Available online at www.sciencedirect.com
ScienceDirect
Physica A 378 (2007) 11–19
PHYSICA A
www.elsevier.com/locate/physa

Percolation theory and fragmentation measures in social networks

Yiping Chen[a], Gerald Paul[a], Reuven Cohen[b], Shlomo Havlin[c,*], Stephen P. Borgatti[d], Fredrik Liljeros[e], H. Eugene Stanley[a]

[a]Center for Polymer Studies, Boston University, Boston, MA 02215, USA
[b]Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215, USA
[c]Minerva Center and Department of Physics, Bar-Ilan University, 52900 Ramat-Gan, Israel
[d]Department of Org. Studies, Boston College, Chestnut Hill, MA 02467, USA
[e]Department of Sociology, Stockholm University, S-106 91 Stockholm, Sweden

Available online 18 December 2006

Abstract

We study the statistical properties of a recently proposed social networks measure of fragmentation F after removal of a fraction q of nodes or links from the network. The measure F is defined as the ratio of the number of pairs of nodes that are not connected in the fragmented network to the total number of pairs in the original fully connected network. We compare this measure with the one traditionally 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, we study Erdős–Rényi (ER) and scale-free (SF) networks under various node removal strategies. We find that for a network obtained after removal of a fraction q of nodes above criticality, $P_\infty \approx (1-F)^{1/2}$. For fixed P_∞ and close to criticality, we show that $1-F$ better reflects the actual fragmentation. For a given P_∞, $1-F$ has a broad distribution and thus one can improve significantly the fragmentation of the network. We also study and compare the fragmentation measure F and the percolation measure P_∞ for a real national social network of workplaces linked by the households of the employees and find similar results.

Keywords: Social network; Fragmentation; Percolation theory

1. Introduction

Complex networks can be used to model many physical, sociological and biological systems and have attracted much attention in recent years [1–14]. Among the problems related to complex networks, the fragmentation of networks has been extensively studied [5–11]. The problem is defined as finding 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.

*Corresponding author. Tel.: +972 3 531 8436; fax: +972 3 535 7678.
E-mail address: havlin@ophir.ph.biu.ac.il (S. Havlin).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.11.074

Examples include random removal (RR) strategy, the high degree removal (HDR) strategy and the high centrality removal strategy [8,15–17].

Recently, a new measure of fragmentation has been developed in social network studies [18]. Suppose a fully connected network of N nodes is fragmented into separate clusters [19] by removing m nodes following a certain strategy. We define $q \equiv m/N$ the ratio of nodes removed and $p \equiv 1 - q$ the ratio 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 to the possible number of pairs in the original fully connected network. Suppose there are m clusters in the fragmented network, since all members of a cluster are, by definition, mutually reachable, the measure F can be written as follows [18]:

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

Here, N_j is the number of nodes in cluster j, m is number of clusters in the fragmented network, and N 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, 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 two removal strategies: the random removal (RR) strategy which removes randomly selected nodes and the high degree removal (HDR) strategy which targets and removes nodes with highest degree. The HDR strategy first removes the node with the highest degree, and then the second highest and so on. These two strategies are commonly used in models representing random and targeted attacks in real world networks [2,5–7].

2. 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 the largest cluster size N_∞ (called the infinite cluster) and N. Many properties have been derived for this measure [5,20,21]. 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 [5–7,20]. The threshold parameter p_c depends on the degree distribution, the network topology, and the removal strategy [5–7,20,21]. 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 [5] while p_c is very high for HDR strategy [6,7]. For $\lambda > 3$ and RR, p_c is finite.

Next, we show simulation results of removing nodes in both strategies (RR and HDR) on ER and scale free networks. Fig. 1 shows the behavior of C ($\equiv 1 - F$) and P_∞ versus q for Erdős–Rényi (ER) and scale-free (SF) networks with RR (Fig. 1(a) and (b)) and HDR (Fig. 1(c) and (d)) 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 C shows a transition similar to P_∞ at $p = p_c$; however, above q_c, C becomes more flat in contrast to P_∞, indicating the effect of connectivity in the small clusters which do not effect P_∞.

