# Quantitative description and modeling of real networks

Andrea Capocci,[1] Guido Caldarelli,[2] and Paolo De Los Rios[3]
[1]*Département de Physique, Université de Fribourg, CH-1700 Fribourg, Switzerland*
[2]*INFM UdR di ROMA1, Dipartimento di Fisica, Università di Roma "La Sapienza," Piazzale A. Moro 2, 00185 Roma, Italy*
[3]*Institut de Physique Théorique, Université de Lausanne, 1015 Lausanne, Switzerland*
(Received 29 May 2002; published 2 October 2003)

We present data analysis and modeling of two particular cases of study in the field of growing networks. We analyze World Wide Web data set and authorship collaboration networks in order to check the presence of correlation in the data. The results are reproduced with good agreement through a suitable modification of the standard Albert-Barabási model of network growth. In particular, intrinsic relevance of sites plays a role in determining the future degree of the vertex.

DOI: 10.1103/PhysRevE.68.047101 PACS number(s): 89.75.Hc, 05.10.−a, 89.20.Hh, 89.65.Ef

The fractal properties of social networks have been largely investigated by the statistical mechanics community in recent times. Many quantities have been recognized as "signatures" of complexity in such networks. In particular, the probability distribution of the degree of the nodes in a social network displays an algebraic decay in several different realizations, including the Internet, the World Wide Web (WWW), the movie actors network, and the science collaboration network [1–4]. Since then, many models have been developed in order to reproduce this particular feature of real networks [3,5].

Properties beyond the degree distribution have also been analyzed: In the internet autonomous systems (IAS) network, the relation between the degree of a node $k$ and the average degree of its neighbors $k_{nn}(k)$ has been measured, showing a decaying behavior of $k_{nn}(k)$ for large $k$; such property is connected to a hierarchical structure of the growth process [6,7]. From a more general point of view, it has been shown that a taxonomy of social networks can be made according to the correlation between the degrees of directly connected nodes [9]. In networks displaying "assortative (disassortative) mixing," the correlation is positive (negative), which corresponds to an increasing (a decreasing) behavior of $k_{nn}(k)$.

Furthermore, a growing number of studies have investigated the clustering properties of social networks, that is, the presence and the abundance of groups of nodes having a strong internal connectivity. In particular, the monitored quantity for a node of degree $k$ is the clustering coefficient $c_k$ that is the average number of edges between nearest neighbors of a node of degree $k$ normalized with respect to the largest possible number of such links, $k(k-1)/2$ (the average is taken over the whole network). With this definition, the clustering coefficient is a number between 0 and 1. Recent surveys on IAS [6,7] have measured the clustering coefficient $c_k$ around nodes of degree $k$. These empirical studies show a decaying behavior of $c_k$ with respect to $k$, as in the case of $k_{nn}(k)$. In order to analyze directed graphs, such as the WWW (where the edges of the network are the hyperlinks that are clearly not undirected), different recipes can be applied. In principle one could explicitly consider the difference between in-going and out-going links. More simply, the clustering coefficient for directed networks is often computed considering each link as if it was undirected, and possible double edges (two nodes could have mutual hyperlinks) are considered as a single undirected link as well. We have adopted this technique to measure both $k_{nn}(k)$ and $c_k$ for a snapshot of the WWW and for the network of scientific collaborations taken from an online database [8], finding, for the WWW, qualitatively the same results as in the IAS (undirected) case studied in Refs. [6,7]. Using standard noise reducing data analysis techniques we find that $k_{nn}(k) \sim k^{-0.76}$ for large $k$, as shown in Fig. 1, and $c_k \sim k^{-1.03}$, see Fig. 2. This behavior is in good agreement with the power laws found in the IAS case [6,7], though the exponents are slightly different [their measurements, which are affected by a weaker noise, yield $k_{nn}(k) \sim k^{-0.5}$ and $c_k \sim k^{-0.75}$]. The IAS and WWW are therefore examples of scale-free networks showing disassortative mixing. The basic model used to describe scale-free networks, namely, the Albert-Barabási model [3], which is based on the "preferential attachment" mechanism, gives networks essentially devoid of correlations: $k_{nn}(k)$ and $c_k$ are roughly independent of $k$, quite differently from what is shown in Refs. [6,7] and again in this work. According to the preferential attachment mechanism, the probability to draw an edge between a new node and an already existing one is proportional to the degree of the lat-

