Bridgeness: a local index on edge significance in maintaining global connectivity

J. Stat. Mech. (2010) P10011

(http://iopscience.iop.org/1742-5468/2010/10/P10011)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Bridgeness: a local index on edge significance in maintaining global connectivity

Xue-Qi Cheng[1,2,5], Fu-Xin Ren[1], Hua-Wei Shen[1], Zi-Ke Zhang[3] and Tao Zhou[2,4]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
[2] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China
[3] Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland
[4] Department of Modern Physics, University of Science and Technology of China, Hefei 230026, People's Republic of China
E-mail: cxq@ict.ac.cn, renfuxin@software.ict.ac.cn, shenhuawei@software.ict.ac.cn, zhangzike@gmail.com and zhutou@ustc.edu

Received 10 August 2010
Accepted 17 September 2010
Published 7 October 2010

Online at stacks.iop.org/JSTAT/2010/P10011
doi:10.1088/1742-5468/2010/10/P10011

Abstract. Edges in a network can be divided into two kinds according to their different roles: some enhance the locality like the ones inside a cluster while others contribute to the global connectivity like the ones connecting two clusters. A recent study by Onnela *et al* uncovered the weak ties effects in mobile communication. In this paper, we provide complementary results on document networks, that is, the edges connecting less similar nodes in content are more significant in maintaining the global connectivity. We propose an index called *bridgeness* to quantify the edge significance in maintaining connectivity, which only depends on local information of the network topology. We compare the bridgeness with content similarity and some other structural indices according to an edge percolation process. Experimental results on document networks show that the bridgeness outperforms content similarity in characterizing the edge significance. Furthermore, extensive numerical results on disparate networks indicate that the bridgeness is also better than some well-known indices on edge

[5] Author to whom any correspondence should be addressed.

1742-5468/10/P10011+10$30.00

significance, including the Jaccard coefficient, degree product and betweenness centrality.

Keywords: random graphs, networks

ArXiv ePrint: 1005.2652

Contents

1. Introduction

Recently, the study of complex networks became a common focus in many branches of science [1]–[4]. Many measurements are used to characterize the role of a node in network structure and function, ranging from simple indices like degree and closeness [5] to complicated centralities like betweenness [6], PageRank score [7] and some random-walk-based indices [8]. In comparison, the study of the edge's role is less extensive. Actually, an edge may play an important role in enhancing the locality or be significant in maintaining the global connectivity. How to differentiate these two roles is an interesting and unsolved problem. A related issue is the famous *weak ties theory* [9], which tells us that the job opportunities and new ideas usually come from people with weak connections. Furthermore, the weak ties can enhance the spreading of rumor [10] and knowledge [11], keep the stability of biological functions [12] and improve the accuracy of network structure prediction [13]. Moreover, experimental and theoretical analyses on a large-scale mobile communication network showed that the weak ties play the leading role in maintaining the global connectivity [14] and they are important in the process of the emergence of social communities [15], which is a common topological characteristic of networks and has been extensively studied in the literature of network theory [16]–[18]. In [14], the strength of a tie is quantified by the time spent on communication, while in this paper we consider a similar problem on document networks where the tie strength is characterized by content similarity. Analogous to the observation of the social communication network, we find that the ties connecting less similar documents are more significant in maintaining global connectivity.

Generally speaking, the strength information, such as the communication time and content similarity, is not easy to be obtained. Therefore, to quantify the edge significance only making use of the information on observed topology is very valuable in practice. For

J. Stat. Mech. (2010) P10011

this purpose, we propose the index bridgeness. To our surprise, this index can better characterize the edge significance in maintaining the global connectivity than the content similarity in document networks. Furthermore, extensive numerical results on disparate networks indicate that the bridgeness is also better than some well-known structural indices on edge significance, including the Jaccard coefficient [19], degree product [20]–[23] and edge betweenness [16].

2. Weak ties phenomenon in document networks

Social networks usually exhibit two important phenomena, being respectively the homophily [24, 25] and weak ties effects [9]. Homophily tells us that connections are more likely to be formed between individuals with close interests, similar social statuses, common characteristics or shared activities. The well-known transitivity or triadic closure phenomenon [26] can be taken as a reflection for homophily. Homophily contributes to the formation of most connections in social networks. In contrast, some connections are formed between less similar individuals, which usually have weaker strength. The weak ties phenomenon refers to the fact that these connections may play crucial roles in maintaining the global connectivity and holding key functions, such as the flow of new information and knowledge.

