Percolation theory applied to measures of fragmentation in social networks

Yiping Chen,[1] Gerald Paul,[1] Reuven Cohen,[2] Shlomo Havlin,[3] Stephen P. Borgatti,[4] Fredrik Liljeros,[5] and H. Eugene Stanley[1]
[1]*Center for Polymer Studies, Boston University, Boston, Massachusetts 02215, USA*
[2]*Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215, USA*
[3]*Minerva Center and Department of Physics, Bar-Ilan University, 52900 Ramat-Gan, Israel*
[4]*Department of Org. Studies, Boston College, Chestnut Hill, Massachusetts 02467, USA*
[5]*Department of Sociology, Stockholm University, S-106 91 Stockholm, Sweden*
(Received 22 October 2006; published 13 April 2007)

We apply percolation theory to a recently proposed measure of fragmentation F for social networks. The measure F is defined as the ratio between the number of pairs of nodes that are not connected in the fragmented network after removing a fraction q of nodes and the total number of pairs in the original fully connected network. We compare F with the traditional measure used in percolation theory, P_∞, the fraction of nodes in the largest cluster relative to the total number of nodes. Using both analytical and numerical methods from percolation, we study Erdős-Rényi and scale-free networks under various types of node removal strategies. The removal strategies are random removal, high degree removal, and high betweenness centrality removal. We find that for a network obtained after removal (all strategies) of a fraction q of nodes above percolation threshold, $P_\infty \approx (1-F)^{1/2}$. For fixed P_∞ and close to percolation threshold ($q=q_c$), we show that $1-F$ better reflects the actual fragmentation. Close to q_c, for a given P_∞, $1-F$ has a broad distribution and it is thus possible to improve the fragmentation of the network. We also study and compare the fragmentation measure F and the percolation measure P_∞ for a real social network of workplaces linked by the households of the employees and find similar results.

DOI: 10.1103/PhysRevE.75.046107 PACS number(s): 89.75.Fb, 89.75.Hc, 89.65.−s

I. INTRODUCTION

Many physical, sociological, and biological systems are represented by complex networks [1–17]. One of the important problems in complex networks is the fragmentation of networks [6–12]. In this problem one studies the statistical properties of the fragmented networks after removing nodes (or links) from the original fully connected network using a certain strategy. Many different removal strategies have been developed for various purposes, e.g., mimicking the real-world network failures, improving the effectiveness of network disintegration, etc. Examples include random removal (RR) strategy, the high degree removal (HDR) strategy, and the high betweenness centrality removal strategy (HBR) [9,18–21]. Note that the best strategy for fragmentation (minimum nodes removal) is also the best for immunization since it represents the minimum number of nodes or links needed to be immunized so that the epidemic cannot spread in the network.

Recently, a new measure of fragmentation has been developed in social network studies [22]. Given a fully connected network of N nodes which is fragmented into separate clusters [23] by removing m nodes following a certain strategy, we define $q \equiv m/N$ as the concentration of nodes removed and $p \equiv 1-q$ as the concentration of existing nodes. The degree of fragmentation F of the network is defined as the ratio between the number of pairs of nodes that are not connected in the fragmented network and the total number of pairs in the original fully connected network. Suppose that after removal there are n clusters in the fragmented network, since all members of a cluster are, by definition, mutually reachable, the measure F can be written as follows [22]:

$$F \equiv 1 - \frac{\sum_{j=1}^{n} N_j(N_j-1)}{N(N-1)} \equiv 1 - C. \tag{1}$$

Here, N_j is the number of nodes in cluster j, n is the number of clusters in the fragmented network, and N is the number of nodes in the original fully connected network. For an undamaged network, $F=0$. For a totally fragmented network, $F=1$. The quantity C defined in Eq. (1) can be regarded as the "connectivity" of the network. When $C=1$ the network is fully connected while for $C=0$ it is fully fragmented.

In this paper, we study the statistical behavior of $F \equiv 1-C$ using both analytical and numerical methods and relate it to the traditional measure of fragmentation, the relative size of the largest cluster P_∞ used in percolation theory. In this way, we are able to obtain analytical results for the fragmentation F of networks. We study three removal strategies: the *random removal (RR) strategy* which removes randomly selected nodes, the *high degree removal (HDR) strategy* which targets and removes nodes with the highest degree, and the *high betweenness centrality removal (HBR) strategy* which targets and removes nodes with the highest betweenness centrality. The HDR (or HBR) strategies first remove the node with the highest degree (or the highest betweenness centrality), and then the second highest, and so on. These three strategies are commonly used in models representing random and targeted attacks in real-world networks [1,6–8,20].