In contrast to Fig. 1(a), the transition in Fig. 1(b) is not so sharp and therefore C and P_∞ do not show a collapse together. The reason is that for $\lambda = 2.5$ there is no transition at $q < 1$ [6] and for $\lambda = 3.5$, P_∞ falls much less sharply compared to ER [22]. For HDR shown in Fig.1(c) and (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 [7].

When $p > p_c$ and not too close to p_c, following percolation theory, the infinite cluster dominates the system and $P_\infty \approx p$, i.e., most of unremoved nodes are connected. Thus, we assume that the small clusters will have a small effect on C compared to the largest one. Using this assumption, Eq. (1) can be

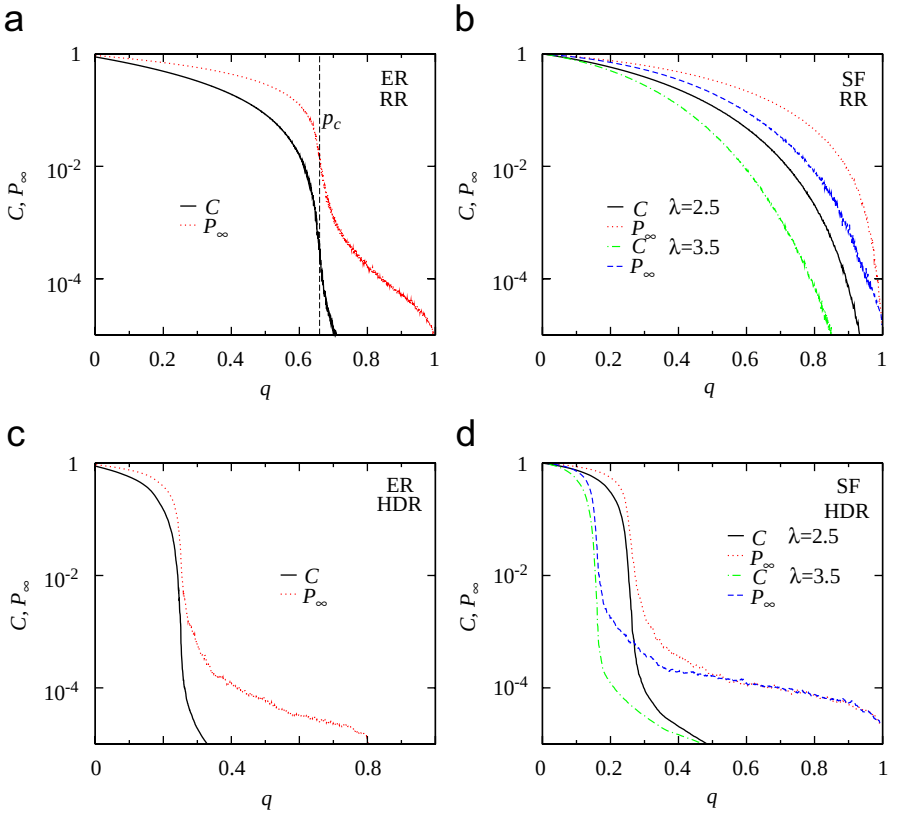

Fig. 1. The behavior of C and P_∞ versus q on ER and SF networks. For ER networks, $N = 200\,000$ and $\langle k\rangle = 3$. For SF networks, $N = 80\,000$. The graphs are (a) RR strategy on ER networks, (b) RR strategy on SF networks, (c) HDR strategy on ER networks and (d) HDR strategy on SF networks.

written as

$$C \equiv 1 - F \equiv \frac{\sum_{j=1}^{i} N_j(N_j - 1)}{N(N-1)} \approx \frac{N_\infty(N_\infty - 1)}{N(N-1)} \approx \frac{N_\infty^2}{N^2} \approx P_\infty^2. \tag{2}$$

Therefore, we expect 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 and $P_\infty \sim \ln(N)/N \to 0$ for large N [7]. Here significant variations between P_∞ and $C^{1/2}$ are expected, as indeed seen in Fig. 2.