Compared with social networks, the homophily and weak ties phenomenon are less studied for document networks. In document networks, each node corresponds to a document with textual content. Typical examples are the World Wide Web and scientific citation networks. Two documents are more likely to be connected if they are relevant, e.g.,webpages presenting the same topic and articles belonging to the same research area. The probability that two documents are connected increases with their content similarity [27] and the probability that three documents form a triangle increases with their trilateral content similarity [28, 29]. These indicate that the content similarity in document networks plays a similar role to the tie strength in social networks.

Inspired by Onnela *et al* [14], we quantify the weak ties phenomenon according to an edge percolation process. Generally speaking, if the weak ties phenomenon exists in terms of content similarity, the network will disintegrate much faster when we remove edges successively in ascending order of content similarity than in descending order. Two quantities are used to describe the percolation process. The first one is the fraction of nodes contained in the giant component, denoted by R_{GC}. A sudden decline of R_{GC} will be observed if the network disintegrates after the deletion of a certain fraction of edges. Another quantity is the so-called *normalized susceptibility*, defined as

$$\tilde{S} = \sum_{s<s_{\max}} \frac{n_s s^2}{N}, \tag{1}$$

where n_s is the number of components with size s, N is the size of the whole network and the sum runs over all components except the largest one. Considering $\tilde{S}$ as a function of the fraction of removed edges f, usually, an obvious peak can be observed that corresponds to the precise point at which the network disintegrates [30, 31].

We use the PNAS citation network to illustrate the weak ties phenomenon, where each node represents a paper published in the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS, http://www.pnas.org), and undirected

Table 1. Basic statistics of the four networks: the PNAS citation network, social network extracted from del.icio.us, political blog network and scientific collaboration network from astrophysicists. N and L denote the number of nodes and edges in the original network, while N_{CG} and L_{CG} are the number of nodes and edges in the giant component.

Networks/measures	N	L	N_{GC}	L_{GC}
PNAS	28 828	40 393	21 132	38 723
Del.icio.us	64 615	107 140	64 615	107 140
Political Blog	1 490	16 715	1 222	16 714
Astro Collaboration	16 706	121 251	14 845	119 652

edges represent the citations. The dataset contains all the papers from 1998 to 2007, and the basic statistics of this network is shown in table 1. We collect keywords presented in titles and abstracts of papers and construct a keyword vector space according to the standard procedures in information retrieval [32]. Here keywords refer to those words which are not stop words such as 'a', 'the', 'of', etc. The content of a paper is then represented as a keyword vector, $\vec{X}$, which gives the frequency of each keyword. The content similarity R between two papers, i and j, is thus defined as

$$R_{ij} = R_{ji} = \frac{\|\overrightarrow{X_i} \cdot \overrightarrow{X_j}\|}{\|\overrightarrow{X_i}\|\|\overrightarrow{X_j}\|}. \tag{2}$$

Figure 1 shows the edge percolation results on the PNAS citation network. As shown in figure 1(a), R_{GC} decays much faster when we first remove the less similar edges. Correspondingly, in figure 1(b), a sharp peak appears in the edge-removing process from the weakest to the strongest ones, indicating a percolation phase transition. While if we first remove stronger edges, no clear peak can be observed. Analogous to the report on mobile communication networks [14], this result strongly supports the weak ties phenomenon in the document network.

3. Bridgeness versus tie strength

As mentioned above, tie strength is a good indicator for the edge's role of maintaining global connectivity. However, the strength information is usually hard to be obtained. In social networks it requires personal information or historical activities, while in document networks it has to crawl the content and calculate the textual similarity. The processes are probably complicated, time-consuming or even infeasible. Therefore, an index depending only on topological information is of great advantage in practice. As we have mentioned, in social and document networks similar or relevant nodes are apt to connect to others and form local clusters. Clique is the simplest structure to describe the locality in a network [29]. A clique of size k is a fully connected subgraph with k nodes [33], and the clique size of a node x or an edge E is defined as the size of the maximum clique that contains this node or this edge [34, 35]. We found that in the social and document networks we observed, clique structures of different sizes are prevalent. Recently, González *et al* [36] have reported the presence of k-clique communities in school friendship networks from the Add Health database and the topology of a c-network formed by communities.