II. THEORY

Traditionally, in analogy to percolation, physicists describe the connectivity of a fragmented network by the ratio $P_\infty \equiv N_\infty/N$ (called the incipient order parameter) between

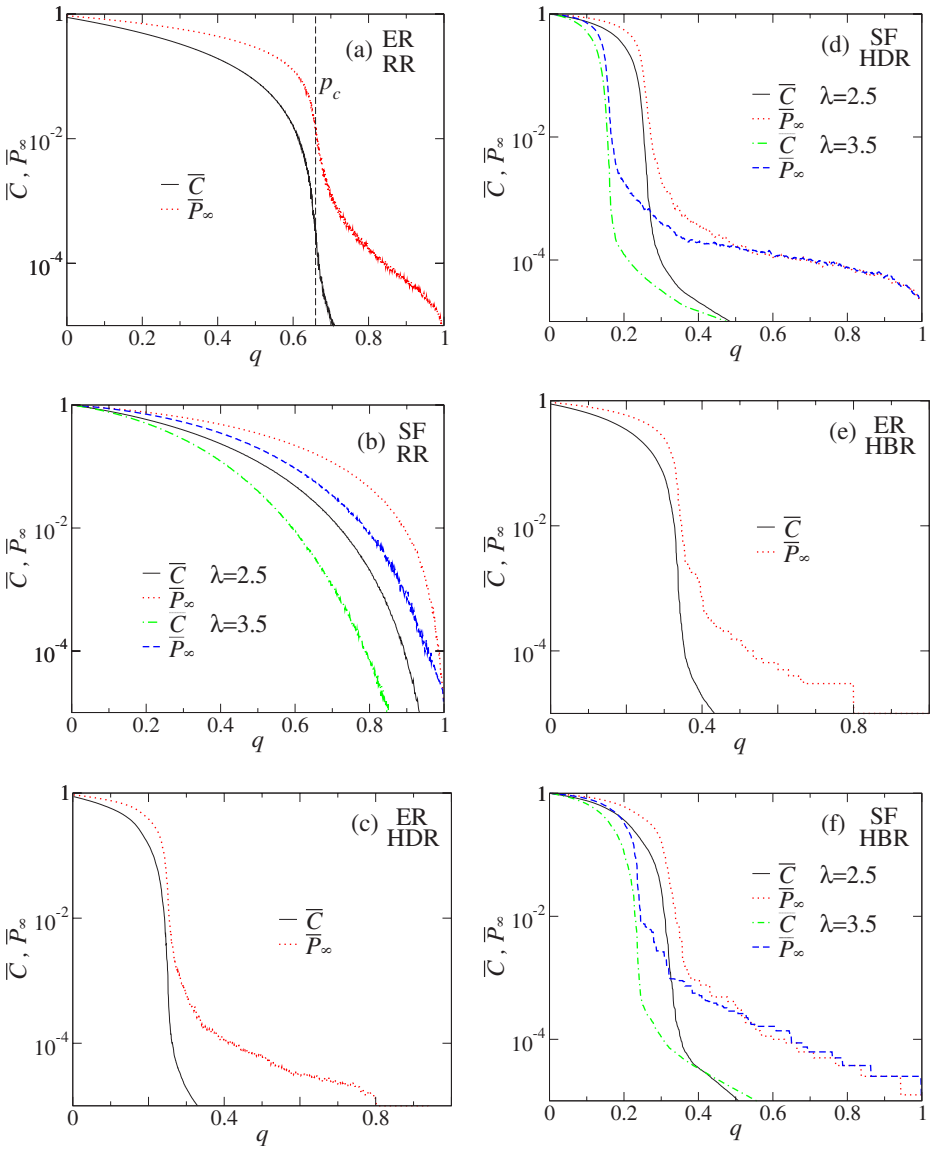

FIG. 1. (Color online) The behavior of $\bar{C}$ and $\bar{P}_\infty$ vs q on ER and SF networks. For ER networks, $N=200\,000$ and $\langle k\rangle=3$. For SF networks, $N=80\,000$. The graphs are (a) RR strategy on ER networks, (b) RR strategy on SF networks, (c) HDR strategy on ER networks, (d) HDR strategy on SF networks, (e) HBR strategy on ER networks, and (f) HBR strategy on SF networks.

the largest cluster size N_∞ (called the infinite cluster) and N.Many properties have been derived for this measure [6,24,25]. For example, in random networks, P_∞ undergoes a second-order phase transition at a threshold p_c. Below p_c, P_∞ is zero for $N\to\infty$, while for $p>p_c$, P_∞ is finite. This occurs for both RR and HDR in random networks and lattice networks [6–8,24,25]. The threshold parameter p_c depends on the degree distribution, the network topology, and the removal strategy [6–8,24,25]. The specific way that P_∞ approaches zero at p_c depends on the network topology and removal strategy but not on details such as p_c. In scale-free networks, where the degree distribution $p(k)\sim k^{-\lambda}$ and $2<\lambda<3$, it has been found that $p_c\to 0$ for RR strategy [6] while p_c is very high for HDR strategy [7,8] and for HBR strategy [20]. For $\lambda>3$ and RR, p_c is finite.

