Published in "Europhysics Letters 89 (5): 58007, 2010"
which should be cited to refer to this work.

Link prediction based on local random walk

Weiping Liu and Linyuan Lü[a]

Department of Physics, University of Fribourg - Chemin du Musée 3, CH-1700 Fribourg, Switzerland

PACS 89.20.Ff – Computer science and technology
PACS 89.75.Hc – Networks and genealogical trees
PACS 89.65.-s – Social and economic systems

Abstract – The problem of missing link prediction in complex networks has attracted much attention recently. Two difficulties in link prediction are the sparsity and huge size of the target networks. Therefore, to design an efficient and effective method is of both theoretical interest and practical significance. In this letter, we proposed a method based on local random walk, which can give competitively good or even better prediction than other random-walk–based methods while having a much lower computational complexity.

http://doc.rero.ch

Introduction. – Recently, the problem of missing link prediction in complex networks has attracted much attention [1–3]. Link prediction aims at estimating the likelihood of the existence of a link between two nodes. For some networks, especially biological networks such as protein-protein interaction networks, metabolic networks and food webs, the discovery of links is costly in the laboratory or the field, and thus the current knowledge of those networks is substantially incomplete [4,5]. Instead of blindly checking all the possible links, predictions based on the observed links and focusing on those links which are most likely to exist can sharply reduce the experimental costs if the predictions are accurate enough [1]. For some others like online friendship networks, very likely but not yet existent links can be suggested to users as recommendations of promising friendships, which can help users in finding new friends and thus enhance their loyalties to the web sites. In addition, the link prediction algorithms can be applied to solve the classification problem in partially labeled networks [6], such as to distinguish the research areas of scientific publications.

Commonly, two nodes are more likely to be connected if they are more similar, where a latent assumption is that the link itself indicates a similarity between the two endpoints and this similarity can be transferred through the links. In this case, the similarity indices are used to quantify the structural equivalence (see, for example, the Leicht-Holme-Newman index [7] and the transferring similarity [8]). However, in some networks the two endpoints of one link are not essentially similar, such as the sexual network [9] and the word co-occurrence networks [10]. In these cases, we can use the regular equivalence (see ref. [11] for regular equivalence), which indicates that two nodes are similar if they have connected to similar nodes. How to predict missing links in such kind of networks is still an open problem to us. Our study focuses on structure equivalence.

Node similarity can be defined by the essential attributes of nodes. For example, if two persons have the same age, sex and job, we can say that they are similar. Another group of similarities is based only on the network structure. An introduction and a comparison of some similarity indices are presented in ref. [2], where the *Common Neighbours* [12], *Jaccard coefficient* [13], *Adamic-Adar Index* [14] and *Preferential Attachment* [15] are the node-dependent indices that require only the information about node degree and the nearest neighborhood, while the *Katz Index* [16], *Hitting Time* [17], *Commute Time* [18], *Rooted PageRank* [19], *SimRank* [20] and *Blondel Index* [21] belong to the path-dependent indices that ask for global knowledge of the network topology. In ref. [22], Zhou *et al.* proposed two new local indices, *Resource Allocation index* and *Local Path index*. Empirical results show that these two indices perform better than nine other known local indices (see ref. [22] for details). In particular, the local path index, asking for a little bit more information than common neighbours, provides competitively accurate prediction compared with the global Katz index [23]. Lü and Zhou [24] studied the link prediction problem in weighted networks, and found that the weak links may play a more important role than strong links (see the well-known weak-ties theory [25] in

[a]E-mail: linyuan.lue@unifr.ch

http://doc.rero.ch

social sciences). Besides, Clauset *et al.* [1] proposed an algorithm based on the hierarchical network structure, which gives good predictions for the networks with hierarchical structures, such as grassland species food web and terrorist association network. In real applications, similarity indices only exploiting local information are more efficient than those based on global information, for their lower computational complexity. However, due to the insufficient information, local indices may be less effective for their lower prediction accuracy. To design an efficient and effective algorithm is a main challenge in link prediction.

In this letter, we define the node similarity based on local random walk, which has lower computational complexity compared with other random-walk–based similarity indices, such as average commute time (ACT) and random walk with restart (RWR). We compare our method with five representative indices, including three local ones (common neighbours, resource allocation and local path indices) and two global ones (ACT and RWR), as well as the hierarchical structure method. Empirical results on five real networks show that our method performs best.

