Structure of a large social network

Gábor Csányi[1,*] and Balázs Szendrői[2,†]
[1]*TCM Group, Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, United Kingdom*
[2]*Department of Mathematics, Utrecht University, P.O. Box 80010, NL-3508 TA Utrecht, The Netherlands*
(Received 28 August 2003; revised manuscript received 13 November 2003; published 31 March 2004)

We study a social network consisting of over 10^4 individuals, with a degree distribution exhibiting two power scaling regimes separated by a critical degree k_{crit}, and a power law relation between degree and local clustering. We introduce a growing random model based on a local interaction mechanism that reproduces the observed scaling features and their exponents. We suggest that the double power law originates from two very different kinds of networks that are simultaneously present in the human social network.

DOI: 10.1103/PhysRevE.69.036131 PACS number(s): 89.75.Da, 89.75.Hc, 89.75.Fb, 89.65.Ef

INTRODUCTION

The ubiquity of networks has long been appreciated: complex systems in the social and physical sciences can often be modeled on a graph of nodes connected by edges. Recently it has also been realized that many networks arising in nature and society, such as neural networks [1], food webs [2], cellular networks [3], networks of sexual relationships [4], collaborations between film actors [1,5] and scientists [6,7], power grids [1,8], Internet routers [9], and links between pages of the World Wide Web [10–12] all share certain universal characteristics very poorly modeled by regular or simple random graphs [13]: they are highly clustered ''small worlds'' [1,14,15] with small average path length between nodes, and they have many highly connected nodes with a degree distribution often following a power law [5,10]. The network of humans with links given by acquaintance ties is one of the most intriguing of such networks [14–17], but its study has been hindered by the absence of large reliable data sets.

The aim of the present work is to introduce a data set describing a large web-based social network, to study its aggregated local and global characteristics, and to deduce some essential features of its structure through modeling. The network exhibits a novel feature, a double power law in the degree distribution; we will argue that this results from the existence of two underlying networks which compete in the formation of the actual observed contact graph. As usual in network science, the range of the observed power laws is limited by system size; however, in case of the model, choosing a sufficiently large system size leads to power laws of arbitrary length.

NEW DATA

The WIW project was started in Budapest, Hungary in April 2002 on the website www.wiw.hu with the general aim of recording social acquaintance among friends. During its 18 months of existence, it has grown into a well-known social phenomenon among Hungarian-speaking Internet users, currently (October 2003) including about 35 000 members.

The basic rules of the website are very simple. Membership is strictly invitation only; existing members can invite an unlimited number of friends to the network via email, who, if they choose to do so, join the network by an initial link connecting to the person who invited them. The list of usernames is fully searchable by all members. Additional links can be recorded among members, representing a single type of ''acquaintance,'' initiated by one member and agreed (reciprocated) by the other. The use of real names is encouraged, and users registering under pseudonyms are unrecognizable by other, genuine users; indeed, there is strong empirical evidence for the lack of proliferation of multiple pseudonyms. The website contains additional services such as messaging, discussion forums, etc.

Because of the relatively short age of the network, links formed between people newly acquainted through the website have a minimal structural effect; in addition, less than 10% of users make active use of the message board and related services. Thus the majority of the links represent genuine preexisting social acquaintance. Hence the WIW develops as a growing subgraph of the underlying social acquaintance network. Indeed, the growth process of the WIW network is essentially equivalent to the ''snowball sampling'' method well known to sociologists [17], and to the crawling methods used to investigate the World Wide Web and other computer networks.

DEGREE DISTRIBUTIONS

Basic network measures of the WIW network, as an anonymous snapshot taken in October 2002, are listed in Table I. Note that by virtue of the invitation mechanism explained above, the WIW network is necessarily connected.

The degree distribution of the network is plotted in Fig. 1. The graph shows two power law regimes

$$P(k) \sim \begin{cases} k^{-1.0} & \text{if } k < k_{\mathrm{crit}} \\ k^{-2.0} & \text{if } k > k_{\mathrm{crit.}} \end{cases}$$

*Corresponding author. Electronic address: gc121@cam.ac.uk
†Electronic address: szendroi@math.uu.nl

TABLE I. Number of nodes, edges, and the average degree, clustering, and path length for the WIW network, and a random network [13] of the same size and edge density.