Next, we show simulation results of removing nodes in all strategies (RR, HDR, and HBR) on ER and scale free networks. Figure 1 shows the behavior of $\bar{C}$ and $\bar{P}_\infty$, the average of $C(\equiv 1-F)$ and P_∞ over 1000 realizations, vs q for Erdős-Rényi (ER) and scale-free (SF) networks with RR [Figs. 1(a) and 1(b)], HDR [Figs. 1(c) and 1(d)], and HBR [Figs. 1(e) and 1(f)] strategies. As seen in Fig. 1(a), the network becomes more fragmented when q increases and both measures drop sharply at $q_c=1-p_c$. Note that $\bar{C}$ shows a transition similar to $\bar{P}_\infty$ at $p=p_c$; however, above q_c, $\bar{C}$ becomes more flat in contrast to $\bar{P}_\infty$, indicating the effect of connectivity in the small clusters which do not affect P_∞.

In contrast to Fig. 1(a), the transition in Fig. 1(b) is not as sharp and therefore $\bar{C}$ and $\bar{P}_\infty$ do not show a collapse together. The reason is that for $\lambda=2.5$ there is no transition at $q<1$ [6] and for $\lambda=3.5$, $\bar{P}_\infty$ falls much less sharply compared to ER [26]. For HDR shown in Figs. 1(c) and 1(d), the transition is again sharp since after removing high degree nodes, the network becomes similar to ER networks, which do not have high degree nodes [8]. A similar behavior is seen for HBR shown in Figs. 1(e) and 1(f) due to the known high correlation between high degree nodes and high betweenness centrality nodes [20].