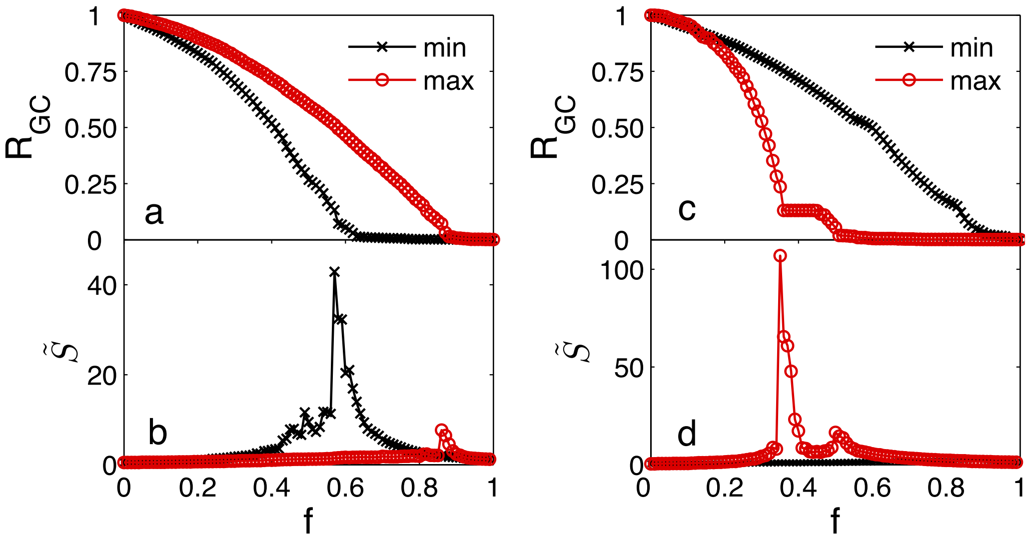

Figure 1. Edge percolation results on PNAS citation network. Plots (a) and (b) are for content similarity, while (c) and (d) are for bridgeness. In (a) and (b), the min- (max-) lines represent the processes where the edges are removed from the least (most) similar to the most (least) similar ones. In (c) and (d), the min- (max-) lines represent the processes where the edges with smaller (larger) bridgeness are removed earlier.

In his work nodes shared by different communities are critical in connecting the c-network. Here we focus on the roles of edges, i.e. edges in cliques mainly contribute to locality while those between cliques are important in connecting the network. Accordingly, we define the bridgeness of an edge as

$$B_E = \frac{\sqrt{S_x S_y}}{S_E}, \tag{3}$$

where x and y are the two endpoints of the edge E. For example, in figure 2, the clique sizes of x, y and z are $S_x = 5$, $S_y = 4$ and $S_z = 3$, and the clique sizes of E, E' and E'' are $S_E = S_{E'} = S_{E''} = 3$.

An interesting and surprising result can be observed from figure 3 that there is a negative correlation between bridgeness and content similarity, namely the weaker the content similarity between two papers is, the stronger its bridgeness will be. Combined with the conclusion in the previous section, we infer that edges with strong bridgeness will play a more important role in maintaining the global connectivity. Figure 1(c) shows that, if the edges with larger bridgeness are removed first (corresponding to the curve labeled by max), the network quickly splits into many pieces, while if we start removing the edges with smaller bridgeness (corresponding to the curve labeled by min), the network gradually shrinks. As shown in figure 1(d), a sharp peak is observed in the former case. In addition, the critical point (corresponding to the peak in $\tilde{S}(f)$) in figure 1(d) is remarkably smaller than the one in figure 1(b), indicating that the bridgeness can even better characterize the edge's significance in maintaining the global connectivity than the content similarity.

