Amaral[1]

[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*Dipartimento di Fisica, INFM UdR, and INFM Center for Statistical Mechanics and Complexity, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, I-00185, Roma, Italy*
[3]*CEA-Service de Physique de la Matière Condensée, BP 12, 91680 Bruyères-le-Châtel, France*

(Received 18 October 2001; published 14 March 2002)

We formulate a general model for the growth of scale-free networks under filtering information conditions—that is, when the nodes can process information about only a subset of the existing nodes in the network. We find that the distribution of the number of incoming links to a node follows a universal scaling form, i.e., that it decays as a power law with an exponential truncation controlled not only by the system size but also by a feature not previously considered, the subset of the network "accessible" to the node. We test our model with empirical data for the World Wide Web and find agreement.

DOI: 10.1103/PhysRevLett.88.138701 PACS numbers: 89.20.Hh, 84.35.+i, 89.75.Da, 89.75.Hc

There is a great deal of current interest in understanding the structure and growth mechanisms of global networks [1–3], such as the World Wide Web (WWW) [4,5] and the Internet [6]. Network structure is critical in many contexts such as Internet attacks [2], spread of an Email virus [7], or dynamics of human epidemics [8]. In all these problems, the nodes with the largest number of links play an important role on the dynamics of the system. It is therefore important to know the global structure of the network as well as its precise distribution of the number of links.

Recent empirical studies report that both the Internet and the WWW have scale-free properties; that is, the number of incoming links and the number of outgoing links at a given node have distributions that decay with power law tails [4–6]. It has been proposed [9] that the scale-free structure of the Internet and the WWW may be explained by a mechanism referred to as "preferential attachment" [10] in which new nodes link to existing nodes with a probability proportional to the number of existing links to these nodes. Here we focus on the *stochastic* character of the preferential attachment mechanism, which we understand in the following way: New nodes want to connect to the existing nodes with the largest number of links—i.e., with the largest degree—because of the advantages offered by being linked to a well-connected node. For a large network it is not plausible that a new node will know the degrees of all existing nodes, so a new node must make a decision on which node to connect with based on what information it has about the state of the network. The preferential attachment mechanism then comes into play as nodes with a larger degree are more likely to become known.

This picture has one underlying and unstated assumption, that the new nodes will process (i.e., gather, store, retrieve, and analyze) information concerning the state of the entire network. For very large networks, such as the WWW or the scientific literature, this would correspond to the unrealistic situation in which new nodes can process an extremely large amount of information—i.e., have unlimited information-processing capabilities. Indeed, it is likely that nodes have limited information-processing capabilities and so must filter incoming information according to their particular "interests." Thus, new nodes of a large growing network will process only information concerning a subset of existing nodes, since there is a cost associated with processing information. The new nodes will then make decisions on with whom to link, based on *filtered* information. From the standpoint proposed here, most models studied in the literature work under the unrealistic assumption of *unfiltered* information—i.e., a new node processes information about all the existing nodes in the network.

Here we consider for the first time the effect on network growth of filtering information due to limited information-processing capabilities. First, we calculate the in-degree distributions of web pages using two databases. The first database, which comprises $\approx 2 \times 10^8$ pages [9], surveys a very significant fraction of the entire WWW, while the second, which comprises $\approx 3 \times 10^5$ pages, lists the University of Notre Dame domain [4]—i.e., the set of URLs containing the string "nd.edu." For the first database, we calculate the cumulative in-degree distributions $P(k) = \sum_{k'>k} p(k')$, where $p(k)$ is the probability distribution. We confirm that the in-degree distribution decays as a power law [9] of the form

$$P(k) \sim k^{-\gamma_{\rm in}}, \qquad (1)$$

with an exponent $\gamma_{\rm in} = 1.25 \pm 0.05$ (Fig. 1). Further, we find an exponential truncation of the scale-free behavior for $k > k_\times \approx 2 \times 10^5$, in contrast with the plateau reported in other studies [2,11]. For the second database, we also find a power law regime with the same exponent, but the exponential truncation appears to be absent, suggesting that the truncation is not due to the finite size of the databases.