Similarity based on local random walk. – Consider an undirected simple network $G(V,E)$, where V is the set of nodes and E is the set of links. Multiple links and self-connections are not allowed. For each pair of nodes, $x, y \in V$, we assign a score, s_{xy}. In this letter, we adopt the simplest framework, that is, to directly set the similarity as the score. All the nonexistent links are sorted in descending order according to their scores, and the links at the top are most likely to exist.

Random walk is a Markov chain that describes the sequence of nodes visited by a random walker [26,27]. This process can be described by the transition probability matrix, P, with $P_{xy} = a_{xy}/k_x$ presenting the probability that a random walker staying at node x will walk to y in the next step, where a_{xy} equals 1 if node x and node y are connected, 0 otherwise, and k_x denotes the degree of node x. Given a random walker starting from node x, denoting by $\pi_{xy}(t)$ the probability that this walker locates at node y after t steps, we have

$$\vec{\pi}_x(t) = P^T \vec{\pi}_x(t-1), \tag{1}$$

where $\vec{\pi}_x(0)$ is an $N \times 1$ vector with the x-th element equal to 1 and others to 0, and T is the matrix transposition. The initial resource is usually assigned according to the importance of nodes [28]. Here, we simply set the initial resource of node x proportional to its degree k_x. Then, after normalization the similarity between node x and node y is

$$s_{xy}^{LRW}(t) = \frac{k_x}{2|E|} \cdot \pi_{xy}(t) + \frac{k_y}{2|E|} \cdot \pi_{yx}(t), \tag{2}$$

where $|E|$ is the number of links in the network. It is obvious that $s_{xy} = s_{yx}$. Note that here we only focus on the few-step random walk instead of the stationary state that can be characterized by the eigenvector centrality [29,30]. In the stationary state, we have $\pi_{xy} = \frac{k_y}{2|E|}$, and thus according to eq. (2), $s_{xy} = \frac{k_x \cdot k_y}{2|E|^2}$, which is equivalent to the preferential attachment index (*i.e.*, $k_x \cdot k_y$) that has been discussed in ref. [22].

One difficulty with all random-walk–based similarity measures is their sensitive dependence to parts of the network far away from target nodes [2]. For example, in a random walk from x to y, the walker has a certain probability to go too far away from both x and y although they may be close to each other. This may lead to a low prediction accuracy since many real-world networks have high clustering coefficients, which tend to cause random walkers to circulate locally rather than escape to other, more distant, parts of the network. A possible way to counteract this dependence is to continuously release the walkers at the starting point, resulting in a higher similarity between the target node and the nodes nearby. By superposing the contribution of each walker (walkers move independently), we obtain the similarity index

$$s_{xy}^{SRW}(t) = \sum_{l=1}^{t} s_{xy}^{LRW}(l), \tag{3}$$

where SRW is the abbreviation for superposed random walk.

Metrics. – To test the algorithm's accuracy, the observed links, E, are randomly divided into two parts: the training set, E^T, and the probe set, E^P. Clearly, $E = E^T \cup E^P$ and $E^T \cap E^P = \emptyset$. We use two standard metrics, AUC[1] [33] and precision [34], to quantify the accuracy of prediction algorithms. The former evaluates the overall ranking resulted from the algorithm, while the later focuses on the top-L candidates. In the present case, AUC can be interpreted as the probability that a randomly chosen missing link (a link in E^P) is given a higher score than a randomly chosen nonexistent link (a link in $U \backslash E$, where U denotes the universal set). In the implementation, among n independent comparisons, if there are n' times the missing link having a higher score and n'' times being of the same score, we have

$$\mathrm{AUC} = \frac{n' + 0.5n''}{n}. \tag{4}$$

If all the scores are generated from an independent and identical distribution, the AUC should be about 0.5. Therefore, the degree to which the AUC exceeds 0.5 indicates how much better the algorithm performs than pure chance. Precision is defined as the ratio of relevant

[1] Actually, AUC is formally equivalent to the *Wilcoxon rank-sum test* [31] and *Mann-Whitney U statistical test* [32]. It is a non-parametric test for assessing whether two independent samples of observations come from the same distribution. Note that, a latent assumption in AUC metric is the independence of the existence of each link, which may be not the case in the real world.