	WIW	Random
V	12388	
E	74495	
$\langle k \rangle$	12.0	
$\langle C \rangle$	0.2	0.001
$\langle l \rangle$	4.5	3.8

The two regimes are separated by a critical degree $k_{\mathrm{crit}} \approx 25$. The exponent $\gamma_2 \approx -2$ of the large-k power law falls in a range that has often been observed before in a variety of contexts [3–12]. The value $\gamma_1 \approx -1$ of the small-k power law exponent is much less common, observed before only in some scientific collaboration networks [6] and food webs [2]. The possibility of a double power law was discussed in Refs. [7,18], but the WIW network is to our knowledge the first data set which demonstrates the existence of double power law behavior. A second snapshot taken in January 2003 confirms these observations, yielding a degree distribution with the same features. Since the network grew by about 50% during this period, the described distribution can be regarded as essentially stationary in time.

The operational rules of the network imply that it contains a distinguished subgraph, the invitation tree, along which membership spreads. The degree distribution of the invitation tree graph is shown in Fig. 2. While for large degrees this curve is compatible with a power law with an exponent $\gamma \approx -3$, the available data are not large enough to warrant a definite conclusion in this regard. Rather, the significance of the invitation network becomes clear below, after we introduce the random growing model. We will show that the model that reproduces the degree distribution of the full WIW network does have a power law in its invitation tree

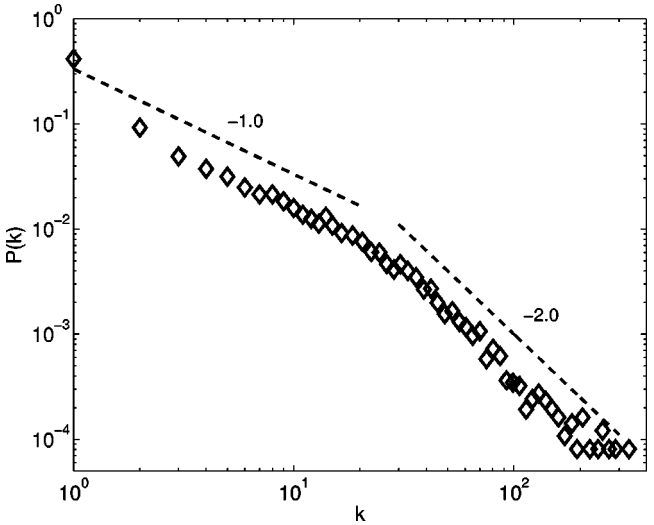

FIG. 1. The degree distribution of the WIW network, with a small-k power law $P(k) \sim k^{-1.0}$ and a large-k power law $P(k) \sim k^{-2.0}$ separated by a critical degree $k_{\mathrm{crit}} \approx 25$.

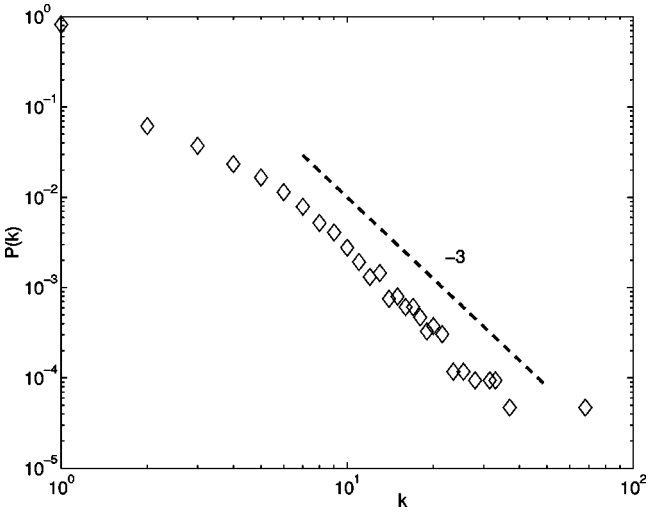

FIG. 2. The degree distribution of the invitation tree of the WIW graph, compatible with a large-k power law $P(k) \sim k^{-3.16}$.

with an exponent of -3 and that changing the adjustable parameter in our model away from a critical value simultaneously destroys the power law in the invitation tree and the small-k power law of the full degree distribution. Therefore, we argue that the power law in the distribution of the invitation network is directly linked with the small-k power law of the total degree distribution.