Following percolation theory, Eq. (1) can be written as

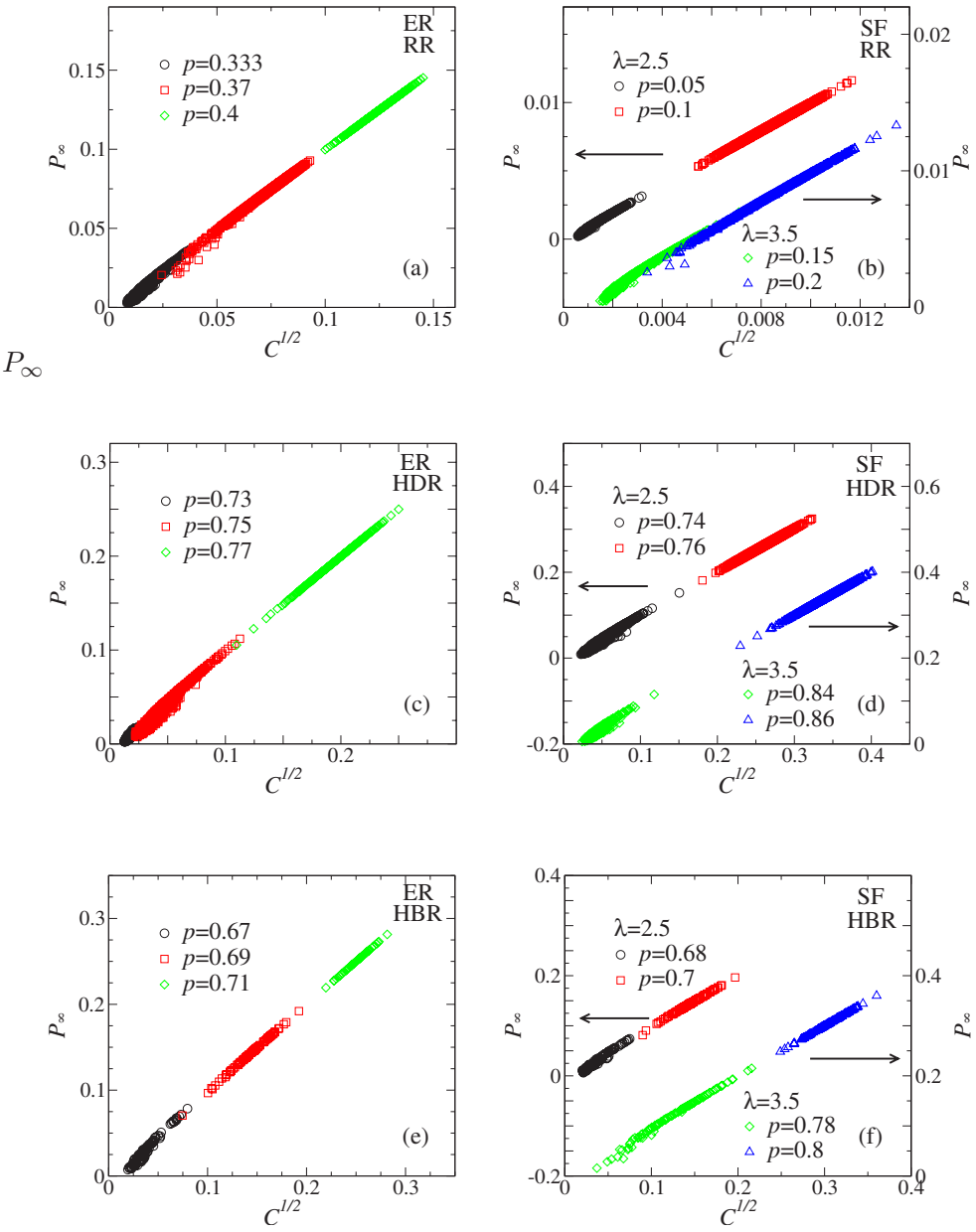

FIG. 2. (Color online) Relationship between $C^{1/2}$ and P_∞ for ER and SF networks with system size $N=50\,000$. For ER networks, the average degree $\langle k\rangle=3$, and for SF networks, $\lambda=2.5$ and 3.5. The graphs are (a) RR strategy on ER networks, (b) RR strategy on SF networks, (c) HDR strategy on ER networks, (d) HDR strategy on SF networks, (e) HBR strategy on ER networks, and (f) HBR strategy on SF networks.

$$C \equiv 1 - F \equiv \frac{\Sigma_{j=1}^{n} N_j(N_j-1)}{N(N-1)} \approx \frac{\Sigma_{j=1}^{n} N_j^2 - \Sigma_{j=1}^{n} N_j}{N^2} = \frac{\Sigma_{j=1}^{n} N_j^2}{N^2} - \frac{1}{N} = P_\infty^2 + \frac{S-1}{N}. \quad (2)$$

To obtain Eq. (2), we denote $j=1$ as the largest cluster and thus $N_1^2/N^2 \equiv P_\infty^2$. The sum $\Sigma_{j=2}^{n} N_j^2/N \equiv S$, where S is the mean cluster size of finite clusters [24,25]. Since S is of order of $\ln N$, $(S-1)/N$ can be neglected for large N. Therefore we expect that P_∞ and C have the relationship $P_\infty \approx C^{1/2}$ when $p > p_c$ (but not too close to p_c). When $p \leqslant p_c$, the infinite cluster loses its dominance in the system with $P_\infty \sim \ln(N)/N \to 0$ and both terms in Eq. (2) are roughly in the same order for large N [8]. Here significant variations between P_∞ and $C^{1/2}$ are expected, as indeed seen in Fig. 2.

III. SIMULATIONS

We test by simulations the relationship $C \sim P_\infty^2$ derived for $p > p_c$ in Eq. (2). In Fig. 2(a) we plot P_∞ vs $C^{1/2}$ for RR strategy in ER networks and for several values of p. As predicted by Eq. (2), the plot of P_∞ vs $C^{1/2}$ yields a linear relationship with slope equal to 1 when $p > p_c = 1/\langle k\rangle = 1/3$. The range of P_∞ and $C^{1/2}$ for $p=0.4$ is due to the variation of P_∞ for a given p and the same variation appears for $C^{1/2}$ showing that the infinite cluster dominates and Eq. (2) is valid. However, when p drops close to $p_c=1/3$, the system approaches criticality and the one-to-one correspondence between $C^{1/2}$ and P_∞ is not as strong. This variation is attributed to the presence of clusters other than the infinite one, which influence C but not P_∞.