3. 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_∞ versus $C^{1/2}$ for RR strategy in ER networks and for several values of p. As predicted by Eq. (2), the plot of P_∞ versus $C^{1/2}$ yields a linear relationship with slope equal to 1 when $p > p_c = 1/\langle k\rangle = \frac{1}{3}$. 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 $p_c = \frac{1}{3}$, the system approaches criticality and the one-to-one correspondence between $C^{1/2}$ and P_∞ is not so strong. This variation is attributed to the presence of clusters other than the infinite one, which influence C but not P_∞.

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 $C^{1/2}$ emerge 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

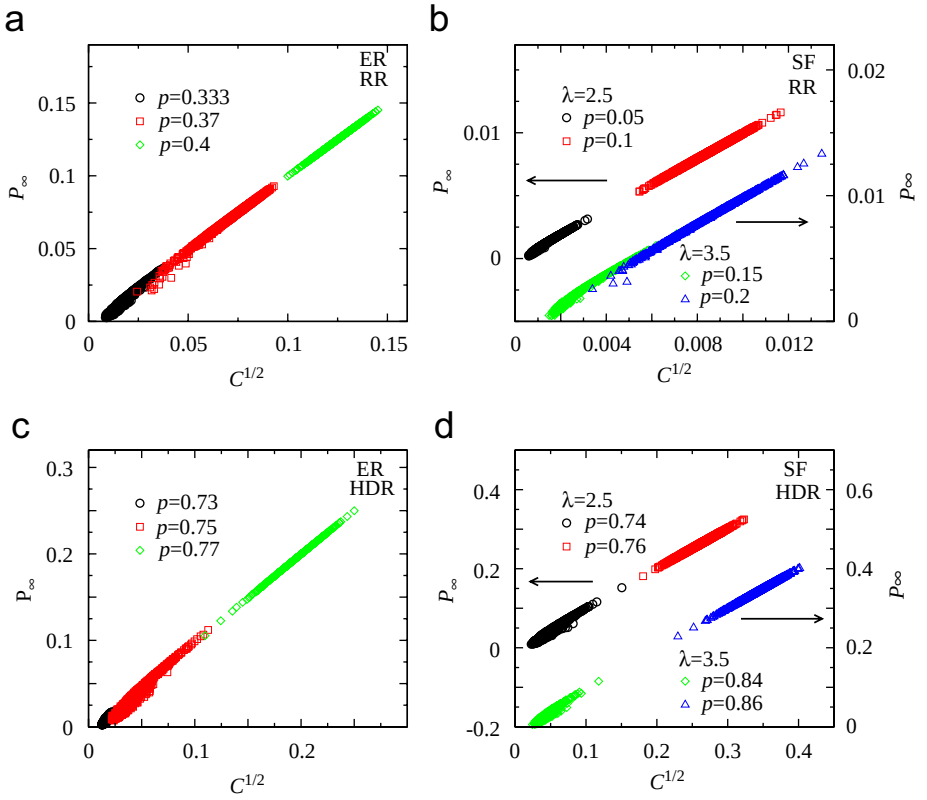

Fig. 2. Relationship between $C^{1/2}$ and P_∞ for ER and SF networks with system size N = 50 000. For ER networks, the average degree $\langle k\rangle = 3$, and for SF networks, $\lambda = 2.5$ and 3.5. The graphs are (a) RR strategy on ER networks, (b) RR strategy on SF networks, (c) HDR strategy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

5×10^{-4}. This observation supports that the SF networks with $\lambda < 3$ are quite robust in sustaining its infinite cluster against random removal [5]. Fig. 2(c) and (d) shows the results for HDR strategy in ER and SF networks. For this targeted strategy, 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 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 it can be clearly observed in the large variation at $P_\infty \approx 0.05$.