J. Stat. Mech. (2010) P10011

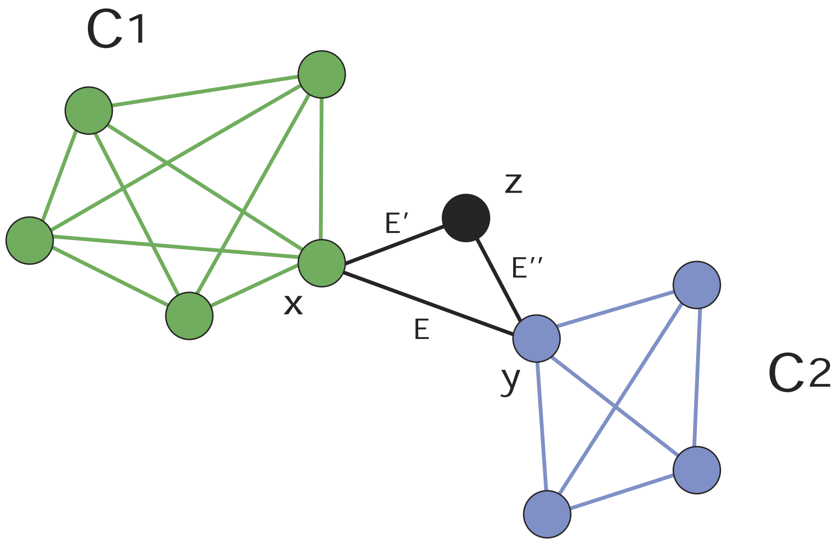

Figure 2. Illustration of the bridgeness. C_1 and C_2 are two cliques of size 5 and 4, and the edge E connecting them is of bridgeness 1.49.

4. Bridgeness versus other indices

In this section we compare the bridgeness with some other indices for edge significance, which are also dependent only on the topological structure. Three representative indices used for comparison are introduced as follows (for more information about how to quantify the edge significance, please see [37]–[39]).

- *Jaccard coefficient* [19]. The Jaccard coefficient is defined as

$$J_E = \frac{|U_x \cap U_y|}{|U_x \cup U_y|}, \tag{4}$$

where x and y are the two endpoints of the edge E and U_x is the set of x's neighboring nodes.

- *Degree product*. The degree product is defined as

$$D_E = k_x k_y, \tag{5}$$

where k_x is the degree of node x. An extended form, $(k_x k_y)^\alpha$, has been recently applied in the studies of biased percolation [21]–[23]. Notice that we only care about the order of the edges, so the values of α (if $\alpha > 0$) play no role in the results and there we choose $\alpha = 1$, the same as in [20].

- *Edge betweenness centrality* [16]. Edge betweenness centrality counts for the number of shortest paths between pairs of nodes passing through the edge as

$$C_E = \sum_{s \neq t} \frac{\sigma_{st}(E)}{\sigma_{st}}, \tag{6}$$

where σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(E)$ is the number of shortest paths from s to t that pass through edge E. Notice that this index is a global index whose computing complexity is remarkably higher than the three local indices—bridgeness, Jaccard coefficient and degree product.