Since the invite network degree distribution is qualitatively different from the total degree distribution, it is reasonable to conclude that there are at least two different types of social linking at play: the network of friends defined by ties strong enough to warrant an invitation is different from the network of acquaintances that drives the mutual recognition, once both parties are registered. We suggest that it is the presence of the two graph processes, the invitation of new members and the recording of acquaintance between already registered members, that is responsible for the two scaling regimes in the degree distribution of the WIW graph.

Note that from the point of view of the node in the network, the two different linking mechanisms can be in play at different times. It is possible that when the node has few links, its link count is mostly affected by the invitation mechanism, but once it acquired many links, the triangle mechanism becomes dominant. Unfortunately time-resolved data for individual nodes were not made available to us, and therefore we could not investigate this issue any further.

CLUSTERING

The density of edges among neighbors of a fixed node v is measured by the local clustering coefficient. For a node v of degree k, the local clustering coefficient $C(v)$ is the number of acquaintance triangles of which v is a vertex, divided by $k(k-1)/2$, the number of all possible triangles. Figure 3 plots $C(k)$, the average of $C(v)$ over nodes of degree k, against the degree, showing the existence of a power law

$$C(k) \sim k^{-0.33}.$$

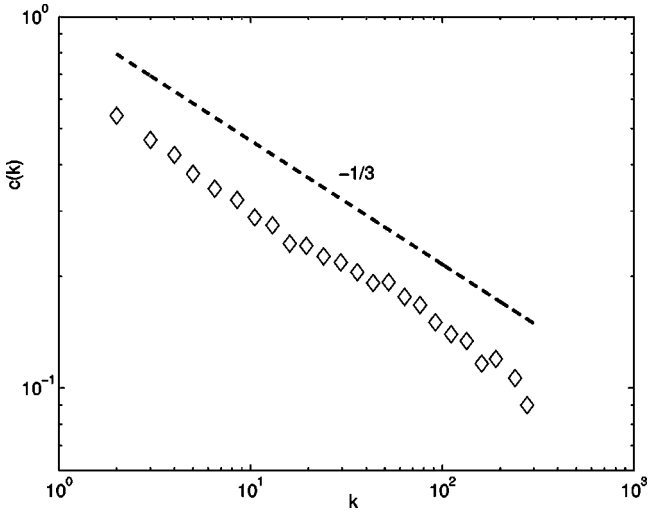

FIG. 3. The correlation between the local clustering coefficient $C(k)$ and the node degree k for the WIW graph, showing a power law $C(k)\sim k^{-0.33}$.

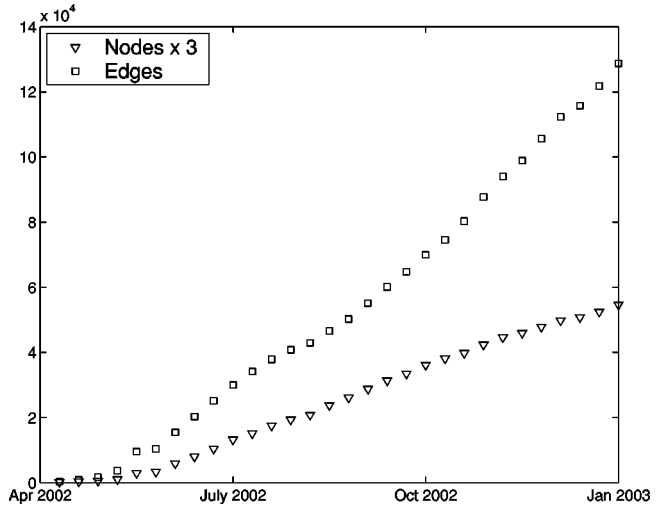

FIG. 4. The time development of the number of nodes and edges of the WIW network. Note that the number of nodes is multiplied by 3 for better visibility.

A relationship $C(k)\sim k^{-\alpha}$ was observed before in Refs. [19,20], but with significantly larger exponents. Such power laws hint at the presence of hierarchical architecture [20]: when small groups organize into increasingly larger groups in a hierarchical manner, the local clustering decreases on different scales according to such a power law.

The average clustering coefficient and average diameter of the WIW network are listed in Table I. The data show that it is a ''small-world'' network in the sense of Ref. [1]: the average path length is almost as short as in a random network [13] of the same size, but the clustering coefficient is two orders of magnitude larger.