 0031-9007/02/88(13)/138701(4)$20.00  

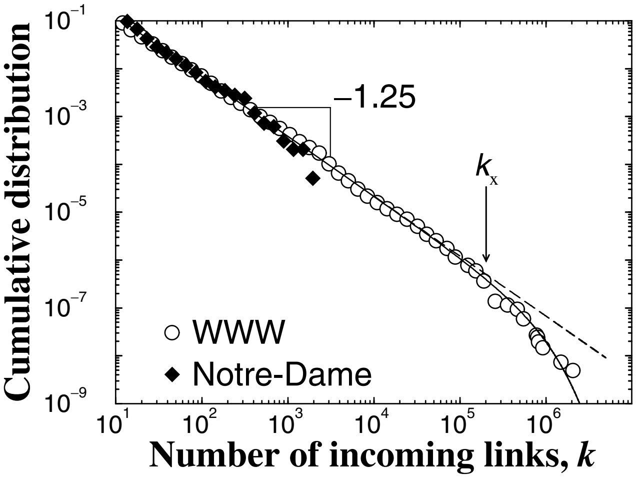


FIG. 1. Distribution of the number of incoming links for the WWW. Cumulative in-degree distribution from two databases, the entire Web [9], and the University of Notre Dame domain [4]. We also plot a power law function with exponent $\gamma_{\text{in}} = 1.25$ (dashed line) and a Yule function [10] of the form $k^{-\gamma_{\text{in}}} \exp(-k/k_x)$ (solid line). A cutoff degree $k_\times \simeq 200\,000$ is visible in the data.

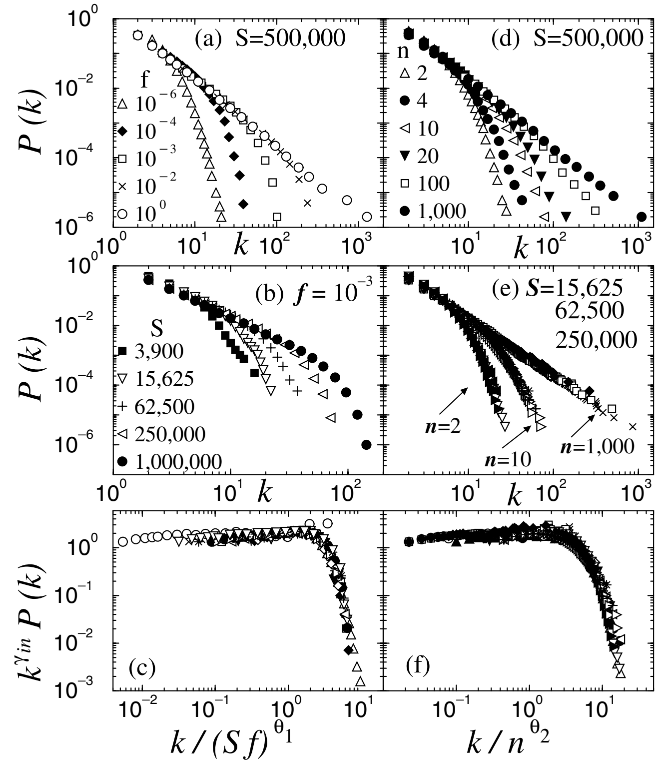


FIG. 2. In-degree cumulative probability distributions $P(k)$ under information filtering. Constant $f$ case: (a) Results for $S = 5 \times 10^5$ and different values of $f$. (b) Results for $f = 10^{-3}$ and different values of $S$. (a) and (b) show that $k_\times$ decreases with $f$ and increases with $S$. (c) Data collapse of the numerical results according to Eq. (3) with $\gamma_{\text{in}} = 1.97 \pm 0.05$ and $\theta_1 = 0.45 \pm 0.04$. Constant $n$ case: (d) Results for $S = 5 \times 10^5$ and different values of $n$ showing the decrease in the cutoff degree $k_\times$ with decreasing $n$. (e) Results for $n = 2, 10$, and 1000 for different values of $S$ showing that $P(k)$ does not depend on $S$. (f) Data collapse according to Eq. (4) with $\gamma_{\text{in}} = 2.00 \pm 0.03$ and $\theta_2 = 0.65 \pm 0.04$.