Table 1: Basic topological features of the giant components of the example networks. N and $|E|$ are the total numbers of nodes and links, respectively. $\langle k\rangle$ is the average degree of the network. $\langle d\rangle$ is the average shortest distance between node pairs. C and r are the clustering coefficient [35] and assortative coefficient [36], respectively. H is the degree heterogeneity, defined as $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$.

Networks	N	$\lvert E\rvert$	$\langle k\rangle$	$\langle d\rangle$	C	r	H
USAir	332	2126	12.807	2.46	0.749	−0.208	3.464
NetScience	379	941	4.823	4.93	0.798	−0.082	1.663
Power	4941	6594	2.669	15.87	0.107	0.003	1.450
Yeast	2375	11693	9.847	4.59	0.388	0.454	3.476
C. elegans	297	2148	14.456	2.46	0.308	−0.163	1.801

items to the number of selected items. In our case, to calculate precision we need to rank all the nonexistent links in decreasing order according to their scores. Then we focus on the top-L (here $L = 100$) links. If there are l links successfully predicted (*i.e.*, in the probe set), then

$$\text{Precision} = \frac{l}{L}. \tag{5}$$

Clearly, a higher value of precision means a higher prediction accuracy.

Data. – We consider five representative networks drawn from disparate fields: i) USAir: The network of the US air transportation system, which contains 332 airports and 2126 airlines. ii) NetScience: A network of coauthorships between scientists who are themselves publishing on the topic of network science [37]. This network contains 1589 scientists, 128 of which are isolated. In fact, it consists 268 components, and the size of the giant component is only 379. iii) Power Grid: An electrical power grid of the western US [35], with nodes representing generators, transformers and substations, and edges corresponding to the high-voltage transmission lines between them. iv) Yeast: A protein-protein interaction network of yeast containing 2617 proteins and 11855 interactions [38]. Although this network is not well connected (it contains 92 components), most of the nodes belong to the giant component, whose size is 2375. v) *C. elegans*: The neural network of the nematode worm *C. elegans*, in which an edge joins two neurons if they are connected by either a synapse or a gap junction [35].

In this letter, we only consider the giant component, because the similarity indices based on local random walk, as well as those well-known indices (except the preferential attachment index) reported in refs. [2,22], will give zero score to a pair of nodes located in two disconnected components. This implies that if a network is unconnected, we actually predict the links in each component separately, and any probe link connecting two components cannot be predicted. Therefore we need to make sure that the training set represents a connected network. Actually, each time before moving a link to the probe set, we first check if this removal will make the training network disconnected. Table 1 summarizes the basic topological features of the giant components of those networks.

Results and discussion. – We compare the LRW index and SRW index with other five similarity indices, including three local ones: Common Neighbour (CN), Resource Allocation index (RA) and Local Path index (LP), and two global ones: Average Commute Time (ACT), Random Walk with Restart (RWR), as well as the Hierarchical Structure method (HSM). A brief introduction of each algorithm is shown as follow:

i) CN: For a node x, let $\Gamma(x)$ denote the set of neighbours of x. By common sense, two nodes, x and y, are more likely to have a link if they have more common neighbours. The simplest measure of this neighbourhood overlap is the directed count, namely

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|. \tag{6}$$

Actually, CN is a kind of localized version of the Katz index which directly sums over the collection of all paths and exponentially damped by length to give the short paths more weights.

ii) RA: Consider a pair of nodes, x and y, which are not directly connected. The node x can send some resource to y, with their common neighbours playing the role of transmitters. Assuming that each transmitter has a unit of resource and will equally distribute to all its neighbours, the similarity between x and y, defined as the amount of resource y received from x, is [22]

$$s_{xy}^{RA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k_z}. \tag{7}$$

Clearly, this measure is symmetric, namely $s_{xy} = s_{yx}$. Note that this index is equivalent to the two-step LRW, where

$$\pi_{xy}(t=2) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k_x \cdot k_z}. \tag{8}$$

Former analysis showed that RA performs best among all the common-neighbour-based indices in the USAir network, NetScience network, Power Grid network, Yeast network and Router network of the Internet [22].

http://doc.rero.ch

Table 2: Comparison of algorithms' accuracy quantified by AUC and Precision. For each network, the training set contains 90% of the known links. Each number is obtained by averaging over 1000 implementations with independently random divisions of training set and probe set. We set the parameters $\varepsilon = 10^{-3}$ in LP and $c = 0.9$ in RWR. The numbers inside the brackets denote the optimal step of LRW and SRW indices. For example, 0.972(2) means the optimal AUC is obtained at the second step of LRW. The highest accuracy in each line is emphasized by boldface. For HSM we generate 5000 samples of dendrograms for each implementation.