To further investigate the characteristics of the variation of C for a given P_∞, we calculate the probability distributions p(C) versus $C/\bar{C}$ for a given P_∞ where $\bar{C}$ is the average value of C and 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 the 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 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 contrary, for the HDR strategy, the broadness of p(C) for ER and SF networks are of the same order due to the fact that for HDR, p_c is also finite for $\lambda = 2.5$. This observation is consistent with the results shown in Fig. 2.

The broadness of p(C) for fixed P_∞ is quantitatively characterized by its standard deviation σ_C. Fig. 4(a) shows the relative standard deviation $\sigma_C/\bar{C}$ for the RR strategy in ER networks, where $\bar{C}$ is the average value of C. For increasing value of P_∞, the infinite cluster gradually gains control of the system and therefore $\sigma_C/\bar{C}$

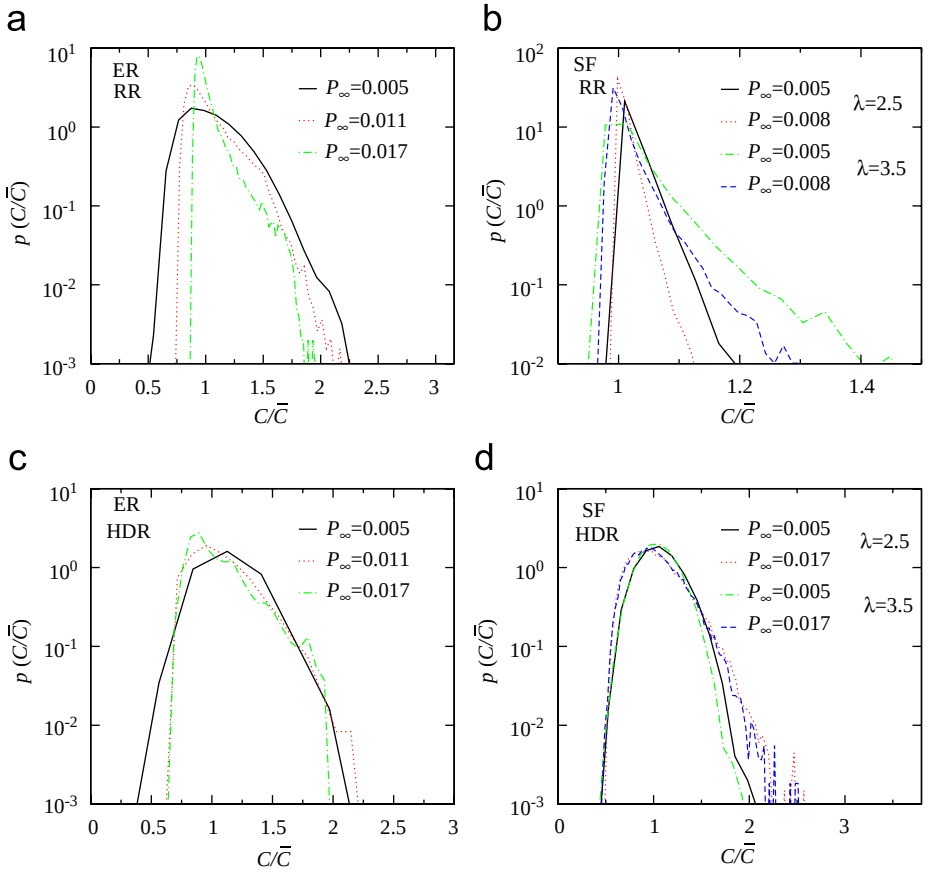

Fig. 3. Probability distributions $p(C/\bar{C})$ versus $C/\bar{C}$ for several values of P_∞ and for ER networks with $\langle k\rangle = 3$, $N = 200\,000$ and SF networks with $N = 80\,000$ and $\lambda = 2.5$ and 3.5. (a) RR strategy on ER networks, (b) RR strategy on SF networks, (c) HDR strategy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