To explain these empirical results, we hypothesize that the authors of new web pages filter some of the information regarding existing web pages, that is, the new nodes make linking decisions under information-filtering conditions. To investigate this process, we consider network growth models in which new nodes process information from only a fraction of existing nodes which one may view as matching the interests of the new nodes. If the fraction $f$ of "interesting" nodes in the network is much less than one, then the attachment of new links is a random process, so the generated network will be a random graph with an exponentially decaying in-degree distribution. In contrast, if $f \approx 1$, then preferential attachment is recovered and the in-degree distribution is scale-free.

We first define the network growth rule: At time $t = 0$, one creates $n_o$ nodes with $n_o - 1$ links each. At each time step, one adds to the network a new node with $n_o - 1$ outgoing links. These $n_o$ links can connect to a randomly selected subset $C$ containing $n(t) = (t + n_o)f$ nodes. The links to the nodes in the subset are selected according to the preferential attachment rule, i.e., the probability that node $i$ belonging to $C$ is selected is proportional to the number of incoming links $k(i)$ to it

$$p(i,t) \equiv \frac{k(i)}{\sum_{j \epsilon C} k(j)}. \tag{2}$$

In Fig. 2(a), we show our numerical results for the in-degree cumulative distributions for networks with $S = 5 \times 10^5$ nodes and $n_o = 1$, for a sequence of $f$ values. For $f = 1$, we reproduce the results reported for the scale-free model [9]—i.e., we observe an in-degree distribution that decays as a power law with an exponent $\gamma_{\text{in}} \approx 2$. For $f < 10^{-2}$, we observe a crossover at $k = k_\times$ from power law behavior to exponential behavior.

To further investigate the effect of changes in $f$ on the cutoff degree $k_\times$, we plot in Fig. 2(b) the in-degree distributions for different network sizes $S$ and a fixed value of $f$. We find that $k_\times$ increases as a power law with $S$. All of our numerical results can be expressed compactly by the scaling form

$$P(k,f,S) \propto k^{-\gamma_{\text{in}}} \mathcal{F}_1\left(\frac{k}{k_\times}\right), \tag{3}$$

with $k_\times \sim (Sf)^{\theta_1}$. We find $\gamma_{\text{in}} = 1.97 \pm 0.05$, $\theta_1 = 0.45 \pm 0.04$ and $\mathcal{F}_1(x) \sim$ const for $x \ll 1$, $\mathcal{F}_1(x) \sim e^{-x}$ for $x \gg 1$. As a test of the scaling form Eq. (3), we plot in Fig. 1(c) the scaled cumulative distribution versus the scaled in-degree. The figure confirms our scaling ansatz, since all data "collapse" onto a single curve, the scaling function $\mathcal{F}_1(x)$.

We consider next a situation in which new nodes are not processing information from a constant fraction $f$ of nodes but from a constant number $n$ of nodes. That is, as the network grows, the new nodes are able to process information about a smaller fraction of existing nodes. This model may be more plausible for networks that have grown to a very large size, since the fraction $f$ of all nodes represents a very large number. In the case of the scientific literature, this effect leads to the fragmentation of a scientific field as it grows [12].

For the constant $n$ case, the fraction of known nodes at time $t$ is $f(t) = n/(t + n_o)$, implying that as the network grows there are two antagonistic trends affecting $k_\times$. The first is a tendency to increase due to the growing size of the network, and the second is a tendency to decrease due to the decreasing value of $f$. Hence, one may hypothesize that there will be a characteristic network size $S_c$ above which $k_\times$ will no longer depend on $S$.