J. Stat. Mech. (2010) P10011

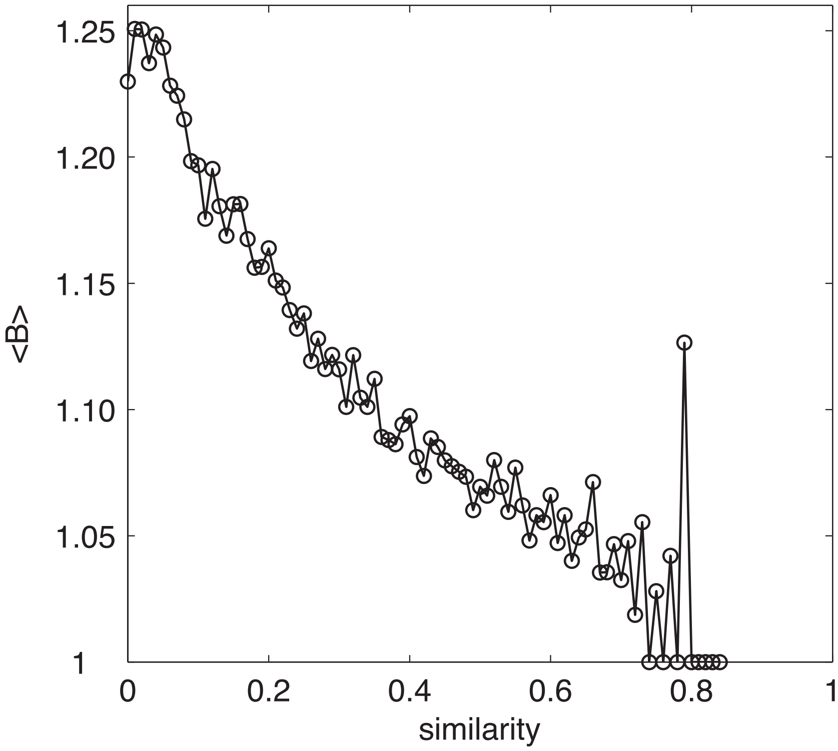

Figure 3. The relation between bridgeness and content similarity in the PNAS citation network. $\langle B \rangle$ is the average bridgeness value of edges with the same content similarity.

Empirical comparison is carried out on four networks. Besides the PNAS citation network, the three others are the social network of del.icio.us, the political blog network and the astrophysics collaboration network. Data were crawled in October 2009 from the famous online social bookmarking service del.icio.us, where a user i is considered to be a fan of another user j if i has seen some bookmarks collected by j, and two users are connected if they are fans of each other. The political blog network is formed by weblogs on US politics and hyperlinks between them [40]. The astrophysics collaboration network is comprised of the coauthorships between scientists posting preprints on the Astrophysics E-Print Archive from 1 January 1995 to 31 December 1999 [41]. The basic statistics of these networks is shown in table 1.

As shown in figure 4, for the PNAS citation network, edge betweenness centrality and bridgeness perform best; for the social network in del.icio.us, the degree product and bridgeness perform best; for the political blog network and the scientific collaboration network, bridgeness performs best. Therefore, one can conclude that bridgeness is remarkably better than the other three indices in characterizing the edge significance in maintaining the global connectivity. As a local index with light computational load, bridgeness is expected to be applied in practice.

5. Conclusion and discussion

In this paper, we investigate the weak ties phenomenon in document networks. Empirical analysis indicates that the weak ties, namely the edges connecting less similar nodes in