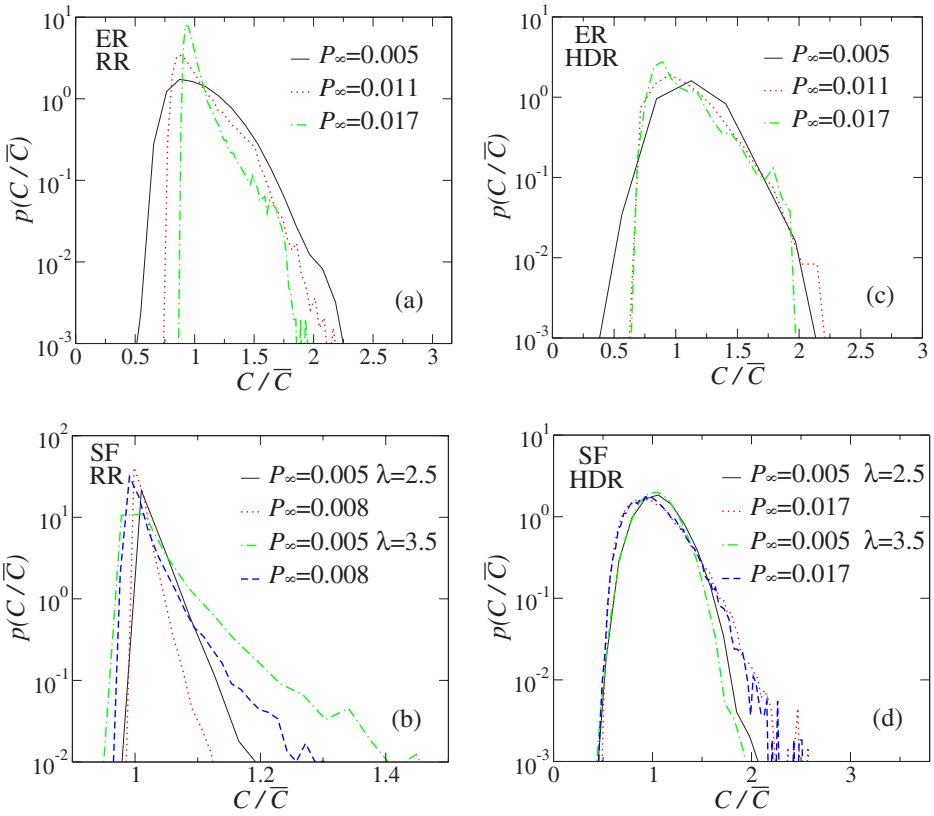

FIG. 3. (Color online) Probability distributions $p(C/\bar{C})$ vs $C/\bar{C}$ for several values of P_∞ and for ER networks with $\langle k\rangle=3$, $N=200\,000$, and SF networks with $N=80\,000$ and $\lambda=2.5$ and 3.5. (a) RR strategy on ER networks, (b) RR strategy on SF networks, (c) HDR strategy on ER networks, and (d) HDR strategy on SF networks.

Similar behavior is observed for RR strategy in SF networks with $\lambda=3.5$ shown in Fig. 2(b). For $\lambda=3.5$, the variation in $C^{1/2}$ emerges close to $p_c=0.2$. However, for $\lambda=2.5$, percolation theory suggests that p_c approaches 0 for large systems. As a result, no significant variation is observed even when P_∞ is as small as 5×10^{-4}. This observation supports that the SF networks with $\lambda<3$ are quite robust in sustaining its infinite cluster against random removal [6]. Figures 2(c)–2(f) show the results for HDR and HBR strategies in ER and SF networks. For these targeted strategies, the variation of $C^{1/2}$ and P_∞ shows up at significantly higher p compared to the random case, indicating that the infinite cluster breaks down easier under HDR and HBR attacks for both ER and SF networks, as seen also in Fig. 1. At this point, the SF network with $\lambda=2.5$ becomes no longer as robust as in the random case, as can be clearly observed in the large variation at $P_\infty\approx0.05$.

To further investigate the characteristics of the variation of C for a given P_∞, we remove nodes until P_∞ equal to a certain value and calculate the probability distributions $p(C/\bar{C})$ vs $C/\bar{C}$. The results are plotted in Fig. 3. In this case, C^*, the most probable value of C, is determined by the fixed infinite cluster size P_∞ with $C^*\approx P_\infty^2$, and the broadness of $p(C)$ comes from presence of clusters other than the infinite one. Because the largest cluster size is fixed, the upper cutoff of $p(C)$ emerges due to the limitation on the sizes of other clusters that by definition must be smaller than the largest cluster. For the RR strategy, the broadness of $p(C)$ for the ER network is bigger than that of SF networks at the same P_∞, especially for $\lambda=2.5$ where the system is always high above criticality and the variation is relatively small. On the con-

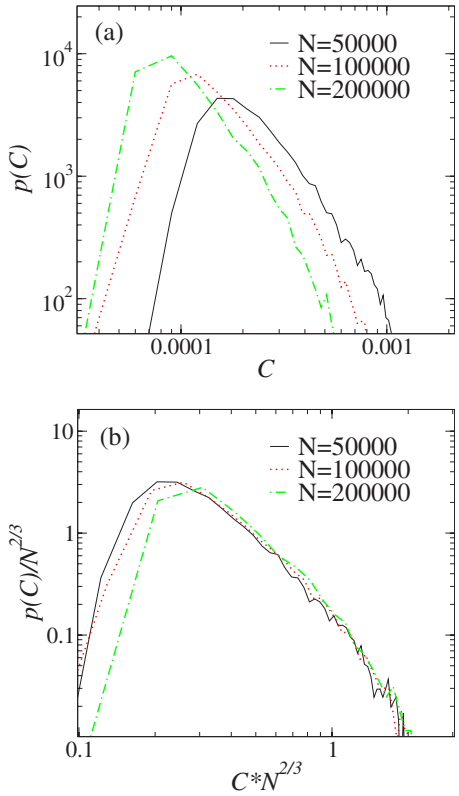

FIG. 4. (Color online) The dependence of $p(C)$ on the system size N with $p=p_c$ for (a) before scaling and (b) after scaling. Simulations are performed on ER networks with $\langle k\rangle=3$.