TIME DEVELOPMENT

As Fig. 4 shows, the number of nodes of the WIW network grew approximately linearly with time. This appears to be related to the fact that the WIW network develops as a subgraph of the underlying social network, and thus the availability of new members is constrained by high clustering of the existing social links: a substantial proportion of the acquaintances of a newly invited member will have been invited already. This conclusion is supported by noting that the average number of successful invitations is very close to one (except for the very first nodes, and the latest ones whose invitations presumably have not yet all been sent out or acted upon). The average number of invitations sent out is about two. Similar observations are noted in Ref. [21] for systems growing on a background network.

Furthermore, the number of edges also grew linearly with time, and thus the edge/node ratio only grew moderately during the existence of the network. This observation is in contradiction with any purely local time-independent edge creation mechanism. If every member of the network actively participates in edge creation independently of its age in the network, the edge/node ratio would also increase linearly with time. This linear growth of the edge/node ratio was not observed in the network, and hence we conclude that the edge creation activity of members necessarily decreased with time.

The fact that the edge/node ratio changes little over time is consistent with the observed stationary nature of the degree distribution. To see this, consider a growing network with $V(t)$ nodes at time t and a time-independent degree distribution $P(k)$ with $\Sigma_k P(k)=1$ and finite first moment $\Sigma_k kP(k)$. At time t, there are

$$n(k,t)=V(t)P(k)$$

nodes of degree k, and hence the number of edges is

$$E(t)=\frac{1}{2}\sum_k kn(k,t)=\frac{V(t)}{2}\sum_k kP(k).$$

Consequently the edge/node ratio $E(t)/V(t)$ is essentially constant, and it only changes because the maximal degree increases. This argument applies to the WIW network with stationary distribution

$$P(k)=\begin{cases}\dfrac{c_1}{k^{\gamma_1}} & \text{if } k<k_{\text{crit}}\\[2ex] \dfrac{c_2}{k^{\gamma_2}} & \text{if } k>k_{\text{crit}}\end{cases}$$

with $\gamma_1\approx -1$, $\gamma_2\approx -2$. This distribution is on the boundary of distributions with finite first moment: the first moment exists for $\gamma_2<-2$ but not otherwise.

NETWORK MODELING

A random graph process based on linear preferential attachment for the creation of new edges was proposed in Ref. [5] to account for the observed power laws in natural networks. Such a process leads indeed to a graph with a power law degree distribution [5,22]. However, this model is by

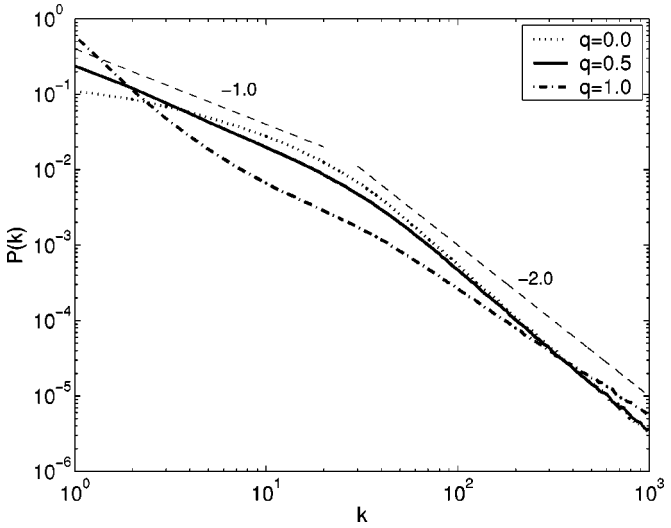

FIG. 5. The dependence of the degree distribution of our model graph on the parameter q, with $m=15$ and $V=2\times10^5$ ($q=0.5,1$) and $V=5\times10^4$ ($q=0$), averaged over 50 graphs; q_{crit} is ≈0.5.

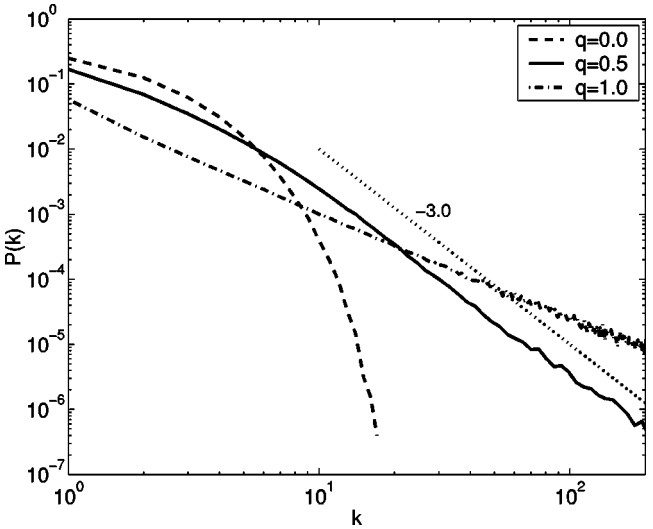

FIG. 6. The degree distribution of the invitation tree in the model, for various values of q.