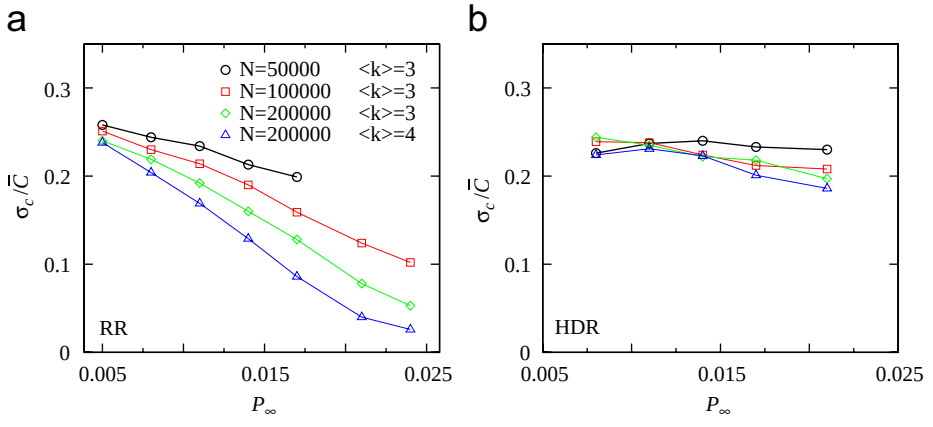

Fig. 4. The dependence of $\sigma_C/\bar{C}$ for (a) RR and (b) HDR on system size N and average degree $\langle k\rangle$ of ER networks.

becomes smaller. It can also be observed that σ_C is smaller for larger system sizes N and larger $\langle k\rangle$. The result for the HDR strategy is shown in Fig. 4(b) and one can observe that in this case, the relative standard deviation of C is much less sensitive to the value of P_∞, as expected in Fig. 3.

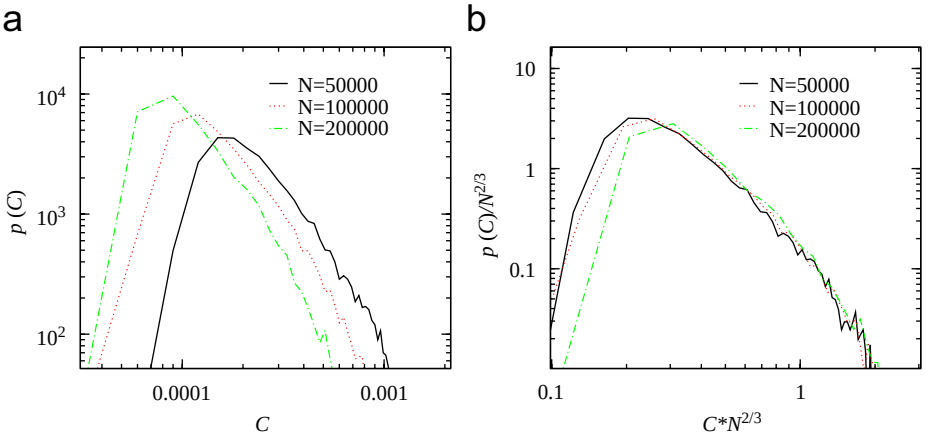

Fig. 5. The dependence of p(C) on the system size N with $p = p_c$ for (a) before scaling and (b) after scaling. Simulations are performed on ER networks with $\langle k \rangle = 3$.

Now we focus on the dependence of p(C) on the system size N at p_c (Fig. 5). From percolation theory and for ER under RR strategy, the infinite cluster size N_∞ at criticality behaves as [23,24]

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

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.

Fig. 5 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. 5a), and find a good collapse when plotted (Fig. 5b) using the scaling form of Eq. (5).