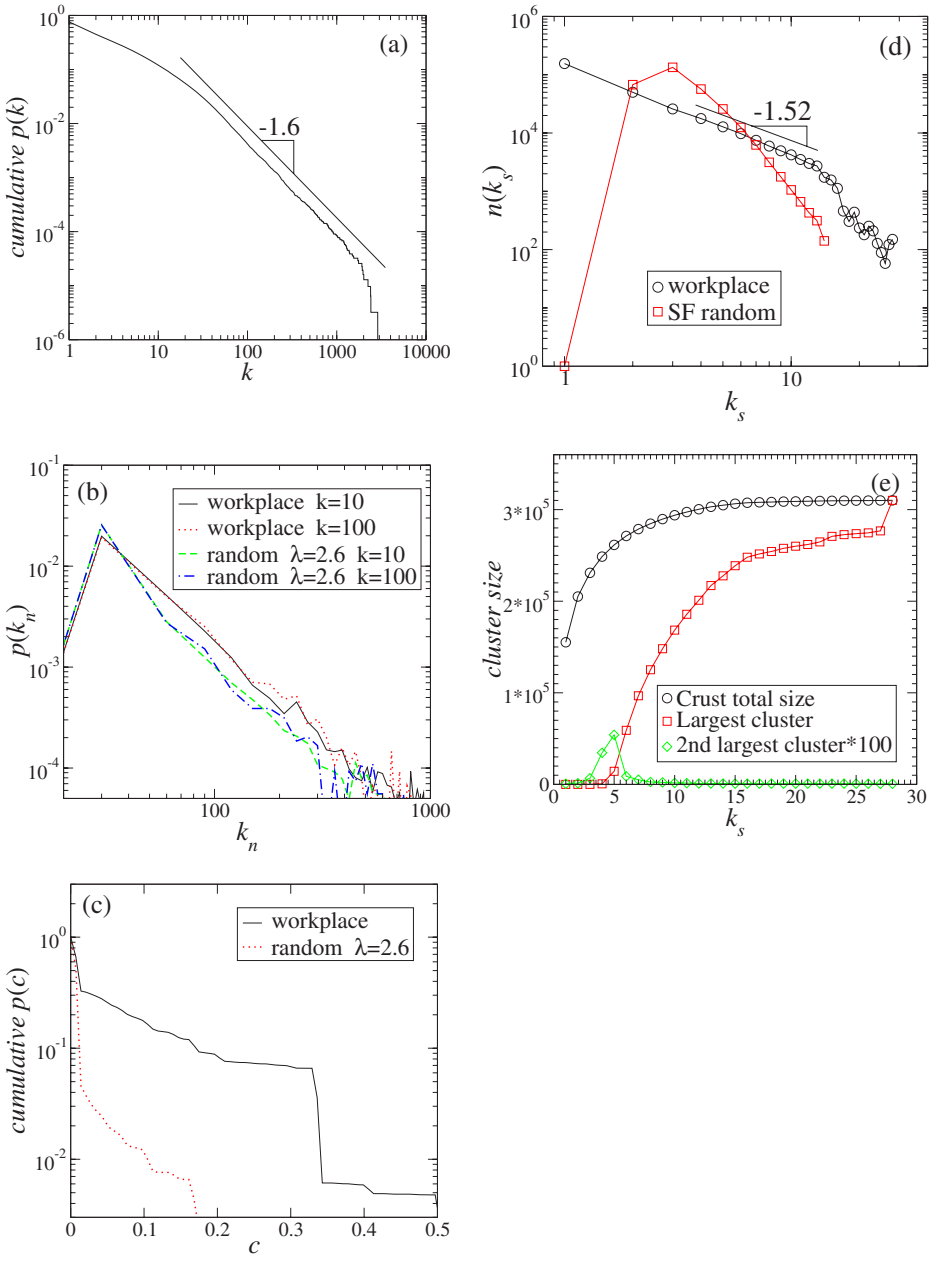

FIG. 5. (Color online) Properties of the Swedish network of workplaces. (a) The cumulative degree distribution (showing $\lambda = 2.6$). (b) The distribution of k_n, the degree of the neighbors of nodes having degree k. (c) The cumulative distribution of clustering coefficient c. (d) Number of nodes in shell k_s. (e) Size of largest and second largest cluster in each k crust. In (b), (c), and (d) the distributions of random SF networks with the same λ and N are plotted for comparison.

trary, for the HDR and HBR strategies, the broadness of $p(C)$ for ER and SF networks are of the same order due to the fact that for HDR and HBR, p_c is also finite for $\lambda = 2.5$. This observation is consistent with the results shown in Fig. 2.

Now we focus on the dependence of $p(C)$ on the system size N at p_c (Fig. 4). From percolation theory and for ER under RR strategy, the infinite cluster size N_∞ at criticality behaves as [27,28]

$$N_\infty \sim N^{2/3}. \tag{3}$$

Since C follows similar behavior as N_∞ at criticality, we expect C for $p = p_c$ to behave as

$$C \equiv 1 - F \approx (N_\infty/N)^2 \sim N^{-2/3}. \tag{4}$$

Thus we expect the probability distribution $p(C)$ with $p = p_c$ to scale as

$$p(C) = N^{2/3} g(CN^{2/3}), \tag{5}$$

where g is a scaling function.