We now test these arguments with numerical simulations. In Figs. 2(d) and 2(e), we show our results for growing networks for which new nodes process information only from $n$ randomly selected existing nodes. We find, in agreement with our scaling arguments, that for $S \gg S_c$ the in-degree distribution obeys the scaling relation

$$P(k, n, S) \propto k^{-\gamma_{\rm in}} \mathcal{F}_2\left(\frac{k}{k_\times}\right), \tag{4}$$

with $k_\times \sim n^{\theta_2}$, $\gamma_{\rm in} = 2.00 \pm 0.03$, $\theta_2 = 0.65 \pm 0.04$, and where the scaling function $\mathcal{F}_2(x)$ has the same limiting behavior as $\mathcal{F}_1(x)$. To test the scaling form Eq. (4), we plot in Fig. 2(f) the scaled cumulative distribution versus the scaled in-degree. This confirms our scaling ansatz since the data collapse onto a single curve, the scaling function $\mathcal{F}_2(x)$.

Comparison of the two scaling relations Eqs. (3) and (4) reveals an unexpected result. By replacing $Sf$ by $n$ in (3) one would naively expect to obtain (4) with $\theta_1 = \theta_2$ and $\mathcal{F}_1(x) = \mathcal{F}_2(x)$. Surprisingly, we find that $\theta_1$ is significantly different from $\theta_2$ and that $\mathcal{F}_1(x)$ is significantly different from $\mathcal{F}_2(x)$. In order to understand this result, consider two growing networks that have reached size $S$. For the first, new nodes process information from a fraction $f$ of existing nodes, while, for the second, new nodes process information from $n = fS$ existing nodes. At a time $t$, prior to the network having reached its final size $S$, there are $t + n_o < S$ sites, and the preferential attachment is acting for the first network on a number of nodes $(t + n_o)f < Sf = n$. The preferential attachment mechanism can operate effectively only when it acts on a number of nodes comparable to $S$, so the fact that for the first network new nodes have always processed information from *fewer* existing nodes suggests the first network will not develop nodes with as large a degree as the second network. Thus, we expect that (i) the two resulting networks have different in-degree distributions, and (ii) the in-degree distribution for $f$ fixed has a sharper truncation and a smaller cutoff than for $n$ fixed, which is indeed what we find.

Our numerical results are in qualitative agreement with empirical data. However, the value of the power law exponent $\gamma_{\rm in} \approx 1.25$ found for the WWW is significantly smaller than the value $\gamma_{\rm in} = 2$ predicted by the model. This fact prompts the question of the effect of the cost of information filtering on models generating an in-degree distribution closer to the empirical results. To answer this question, we investigate two possible explanations for the observed value $\gamma_{\rm in} \approx 1.25$.

*(i) Effect of out-degree distribution on* $\gamma_{\rm in}$.—The scale-free model [9] is missing an important ingredient: a heterogeneous distribution of number of outgoing links. Indeed, the out-degree distribution considered so far is restricted to a single value $m = n_o - 1$, i.e., $p_{\rm out}(m) = \delta_{m,n_o-1}$, while for the empirical data of the WWW it decays as a power law of the form $p_{\rm out}(m) \sim m^{-\gamma_{\rm out}}$ with $\gamma_{\rm out} = 1.68 \pm 0.05$. We show in Fig. 3 the computed value of the exponent $\gamma_{\rm in}$ of the in-degree distribution as a function of $\gamma_{\rm out}$ [13]. We find that $\gamma_{\rm in}$ increases approximately linearly with increasing values of the exponent $\gamma_{\rm out}$ until it reaches the limiting value $\gamma_{\rm in} = 2$. For $\gamma_{\rm out} \approx 1.7$, which is the empirically observed value for the WWW, we find $\gamma_{\rm in} \approx 1.8$, which does not agree with the empirical value of 1.25, so the power law decaying out-degree distribution alone cannot explain the results obtained for the WWW.