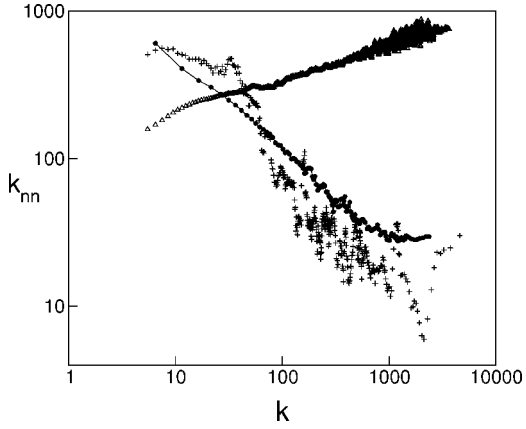


FIG. 1. Average degree $k_{nn}(k)$ of nearest neighbors of a node with degree $k$, as a function of $k$. Triangles refer to the actor collaboration network, plus symbols refer to the WWW empirical survey (10-points averaged), and circles to simulations of our model.

 

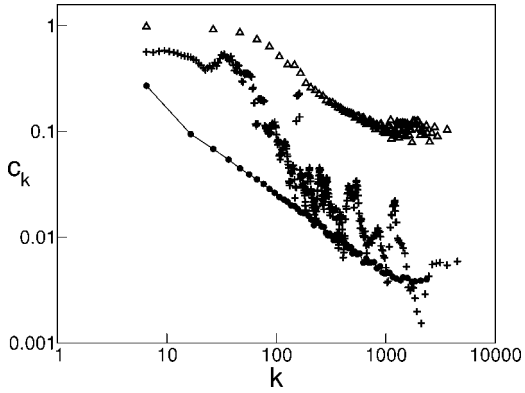


FIG. 2. Clustering coefficient $c_k$ around a node of degree $k$ as a function of $k$. Circles refer to the actor collaboration network, plus symbols refer to the WWW empirical survey (10-points averaged), and squares to the simulation of our model.

ter. The information used to decide about connections is therefore purely topological. In real networks, rather, the choice to establish a connection depends also on some other properties of the involved nodes. For example, social networks showing assortative mixing typically describe systems where a mutual consent to establish the link is needed, like in author collaboration networks and actor networks (where the further external intervention of the actors' agents and of the film producer is important). For the WWW, which shows disassortative mixing, we expect that the probability that a newcomer node connects to an older one does not depend only on the degree of the latter, but also on its intrinsic qualities.

To check if our hypothesis is true, we introduce a growing undirected network model. Sites are added at a discrete pace, and each site has an intrinsic ''relevance,'' which is a random variable drawn from a uniform distribution in the range $[0,1]$ (models where intrinsic node variables determine the structure of a network have recently been introduced [10]). In our interpretation, a link is the relevance attributed to the pointed node, in the spirit of Refs. [11–13]. In the WWW, for example, a relevant web page rarely points to a nonrelevant one, suggesting a relevance-driven connectivity concentration. To implement such a policy, in our model a node added at time $t$ with a relevance $r_t$ can be connected only to nodes having a relevance higher than $r_t$, with linear preferential attachment: the probability of acquiring a new link is proportional to the actual degree.

This implies that an existing node $i$ with a relevance $r_i$ and degree $k_i$ has a probability $p_i=\Theta(r_i-r_t)k_i/[\Sigma_{s=1}^{t}k_s\Theta(r_s-r_t)]$ of acquiring a new link, where $\Theta(x)=1$ for $x>0$ and $\Theta(x)=0$ otherwise. Finally, we assume that a newly added node is connected to $m$ existing nodes according to the described rule.

Let us call $k_i(t)$ the degree at time $t$ of the node $i$ introduced at time $i$, whose relevance is $r_i$. At each time step, there is a probability $r_i$ that the newly introduced node has a relevance $r_t<r_i$, since $r_t$ is drawn from a uniform distribution between 0 and 1. Then, the probability of increasing by 1 the degree $k_i(r_i,t)$ is approximately given by

$$\langle p_i\rangle_{r_t}\simeq\frac{r_i k_i(t)}{\left\langle\sum_{s:r_s>r_t}^{1,t-1}k_s(t)\right\rangle_{r_t}}, \tag{1}$$

where $\langle\cdots\rangle_{r_t}$ denotes the average over all the realizations of $r_t$. In the following, we will neglect the explicit time dependence whenever unnecessary. We can write a rate equation for the degree, following the reasoning made in Ref. [14]:

$$\dot{k}_i=\frac{m r_i k_i(t)}{\left\langle\sum_{s:r_s>r_t}^{1,t-1}k_s(t)\right\rangle_{r_t}}. \tag{2}$$