4. Real networks

The structure ER networks and SF networks that we have been studying so far are random and only determined by the degree distribution of the network. Research has shown that real networks often exhibit structural properties of importance for the percolation threshold such as high level of clustering, assortativity and fractality that these types of networks do not exhibit [12,25]. We therefore test our results about the correlation between C and P_∞ on a large real social network. The network we use is extracted from a data set obtained from Statistics Sweden [26] 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 couple without kids and other individuals sharing household are not registered in the data set as household. This kind of network has been shown to be of importance for the spreading of influenza [27] and is also likely to be of importance for the spread of information and rumors in society. The network consists of 310136 nodes (workplaces) and 906260 links (employees sharing the same households) and, as shown in Fig. 6(a), is approximately a SF network with $\lambda \approx 2.6$ and an exponential cut off. The network shows almost no degree-correlation (assortativity) preference (Fig. 6(b)). However, the workplace network clustering coefficient c is significantly higher than the random SF network with same λ and N (Fig. 6(c)). The average of c is 0.048 for the workplace network versus 3.2×10^{-4} for the random SF networks, which is consistent with the

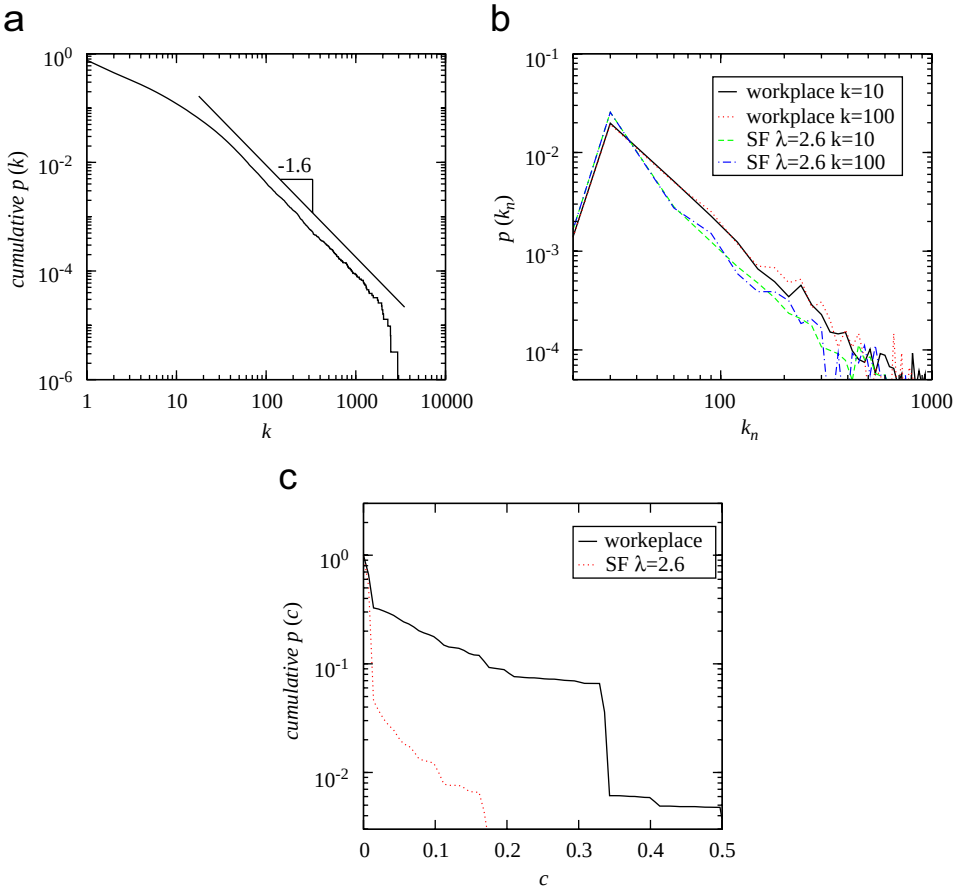

Fig. 6. Properties of the Swedish network of workplaces. (a) The cumulative degree distribution (showing $\lambda = 2.6$). (b) The distribution of k_n, the degree of the neighbors of nodes having degree k. (c) The cumulative distribution of clustering coefficient c. In (b) and (c) the distributions of random SF networks with the same λ and N are plotted for comparison.