Figure 4(b) supports this scaling relationship. We calculate $p(C)$ for RR strategy at criticality on ER networks with N values of 50 000, 100 000, 200 000, and $\langle k \rangle = 3$ [shown in Fig. 4(a)], and [34] find a good collapse when plotted [Fig. 4(b)] using the scaling form of Eq. (5).

IV. REAL NETWORKS

The ER networks and the SF networks that we have been studying are a random ensemble of networks which are only determined by their degree distribution. It is known that many real networks often exhibit important structural properties relevant for percolation properties such as a high level of clustering, assortativity, and fractality that random net-

works do not exhibit [13,29]. We therefore test our results about the relation between C and P_∞ on an example of a large real social network. The network we use is extracted from a data set obtained from Statistics Sweden [30] and consists of all geographical workplaces in Sweden that can be linked with each other by having at least one employee from each workplace sharing the same household. Household is defined as a married couple or a couple having kids together that are living in the same flat or house. Unmarried couples without kids and other individuals sharing a household are not registered in the data set as households. This kind of network has been shown to be of importance for the spreading of influenza [31] and are also likely to be important for spreading information and rumors in society. The network consists of 310 136 nodes (workplaces) and 906 260 links (employees sharing the same households) and, as shown in Fig. 5(a), is approximately a SF network with $\lambda \approx 2.6$ and an exponential cutoff. The network shows almost no degree-degree correlation (assortativity) [Fig. 5(b)]. However, the workplace network clustering coefficient c is significantly higher than that of a random SF network with same λ and N [Fig. 5(c)]. The average of c is 0.048 for the workplace network vs 3.2×10^{-4} for the random SF networks, which is consistent with the earlier social network studies [32,33]. Figure 5(d) shows the node distribution $n(k_s)$ of k shell (k_s) in the network compared to that of a random SF network with the same λ and $\langle k\rangle$ [34]. The one-shell is obtained by pruning all nodes with degree 1 (or less) away from the network until no more such nodes remain. A similar procedure is performed recursively for larger degrees to get other k shells. It is seen that in the workplace network there exist significantly more shells and the large shells are more occupied compared to random SF networks. The distribution $n(k_s)$ shows a power-law behavior with slope −1.52. This indicates the structure of this real network. Figure 5(e) shows the crust total size, the largest cluster size, and the second largest cluster size as a function of shell k_s. The k crust is defined as the union of all shells with indices smaller or equal to k. It is seen that the largest cluster has two transitions. One around $k_s=5$ and the other at $k_s=27$. At $k_s>5$, the largest cluster increases from zero to a finite fraction of the network. This transition is related to the HDR seen in Fig. 6(d) (see also [8]). The second transition at $k_s=27$ defines the nucleus of the workplace network which includes about 100 nodes [see Fig. 5(d), $n(28)\approx100$] which are well connected to each other. The jump of the largest cluster from $k_s=27$ to $k_s=28$ from 2.8×10^5 nodes to 3.1×10^5 nodes (i.e., 3×10^4 nodes) is due to nodes which are connected only to the nucleus. These nodes are called dendrites. Figure 5(e) is very similar to the Medusa model [34] suggested for the AS topology of the Internet. Figures 6(a) and 6(b) show simulation results for several values of p for P_∞ vs $C^{1/2}$. The curves are linear, similar to Fig. 2 for our model networks. Moreover, Figs. 6(c) and 6(d) show that $\bar{C}^{1/2}$ and $\bar{P}_\infty$ are almost identical above the criticality threshold p_c for a typical configuration after both RR and HDR. For p below criticality, differences appear which are especially obvious for HDR strategy where $q_c=1-p_c$ is relatively small. While $\bar{P}_\infty$ rapidly decreases to a very small value (below 10^{-5}), a plateau shows up in the