*(ii) Effect of fitness on* $\gamma_{\rm in}$.—The preferential attachment mechanism is modified by a "fitness" factor [11]: Nodes have different fitness, and fitter nodes are more likely to receive incoming links than less fit nodes with the same value of $k$. Uniformly distributed fitness is known to lead to a smaller exponent $\gamma_{\rm in} = 1.255$ [11], which is quite close to the value measured for the WWW. Hence, we assign to each node a fitness $\eta(i)$ [11], reflecting the fact that for equal values of $k$ some nodes are more "attractive" than others [14]. The probability that a new node will link to node $i$ is

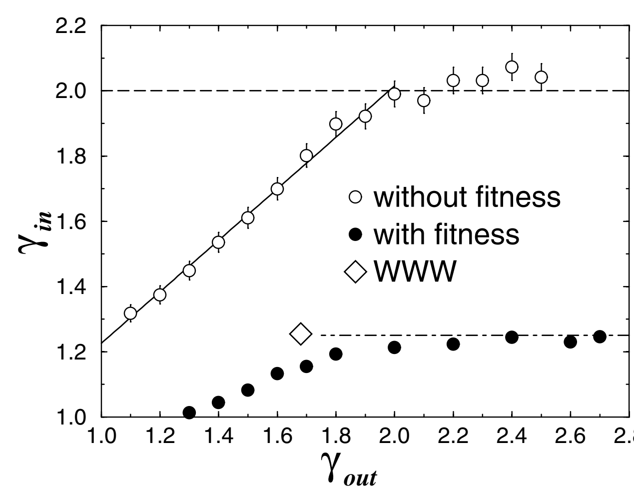


FIG. 3. Dependence of the in-degree distribution exponent $\gamma_{\rm in}$ on the out-degree distribution exponent $\gamma_{\rm out}$. We show results for models (i) without fitness [$\eta(i) = $ const] and (ii) with fitness [$\eta(i)$ uniformly distributed]. For the former case, $\gamma_{\rm in}$ increases initially approximately linearly with $\gamma_{\rm out}$, and then saturates at $\gamma_{\rm in} \approx 2$ for $\gamma_{\rm out} > 2$. This saturation of $\gamma_{\rm in}$ is to be expected as $\gamma_{\rm in} = 2$ for the case of a peaked distribution of $n_o$. For the latter case, $\gamma_{\rm in}$ increases approximately linearly with $\gamma_{\rm out}$ initially, and then saturates at $\gamma_{\rm in} \approx 1.25$ for $\gamma_{\rm out} > 1.9$. This saturation is to be expected as $\gamma_{\rm in} = 1.255$ for the case of a peaked distribution of $n_o$ [11].

$$p(i,t) \equiv \frac{\eta(i)k(i)}{\sum_{j=1}^{t+n_o} \eta(j)k(j)} \,. \tag{5}$$

We consider here the case in which $\eta(i)$ is a uniformly distributed random variable [15]. Figure 3 shows that the in-degree distribution decays as a power law with values of $\gamma_{\rm in} < 1.25$. For $\gamma_{\rm out} > 1.9$, the exponent approaches the limiting value $\gamma_{\rm in} \approx 1.25$. Interestingly, for $\gamma_{\rm out} \approx 1.7$, the empirical value for the WWW, we find $\gamma_{\rm in} \approx 1.2$, in agreement with the empirical value $\gamma_{\rm in} \approx 1.25$.

Our results for the model with fitness show that information filtering and node fitness are both necessary in order to approximate the empirical results. An open question is which type of filtering is more appropriate for the WWW, constant $f$ or constant $n$? To answer this question one would need WWW data for a different sample size, which are not presently available to us. However, due to the sheer size of the WWW, it seems plausible that constant $n$ would be the more appropriate case.

Our key finding is that limited information-processing capabilities have a significant and quantifiable effect on the large-scale structure of growing networks. We find that information filtering leads to an exponential truncation of the in-degree distribution for networks growing under conditions of preferential attachment. Surprisingly, we find simple scaling relations that predict the in-degree distribution in terms of (i) the information-processing capabilities available to the nodes, and (ii) the size of the network.