earlier social network studies [28,29]. Fig. 7(a) and (b) shows simulation results for several values of p for P_∞ versus $C^{1/2}$. The curves are linear, similar to Fig. 2 for our model networks. Moreover, Fig. 7(c) and (d) shows that $C^{1/2}$ and P_∞ are almost identical above the criticality thresholdp_c for a typical configuration after either 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 P_∞ rapidly decreases to a very small value (below 10^{-5}), a plateau shows up in the curve of $C^{1/2}$ due to the influence of the small clusters.

5. 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 percolation theory. For p above criticality, C and P_∞ are highly correlated and $C \approx P_\infty^2$. Close to criticality, for $p \geqslant p_c$ 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 better precision 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. For an alternative measure of connectivity of networks see Ref. [30].

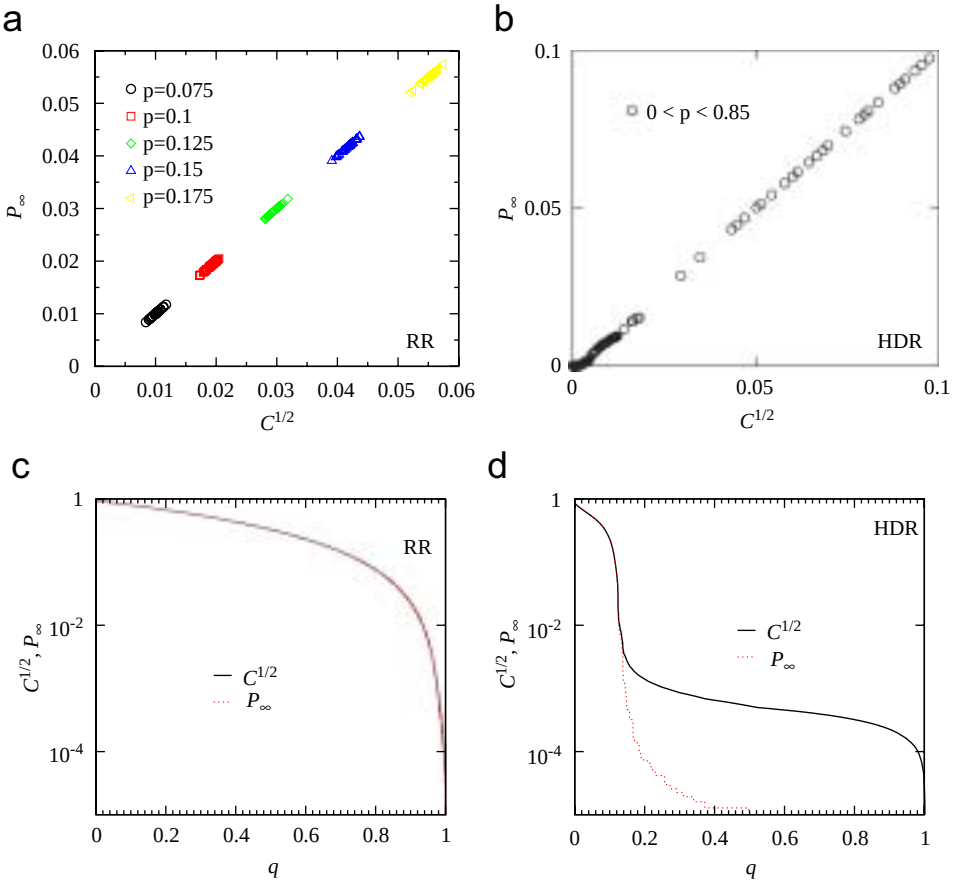

Fig. 7. P_∞ versus $C^{1/2}$ for (a) RR strategy and (b) HDR strategy and plot $C^{1/2}$, P_∞ versus q for (c) RR strategy and (d) HDR strategy for the Swedish network of workplaces with N = 310136 nodes.