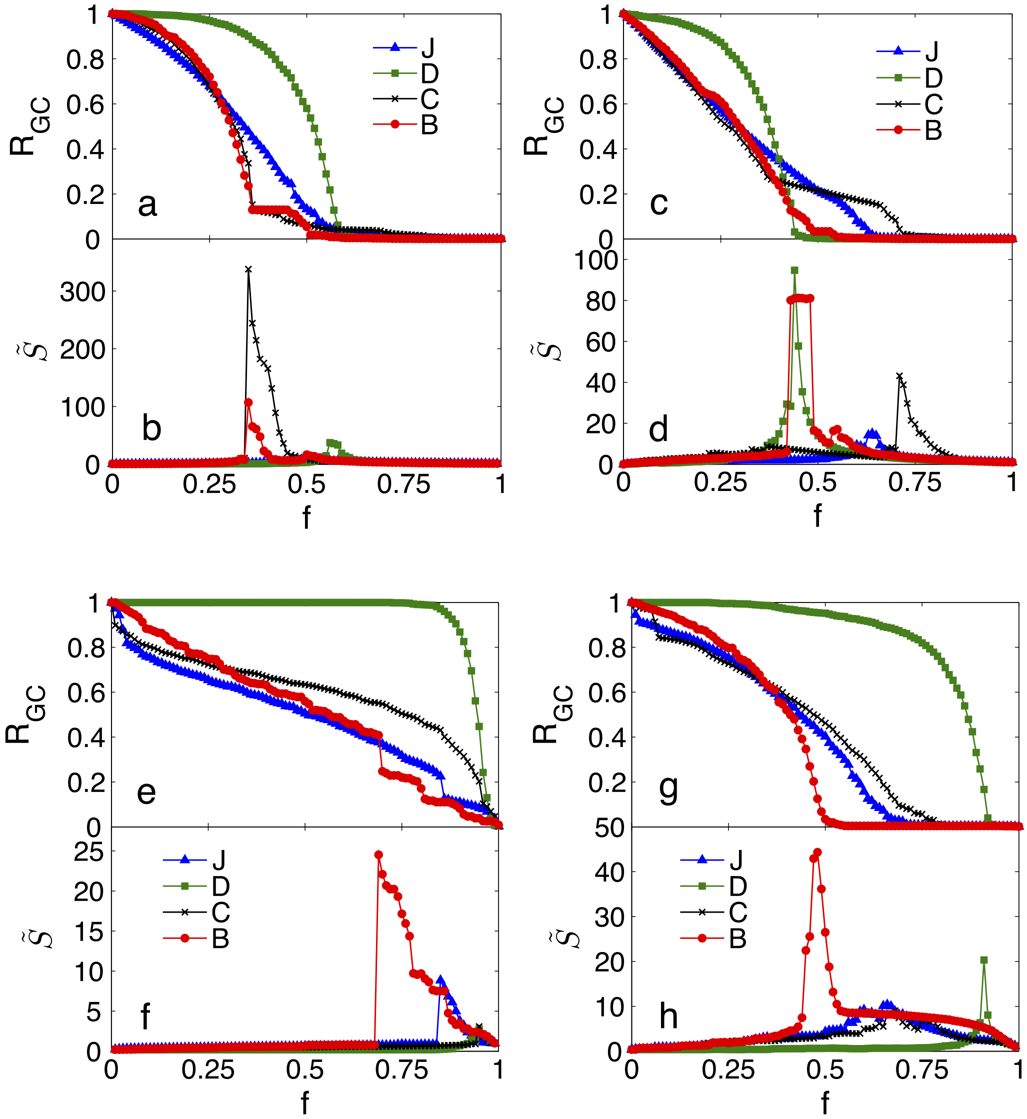

Figure 4. Edge percolation results on four real networks. Plots (a) and (b) are for the PNAS citation network, (c) and (d) for the social network of del.icio.us, (e) and (f) for the political blog network, and (g) and (h) for the astrophysics collaboration network. In each plot, four curves are corresponding to the four structural indices, J ($\triangle$), D ($\square$), C ($\times$) and B ($\bigcirc$). For the case of Jaccard coefficient, the edges are removed in ascending order, while for the other three cases, they are removed in descending order.

content, play a more significant role in maintaining the global connectivity. Inspired by the strong correlation between the existence of an edge and the strength of that edge [27, 29], we believe the edge significance in maintaining the global connectivity can be well characterized by some indices depending solely on the topological structure. Accordingly, we propose a local structural index, called bridgeness. Compared with both the content similarity and three well-known topological indices, bridgeness always leads to an earlier network disintegration in the edge percolation process, indicating that

J. Stat. Mech. (2010) P10011

bridgeness performs best in characterizing the edge significance. This will help us in some real-life applications such as controlling the spreading of diseases or rumor and withstanding targeted attacks on network infrastructures.

The percolation process is carried out on real-world networks and the results illustrate that the networks are fragile when removing edges with larger bridgeness. These networks are monopartite, i.e. nodes in these networks are of no distinction. It will be necessary to generalize the definition of bridgeness when facing other network configurations such as bipartite networks, as Lind *et al* did in [42]. Moreover, empirical analysis in more large-scale networks will also be part of our future work.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under grant nos. 60873245 and 60933005. Z-KZ acknowledges the Swiss National Science Foundation under grant no. 200020-121848. TZ acknowledges the National Natural Science Foundation of China under grant nos. 10635040 and 90924011.