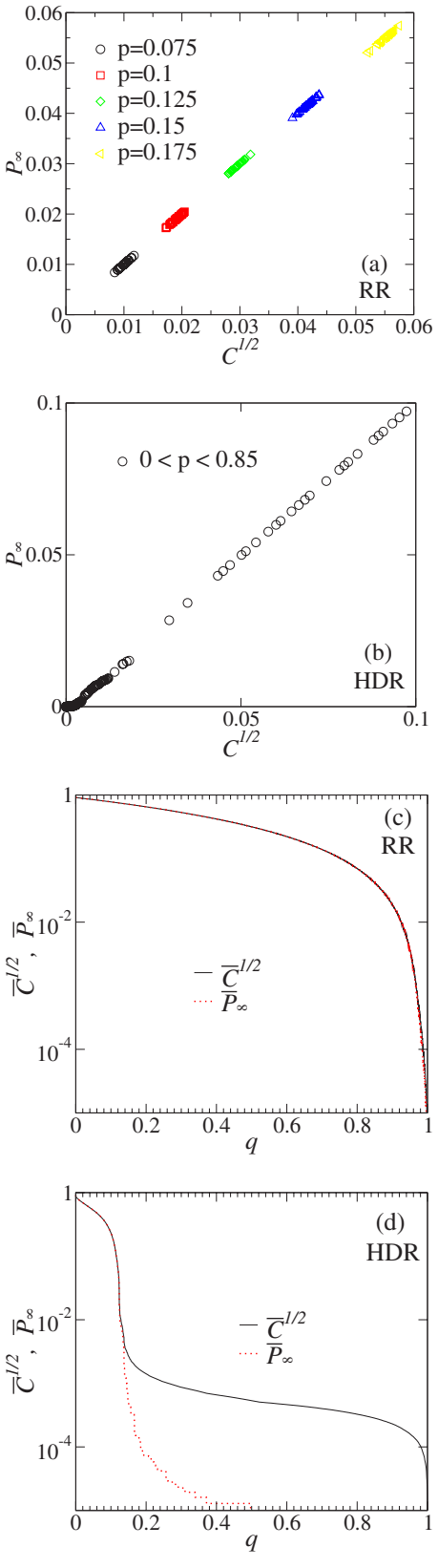

FIG. 6. (Color online) P_∞ vs $C^{1/2}$ for (a) RR strategy and (b) HDR strategy and plot $\bar{C}^{1/2}$, $\bar{P}_\infty$ vs q for (c) RR strategy and (d) HDR strategy for the Swedish network of workplaces with $N=310\,136$ nodes.

curve of $C^{1/2}$ due to the influence of the small clusters.

V. SUMMARY

In summary, we study the measure for fragmentation $F \equiv 1-C$ proposed in social sciences and relate it to the traditional P_∞ used in physics in percolation theory. For p above criticality, C and P_∞ are highly correlated and $C \approx P_\infty^2$. Close to criticality, for $p \geq p_c$ and below p_c, variations between C and P_∞ emerge due to the presence of the small clusters. For systems close to or below criticality, F gives a better measure for fragmentation of the whole system compared to P_∞. We study the probability distribution $p(C)$ for a given P_∞ and find that $p(C)$ at $p=p_c$ obeys the scaling relationship $p(C) = N^{2/3} g(CN^{2/3})$ for both RR strategy on ER network, and for HDR on scale-free networks.

ACKNOWLEDGMENTS

We thank ONR, European NEST project DYSONET, and Israel Science Foundation for financial support. The study was approved by the Regional Ethical Review board in Stockholm (record 2004/2:9).

[1] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **406**, 378 (2000).
[2] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, England, 2004).
[3] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[5] V. Paxon, IEEE/ACM Trans. Netw. **5**, 601 (1997).
[6] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).
[7] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).
[8] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **86**, 3682 (2001).
[9] A. X. C. N. Valente, A. Sarkar, and H. A. Stone, Phys. Rev. Lett. **92**, 118702 (2004).
[10] G. Paul, T. Tanizawa, S. Havlin, and H. E. Stanley, Eur. Phys. J. B **38**, 187 (2004).
[11] F. Chung and L. Lu, Ann. Comb. **6**, 125 (2002).
[12] Z. Burda and A. Krzywicki, Phys. Rev. E **67**, 046118 (2003).
[13] C. Song *et al.*, Nature (London) **433**, 392 (2005).
[14] L. C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science* (Empirical, 2004).
[15] S. Wasserman, K. Faust, D. Iacobucci, and M. Granovetter, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, England, 1994).
[16] G. Paul, S. Sreenivasan, and H. E. Stanley, Phys. Rev. E **72**, 056130 (2005).
[17] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[18] T. Tanizawa, G. Paul, R. Cohen, S. Havlin, and H. E. Stanley, Phys. Rev. E **71**, 047101 (2005).
[19] R. Pastor-Satorras and A. Vespignani, Phys. Rev. E **65**, 036104 (2002).
[20] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, Phys. Rev. E **65**, 056109 (2002).
[21] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[22] S. P. Borgatti, Comput. Math. Org. Theory **12**, 21 (2006).
[23] Group of connected nodes known as "component" in the language of sociology.
[24] A. Bunde and S. Havlin, *Fractals and Disordered Systems* (Springer, New York, 1995).
[25] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1994).
[26] R. Cohen, D. ben-Avraham, and S. Havlin, Phys. Rev. E **66**, 036113 (2002).
[27] P. Erdős and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).
[28] R. Cohen, S. Havlin, and D. ben-Avraham, in *Handbook of Graphs and Networks*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, New York, 2002), Chap. 4.
[29] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[30] WWW.SCB.SE
[31] C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell, Science **312**, 447 (2006).
[32] G. Csányi and B. Szendrői, Phys. Rev. E **69**, 036131 (2004).
[33] K. Klemm and V. M. Eguíluz, Phys. Rev. E **65**, 057102 (2002).
[34] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, e-print arXiv:cond-mat/0601240.