definition ''macroscopic,'' requiring global information about the entire network in every step. This assumption is realistic for the World Wide Web or some collaboration networks, where all nodes are ''visible'' from all others. For human social networks however, it is reasonable to assume some degree of locality in the interactions. Also, the original scale-free models are not applicable to networks with high, degree-dependent clustering coefficients. These problems motivated the introduction of models which use local triangle creation mechanisms [15,23–26], which increase clustering in the network. These models have degree-dependent local clustering, and can also lead to power law degree distributions, though no existing model of this kind shows a double power law.

We now present a model to account for the observed properties of the WIW network. As mentioned above, the WIW can be viewed as a growing subgraph of the underlying social acquaintance graph. This suggests a model obtained by a two-step process, first modeling the underlying graph, and then implementing a growth process. The lack of available data on the underlying graph, however, prevents us from following this program directly. We build instead a growing graph in a single process, choosing the local triangle mechanism as our basic edge creation method. This models the social introduction of two members of the WIW network by a common friend some time in the past, such edges being recorded in the WIW network itself gradually over time. The invitation of new members is modeled by sublinear preferential attachment [5], motivated by experimental results on scientific collaboration networks [7,27], where the data permit the analysis of initial edge formation. We also impose the constant edge/node ratio to be consistent with the observed stationary degree distribution. Note that this constant has to be tuned from the shape of observed distributions, and cannot be inferred from the WIW data directly. The reason for this is that the WIW network has a disproportionate number of nodes of degree 1 (Fig. 1), representing people who once responded to the invitation but never returned, which distorts the edge/node ratio without significantly affecting any other property of the graph.

The precise description of our process is as follows.

(1) We begin with a small sparse graph.

(2) New nodes arrive at a rate of one per unit time and are attached to an earlier node chosen with a probability distribution giving weight k^q to a node of degree k, where $q\geqslant0$ is a parameter.

(3) Internal edges are created as follows: we select two random neighbors of a randomly chosen node v, and if they are unconnected, we create an edge between them. Otherwise, we select two new neighbors of the same node v and try again. If all neighbors of v are already connected to each other, we pick a new v.

(4) A constant number of internal edges is created per unit time, so that the edge/node ratio equals a constant m after each time step.

The degree distribution of graphs generated by our process is shown in Fig. 5. We found a very robust large-k power law of exponent $\gamma_2\approx-2$, essentially independently of the invitation mechanism. We measured the joint probability distribution of the degrees k,k' of nodes connected by new internal edges, and found that for large values, it was proportional to kk'. This directly leads to a power law exponent of -2 via standard mean-field arguments [7]. On the other hand, the small-k behavior was found to be sensitive to the invitation mechanism; Fig. 5 shows that a second power law only appears at a critical $q=q_{\mathrm{crit}}$. The value of q_{crit} depends on the edge/node ratio.

To explore the hypothesis that the low degree power law is indeed related to the invitation mechanism, we plot in Fig. 6 the degree distribution of the invitation tree of the model network for various values of the parameter q. For $q=q_{\mathrm{crit}}$, we obtain a scale-free distribution with exponent -3, similar to what was observed for the real network. Decreasing q leads to a much sharper drop in the curve, with an exponential tail for $q=0$, whereas increasing q above the critical value leads to a gelation-type behavior: new nodes