To evaluate the denominator in the right-hand side of the above equation, we have to compute

$$\left\langle\sum_{s:r_s>r_t}^{1,t-1}k_s(t)\right\rangle_{r_t}=\int_0^{r_i}dr_t A(r_t,t), \tag{3}$$

where we defined the decreasing funtion of $r_t$:

$$A(r_t,t)=\sum_{s:r_s>r_t}^{1,t-1}k_s(t). \tag{4}$$

Since $A(0,t)=2mt$ and $A(1,t)=0$, we assume the ansatz $A(r_t,t)=2mt(1-r_t^{\alpha})$ for the functional form of $A(r_t,t)$, for $t$ large enough. By this ansatz, we can compute the right-hand side of Eq. (3),

$$\int_0^{r_i}dr_t A(r_t,t)=2mtr_i C(r_i), \tag{5}$$

where

$$C(r)=1-\frac{r^{\alpha}}{1+\alpha}. \tag{6}$$

Therefore, the solution of the rate Eq. (2) is

$$k_i(t)=m\left(\frac{t}{i}\right)^{r_i/[2C(r_i)]} \tag{7}$$

for the time evolution of the degree, following the same reasoning as in Ref. [3].

Let us now call $K(r,t)dr$ the sum of the degrees of the nodes with relevance between $r$ and $r+dr$, at time $t$. At each time step, $dr$ nodes on average are introduced with such a relevance. Equation (7) gives us the degree acquired by each of these nodes. To obtain $K(r,t)$ we have to sum over all time steps from 1 to $t$, and we get

$$drK(r,t)=dr\sum_{s=1}^{t}k_s(t). \tag{8}$$

Approximating the sum by an integral, and replacing $k_s(t)$ by Eq. (7), we get

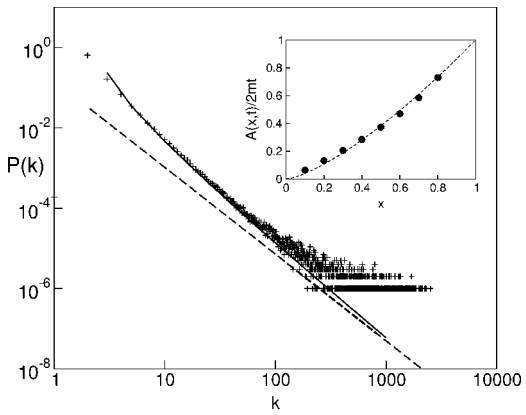


FIG. 3. Degree of PDF in our network model made of $10^4$ nodes, with $m=2$. Plus symbols refer to numerical simulation. The solid line is obtained by plotting Eq. (11). The dashed line is proportional to $k^{-2.16}$. Inset: the function $[A(x,t)]/(2mt)$ plotted for $t=10^4$ and $m=2$. The dashed line represents $x^{1.38}$, displayed here to check the validity of our ansatz.

$$K(r,t)=\frac{mt}{1-\dfrac{r}{2C(r_s)}}. \tag{9}$$

We can estimate $\alpha$ by integrating $K(r,t)$ over all $r$, thus obtaining the total sum of the nodes' degrees:

$$\int_0^1 drK(r,t)=2mt. \tag{10}$$

This equation can be numerically shown to yield $\alpha=1.3837$. The ansatz on $A(r_t,t)$ is verified in simulations of the model, as shown in Fig. 3

Following Ref. [14], the time evolution of the individual degrees allows us to compute their statistical distribution $P(k)$; we obtain

$$P(k)=\frac{1}{k}\int_0^1 dr\left(\frac{k}{2}\right)^{-B(r)}B(r), \tag{11}$$

where $B(r)=(2/r)[1-(r^\alpha/1+\alpha)]$, which displays a power-law behavior for large $k$. We can estimate the power-law exponent of the degree distribution $P(k)$ finding upper and lower bounds for its integral expression. Indeed, we find that

$$F(r)=e^{-2\ln(k/2)B(r)}B(r) \tag{12}$$

is such that the integrand is monotonically growing. Therefore it is easily seen that

$$P(k)<\frac{1}{k}e^{-2\ln(k/2)B(1)}B(1)\sim k^{-(3\alpha+1)/(\alpha+1)}. \tag{13}$$

As for the lower bound, we first observe that the integrand is monotonically increasing, with positive second derivative. So,

$$F_1(r)=F(1)-F'(1)(1-r) \tag{14}$$

is such that $F_1(r)<F(r)$ for $0\leqslant r\leqslant 1$. Then if we extend the integral from $r_1$ $[F_1(r_1)=0]$ to 1, we surely find an underestimation for $P(k)$. In particular we find

$$P(k)>k^{-(3\alpha+1)/(\alpha+1)}\frac{1}{\dfrac{2\alpha}{\alpha+1}\ln(k/2)-1}. \tag{15}$$

The asymptotic behavior of $P(k)$ is therefore $k^{-(3\alpha+1)(\alpha+1)}$ with at most logarithmic corrections.