Acknowledgments

We thank ONR, European NEST project DYSONET, and Israel Science Foundation for financial support.

References

[1] G. Paul, S. Sreenivasan, H.E. Stanley, Phys. Rev. E 72 (2005) 056130.
[2] R. Albert, H. Jeong, A.-L. Barabási, Nature, London 406 (2000) 378.
[3] M.E.J. Newman, Phys. Rev. Lett. 89 (2002) 208701.
[4] V. Paxon, IEEE/ACM Trans. Networking 5 (1997) 601.
[5] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Phys. Rev. Lett. 85 (2000) 4626.
[6] D.S. Callaway, M.E.J. Newmann, S.H. Strogatz, D.J. Watts, Phys. Rev. Lett. 85 (2000) 5468.
[7] R. Cohen, et al., Phys. Rev. Lett. 86 (2001) 3682.
[8] A. Valente, A. Sarkar, H.A. Stone, Phys. Rev. Lett. 92 (2004) 118702.
[9] G. Paul, T. Tanizawa, S. Havlin, H.E. Stanley, Eur. Phys. J. B 38 (2004) 187.
[10] F. Chung, L. Lu, Ann. Combinatorics 6 (2002) 125.
[11] Z. Burda, A. Krzywicki, Phys. Rev. E 67 (2003) 046118.
[12] C. Song, et al., Nature 433 (2005) 392.
[13] L.C. Freeman, The Development of Social Network Analysis: A Study in the Sociology of Science, Empirical, 2004.
[14] S. Wasserman, K. Faust, D. Iacobucci, M. Granovetter, Social Network Analysis: Methods and Applications, Cambridge, 1994.
[15] T. Tanizawa, G. Paul, R. Cohen, S. Havlin, H.E. Stanley, Phys. Rev. E 71 (2005) 047101.
[16] R. Pastor-Satorras, A. Vespignani, Phys. Rev. E 65 (2002) 036104.

[17] P. Holme, B.J. Kim, C.N. Yoon, S.K. Han, Phys. Rev. E 65 (2002) 056109.
[18] S.P. Borgatti, Comp. Math. Org. Theory 12 (2006) 21.
[19] Group of connected nodes known as ''component'' in the language of sociology.
[20] A. Bunde, S. Havlin, Fractals and Disordered Systems, Springer, Berlin, 1995.
[21] D. Stauffer, A. Aharony, Introduction to Percolation Theory, Taylor & Francis, London, 1994.
[22] R. Cohen, et al., Phys. Rev. E 66 (2002) 036113.
[23] P. Erdős, A. Rényi, Publ. Math. (Debrecen) 6 (1959) 290.
[24] R. Cohen, S. Havlin, D. Ben-Avraham, Structural properties of scale free networks, in: S. Bornholdt, H.G. Schuster (Eds.), Handbook of Graphs and Networks, vol. 4, Wiley-VCH, New York, 2002.
[25] M.E.J. Newman, SIAM Rev. 45 (2003) 167.
[26] ⟨WWW.SCB.SE⟩.
[27] C. Viboud, O.N. Bjørnstad, D.L. Smith, L. Simonsen, M.A. Miller, B.T. Grenfell, Science 312 (2006) 447.
[28] G. Csányi, B. Szendrői, Phys. Rev. E 69 (2004) 036131.
[29] K. Klemm, V.M. Eguíluz, Phys. Rev. E 65 (2002) 057102.
[30] E.J. Bienenstock, Balancing Efficiency and Vulnerability in Social Networks, this volume.