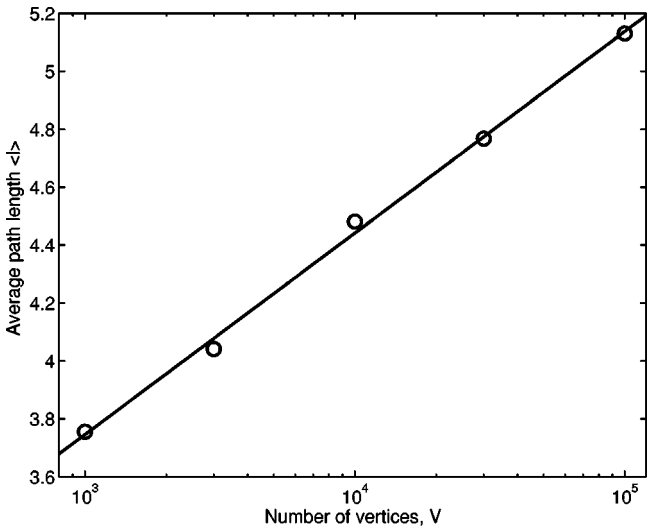

FIG. 7. The average path length as a function of the number of nodes for the graphs created by our process.

connect only to very large degree nodes.

The small-world property in a random graph process is characterized by logarithmic growth in the average path length $\langle l \rangle$ as a function of system size V. As Fig. 7 shows, the average path length indeed scales logarithmically in our process.

CONCLUSION

We have presented and analyzed a large data set of a human acquaintance network with a stable degree distribution which exhibits an interesting feature: two power law regimes with different exponents. The observed approximately constant edge/node ratio is a consequence of the stability of the degree distribution, and implies that the average activity of members is time dependent, whereas the growth of the number of nodes is constrained by the underlying social network. We introduced a random growing model which reproduces the observed degree distribution extremely well, and concluded that the small-k power law is related to the scale-free nature of the invitation tree, whereas the large-k power law is a result of the triangle mechanism of social introductions. Our results show that human social networks are likely to be composed of several networks with different characteristics, and directly observable processes will exhibit a mixture of features resulting from distinct underlying mechanisms. Multiple layers of networks are indeed likely to exist outside of the context of social networks as well; for example, technological networks might contain evidence of development at different times, transport networks obviously decompose into networks defined by range and transport type, collaboration networks have links of different strengths and the list could be continued. In our view, the interaction between layers in different classes of networks deserves further investigation.

ACKNOWLEDGMENTS

We thank Zsolt Várady and Dániel Varga for access to the data of the WIW network and Risi Kondor for helpful discussions.

[1] D.J. Watts *et al.*, Nature (London) **393**, 440 (1998).
[2] J.M. Montoya and R.V. Sole, J. Theor. Biol. **214**, 405 (2002).
[3] H. Jeong *et al.*, Nature (London) **407**, 651 (2000).
[4] F. Liljeros *et al.*, Nature (London) **411**, 907 (2001).
[5] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[6] M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[7] A.-L. Barabási *et al.*, Physica A **311**, 590 (2002).
[8] L.A.N. Amaral, A. Scala, M. Barthelemy, and H.E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11 149 (2000).
[9] M. Faloutsos *et al.*, Comput. Commun. Rev. **29**, 251 (1999).
[10] R. Albert *et al.*, Nature (London) **401**, 130 (1999).
[11] R. Kumar, P. Raghavan, S. Rajalopagan, and A. Tomkins, in *Proceedings of the Ninth ACM Symposium on Principles of Database Systems* (Association for Computing Machinery, New York, 2000), p. 1.
[12] A. Broder *et al.*, Comput. Netw. **33**, 309 (2000).
[13] P. Erdős and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).
[14] S. Milgram, Psychol. Today **2**, 60 (1967).
[15] D.J. Watts, *Small Worlds* (Princeton University Press, Princeton, NJ, 1999).
[16] I. de Sola Pool and M. Kochen, Soc. Networks **1**, 1 (1978).
[17] S. Wassermann and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[18] M.E.J. Newman, Phys. Rev. E **64**, 016131 (2001).
[19] A. Vazquez *et al.*, Phys. Rev. E **65**, 066130 (2002).
[20] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[21] R. Pastor-Satorras *et al.*, Phys. Rev. E **63**, 066117 (2001).
[22] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, Random Struct. Algorithms **18**, 279 (2001).
[23] E.M. Jin *et al.*, Phys. Rev. E **64**, 046132 (2001).
[24] A. Vazquez, Europhys. Lett. **54**, 430 (2001).
[25] P. Holme and B.J. Kim, Phys. Rev. E **65**, 026107 (2002).
[26] J. Davidsen *et al.*, Phys. Rev. Lett. **88**, 128701 (2002).
[27] M.E.J. Newman, Phys. Rev. E **64**, 025102 (2001).