We numerically checked that $P(k)$ is a power law with a rather weak correction that slows down the decay, as displayed in Fig. 3. Neglecting the correction, the best approximating exponent of the probability density function (PDF) is about $-2.16$, which confirms the above computation. Indeed, we have $(3\alpha+1)/(\alpha+1)=2.16$. This value, moreover, is close to the exponents measured in real networks, which lie in the range 2–2.4.

In the simulation of the model, $k_{nn}(k)$ and $c_k$ have also been numerically investigated. Unfortunately, we could not find an analytical description of these two quantities. As required by real data, $k_{nn}(k)$ and $c_k$ decay algebraically with respect to $k$. For the nearest-neighbors degree, we approximately measured $k_{nn}(k)\simeq k^{-0.57}$, as shown in Fig. 1. The value of the exponent agrees with the measurement reported in Refs. [6,7], which yields $k_{nn}(k)\simeq k^{-\nu_k}$ with $\nu_k=0.5\pm 0.1$. As for the clustering coefficient $c_k$, simulations reported in Fig. 2 show that $c_k\simeq k^{-0.72}$. The same relation, measured in Refs. [6,7], in the IAS networks case, reads $c_k\simeq k^{-\omega}$ with $\omega=0.75\pm 0.03$.

The qualitative behavior of these quantities is reproduced in our extremely simple model. As a comparison, let us recall that, without an intrinsic relevance, a simple growing network model with preferential attachment shows no correlation between the degrees of two linked nodes. In addition, in this model the clustering coefficient around a node does not depend on the degree of the node [7,9]. An improvement in approximating real data could be achieved by adding other microscopic interactions to the dynamics of our toy model, such as rewiring and elimination and links, or by merging nodes, as already done in former works [15–17] in the search for a better approximation of the scale-free degree distribution.

We believe that our analysis has pointed out some key structural features of social networks, by the observation of the correlation and the clustering of the connectivity in networks. In particular, the nontrivial behavior of the nearest-neighbor average degree and of the connectivity coefficient have been measured in some real examples. We also provided a toy model of a growing network with preferential attachment, where nodes only connect to more relevant ones. We have shown numerically and analytically, as far as we could, that our model reproduces to a good approximation the statistical properties of real networks, including the correlations in the connectivity. We believe that this approach suggests new empirical measurements to be carried out on real networks, and that new ingredients and further analytical steps are needed toward the comprehension of these complex systems.

The authors wish to thank M.E.J. Newman, A.-L. Barabasi, R. Albert, and H. Jeong for providing free-access data about networks. They acknowledge support from EC-Fet Open Project No. COSIN IST-2001-33555, and the OFES-Bern(CH). G.C. and P.D.L.R. finally wish to acknowledge the Herbette Foundation for financial support.

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[2] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).
[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[4] M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[5] B.A. Huberman and L.A. Adamic, Nature (London) **401**, 131 (1999).
[6] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **88**, 108701 (2002).
[7] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[8] Raw data about the WWW graph and the actors collaboration network are available at http://www.nd.edu/~networks
[9] M.E.J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[10] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[11] S. Brin and L. Page, Proceedings of the seventh International World Wide Web Conference (WWW7), 1998 (unpublished).
[12] M. Lifantsev, in *Proceedings of the International Conference on Internet Computing*, edited by Peter Graham and Muthucumaru Maheswaran, (CSREA Press, Las Vegas, 2000), pp. 143–148.
[13] J. Kleinberg, in Proceedings of the ninth ACM-SIAM Symposium on Discrete Algorithms, 1998 (unpublished).
[14] G. Bianconi and A.-L. Barabási, Europhys. Lett. **54**, 436 (2001).
[15] R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).
[16] S.N. Dorogotsev and J.F.F. Mendes, Europhys. Lett. **52**, 33 (2000).
[17] A. Capocci, G. Caldarelli, R. Marchetti, and L. Pietronero, Phys. Rev. E **64**, 035105 (2001)