References

[1] Albert R and Barabási A-L, 2002 *Rev. Mod. Phys.* **74** 47
[2] Newman M E J, 2003 *SIAM Rev.* **45** 167
[3] Boccaletti S, Latora V, Moreno Y, Chavez M and Huang D-U, 2006 *Phys. Rep.* **424** 175
[4] Newman M E J, Barabási A-L and Watts D J, 2006 *The Structure and Dynamics of Networks* (Princeton, NJ: Princeton University Press)
[5] Sabidussi G, 1966 *Psychometrika* **31** 581
[6] Freeman L C, 1977 *Sociometry* **40** 35
[7] Brin S and Page L, 1998 *Comput. Netw. ISDN Syst.* **30** 107
[8] Liu W and Lü L, 2010 *Europhys. Lett.* **89** 58007
[9] Granovetter M, 1973 *Am. J. Sociol.* **78** 1360
[10] Lai G and Wong O, 2002 *Soc. Netw.* **24** 49
[11] Levin D Z and Cross R, 2004 *Manag. Sci.* **50** 1477
[12] Csermely P, 2004 *Trends Biochem. Sci.* **29** 331
[13] Lü L and Zhou T, 2010 *Europhys. Lett.* **89** 18001
[14] Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J and Barabási A-L, 2007 *Proc. Nat. Acad. Sci.* **104** 7332
[15] Kumpula J M, Onnela J-P, Saramäki J, Kaski K and Kertész J, 2007 *Phys. Rev. Lett.* **99** 228701
[16] Girvan M and Newman M E J, 2002 *Proc. Nat. Acad. Sci.* **99** 7821
[17] Cheng X-Q and Shen H-W, 2010 *J. Stat. Mech.* P04024
[18] Shen H-W, Cheng X-Q and Fang B-X, 2010 *Phys. Rev.* E **82** 016114
[19] Jaccard P, 1901 *Bull. Soc. Vaudoise Sci. Nature* **37** 547
[20] Holme P, Kim B J, Yoon C N and Han S K, 2002 *Phys. Rev.* E **65** 056109
[21] Giuraniuc C V, Hatchett J P L, Indekeu J O, Leone M, Pérez Castillo I, Van Schaeybroeck B and Vanderzande C, 2005 *Phys. Rev. Lett.* **95** 098701
[22] Moreira A A, Andrade J S, Herrmann H J and Indekeu J O, 2009 *Phys. Rev. Lett.* **102** 018701
[23] Hooyberghs H, Van Schaeybroeck B, Moreira A A, Andrade J S Jr, Herrmann H J and Indekeu J O, 2010 *Phys. Rev.* E **81** 011102
[24] Lazarsfeld P and Merton R K, 1954 *Freedom and Control in Modern Society* (New York: Van Nostrand) pp 18–66
[25] McPherson J M, Smith-Lovin L and Cook J, 2001 *Annu. Rev. Sociol.* **27** 415
[26] Rapoport A, 1953 *Bull. Math. Biophys.* **15** 523
[27] Menczer F, 2004 *Proc. Nat. Acad. Sci.* **101** 5261
[28] Cheng X-Q, Ren F-X, Cao X-B and Ma J, 2007 *Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence* (Washington, DC: IEEE Press) pp 81–4
[29] Cheng X-Q, Ren F-X, Zhou S and Hu M-B, 2009 *New J. Phys.* **11** 033019
[30] Stauffer D and Aharony A, 1994 *Introduction to Percolation Theory, 2nd* (London: CRC Press)

[31] Bunde A and Havlin S, 1996 *Fractals and Disordered Systems, 2nd* (New York: Springer) p 51
[32] Salton G, 1989 *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Reading, MA: Addison-Wesley)
[33] Xiao W-K, Ren J, Qi F, Song Z-W, Zhu M-X, Yang H-F, Jin H-J, Wang B-H and Zhou T, 2007 *Phys. Rev.* E **76** 037102
[34] Shen H-W, Cheng X-Q, Cai K and Hu M-B, 2009 *Physica* A **388** 1706
[35] Shen H-W, Cheng X-Q and Guo J-F, 2009 *J. Stat. Mech.* P07042
[36] González M C, Herrmann H J, Kertész J and Vicsek T, 2007 *Physica* A **379** 307
[37] Liben-Nowell D and Kleinberg J, 2007 *J. Am. Soc. Inform. Sci. Technol.* **58** 1019
[38] Fouss F, Pirotte A, Renders J-M and Saerens M, 2007 *IEEE Trans. Knowl. Data Eng.* **19** 355
[39] Zhou T, Lü L and Zhang Y-C, 2009 *Eur. Phys. J.* B **71** 623
[40] Adamic L A and Glance N, 2005 *Proc. WWW'2005 Workshop on the Weblogging Ecosystem* (New York: ACM Press) p 43
[41] Newman M E J, 2001 *Proc. Nat. Acad. Sci.* **98** 404
[42] Lind P G, González M C and Herrmann H J, 2005 *Phys. Rev.* E **72** 056127