We also quantify the effect of a heterogeneous out-degree distribution on the in-degree distribution of networks growing under conditions of preferential attachment. We find that, for a power law decaying out-degree distribution with exponents $\gamma_{\rm out} < 2$, the exponent $\gamma_{\rm in}$ characterizing the tail of the in-degree distribution will take values smaller than those predicted by theoretical calculations [2,3].

The exponential truncation we find may have dramatic effects on the dynamics of the system, especially for processes where the nodes with the largest degree have important roles. This is the case, for example, for virus spreading [7], where for networks with exponentially truncated in-degree distributions there is a nonzero threshold for the appearance of an epidemic. In contrast, scale-free networks are prone to the spreading and the persistence of infections no matter how small the spreading rate. Our finding of a mechanism leading to an exponential truncation even for systems where before none was expected [16] indicates that the most connected nodes will have a smaller degree than predicted for scale-free networks leading, possibly, to different dynamics, e.g., for the initiation and spread of epidemics.

In the context of network growth, the impossibility of knowing the degrees of all the nodes comprising the network due to the filtering process—and, hence, the inability to make the optimal, rational, choice—is not altogether unlike the "bounded rationality" concept of Simon [17]. Remarkably, it appears that, for the description of WWW growth, the preferential attachment mechanism, originally proposed by Simon [10], must be modified along the lines of another concept also introduced by him—bounded rationality [17].

We thank R. Albert, P. Ball, A.-L. Barabási, M. Buchanan, J. Camacho, and R. Guimerà for stimulating discussions and helpful suggestions. We are especially grateful to R. Kumar for sharing his data. We thank NIH/NCRR (P41 RR13622) and NSF for support.

[1] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. (to be published).
[4] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[5] B. A. Huberman and L. A. Adamic, Nature (London) **401**, 131 (1999); R. Kumar *et al.,* in *Proceedings of the 25th International Conference on Very Large Databases* (Morgan Kaufmann Publishers, San Francisco, 1999), p. 639; A. Broder *et al.,* Comput. Netw. **33**, 309 (2000); P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000); S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, *ibid.* **85**, 4633 (2000); A. Vazquez, Europhys. Lett. **54**, 430 (2001).
[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999); G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000); A. Medina, I. Matta, and J. Byers, Comput. Commun. Rev. **30**, 18 (2000); R. Pastor-Satorras, A. Vazquez, and A. Vespignani, arXiv:cond-mat/0105161; L. A. Adamic *et al.,* Phys. Rev. E **64**, 046135 (2001).
[7] F. B. Cohen, *A Short Course on Computer Viruses* (Wiley, New York, 1994); R. Pastor-Satorras and A. Vespignagni, Phys. Rev. Lett. **86**, 3200 (2001); Phys. Rev. E **63**, 066117 (2001).
[8] F. Liljeros, C. R. Edling, L. A. Nunes Amaral, H. E. Stanley, and Y. Åberg, Nature (London) **411**, 907 (2001).
[9] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[10] Y. Ijiri and H. A. Simon, *Skew Distributions and the Sizes of Business Firms* (North-Holland, Amsterdam, 1977).
[11] G. Bianconi and A.-L. Barabasi, Europhys. Lett. **54**, 436 (2001).
[12] A. F. J. Van Raan, Scientometrics **47**, 347 (2000).
[13] We consider a modification to the network growth rule described earlier in the paper: at each time step $t$, the new node establishes $m$ new links, where $m$ is drawn from a power law distribution with exponent $\gamma_{\rm out}$.
[14] For $\eta(i) =$ const, one recovers the scale-free model of Ref. [9].
[15] It is known [11] that, for an exponential or fat-tailed distribution of fitness, the structure of the network becomes much more complex; in particular, the in-degree distribution is no longer a power law. Hence, we do not consider in this manuscript other shapes of the fitness distribution.
[16] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11 149 (2000).
[17] H. A. Simon, *Models of Bounded Rationality: Empirically Grounded Economic Reason* (MIT Press, Cambridge, 1997).