AUC	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.954	0.972	0.952	0.901	0.977	0.904	0.972(2)	**0.978**(3)
NetScience	0.978	0.983	0.986	0.934	**0.993**	0.930	0.989(4)	0.992(3)
Power	0.626	0.626	0.697	0.895	0.760	0.503	0.953(16)	**0.963**(16)
Yeast	0.915	0.916	0.970	0.900	0.978	0.672	0.974(7)	**0.980**(8)
C. elegans	0.849	0.871	0.867	0.747	0.889	0.808	0.899(3)	**0.906**(3)
Precision	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.59	0.64	0.61	0.49	0.65	0.28	0.64(3)	**0.67**(3)
NetScience	0.26	0.54	0.30	0.19	**0.55**	0.25	0.54(2)	0.54(2)
Power	0.11	0.08	**0.13**	0.08	0.09	0.00	0.08(2)	0.11(3)
Yeast	0.67	0.49	0.68	0.57	0.52	0.84	**0.86**(3)	0.73(9)
C. elegans	0.12	0.13	**0.14**	0.07	0.13	0.08	**0.14**(3)	**0.14**(3)

http://doc.rero.ch

iii) LP: This index takes consideration of local paths, with wider horizon than CN. It is defined as [23]

$$S^{LP} = A^2 + \epsilon A^3, \tag{9}$$

where S denotes the similarity matrix, A is the adjacency matrix and ϵ is a free parameter. Clearly, this measure degenerates to CN when $\epsilon = 0$. References [22,23] show that LP is a good trade-off between effectiveness and efficiency.

iv) ACT: Denote by $m(x, y)$ the average number of steps required by a random walker starting from node x to reach node y, the average commute time between x and y is $n(x, y) = m(x, y) + m(y, x)$, which can be computed in terms of the Pseudoinverse of the Laplacian matrix L^+ (see footnote [2]), as [39]

$$n(x, y) = |E| \cdot (l^+_{xx} + l^+_{yy} - 2l^+_{xy}), \tag{10}$$

where l^+_{xy} denotes the corresponding entry in L^+. Assume that two nodes are considered to be more similar if they have a smaller average commute time, then the similarity between the nodes x and y can be defined as the reciprocal of $n(x, y)$, namely (the constant factor $|E|$ is removed)

$$s^{ACT}_{xy} = \frac{1}{l^+_{xx} + l^+_{yy} - 2l^+_{xy}}. \tag{11}$$

v) RWR: This index is a direct application of the PageRank algorithm [19]. Considering a random walker starting from node x, who will iteratively move to a random neighbour with probability c and return to node x with probability $1 - c$, and denoting by q_{xy} the probability this walker locates at node y in the steady state, then we have

$$\vec{q_x} = cP^T\vec{q_x} + (1 - c)\vec{e_x}, \tag{12}$$

where $\vec{e_x}$ is an $N \times 1$ vector with the x-th element equal to 1 and others to 0. The solution is straightforward, as

$$\vec{q_x} = (1 - c)(I - cP^T)^{-1}\vec{e_x}. \tag{13}$$

Accordingly, the RWR index is defined as

$$s^{RWR}_{xy} = q_{xy} + q_{yx}. \tag{14}$$

vi) HSM: The hierarchical structure of a network can be represented by a dendrogram with N leaves and $N - 1$ internal nodes. Each internal node r is associated with a probability p_r and the connecting probability of a pair of nodes is equal to p_m where m is the lowest common ancestor of these two nodes. To predict missing links with this method we first sample a large number of dendrograms with probability proportional to their likelihood. And then calculate the mean connecting probability $\langle p_{ij} \rangle$ by averaging the corresponding probability p_{ij} over all sampled dendrograms. A higher $\langle p_{ij} \rangle$ indicates a higher probability that nodes i and j are connected [1].

The results of these eight methods on five real networks are shown in table 2. For each network, the training set contains 90% of the known links. Generally speaking, the global indices perform better than the local ones. And our proposed indices, LRW and SRW, can give overall better predictions than the other methods for both AUC and precision. Compared with LRW index, the SRW index can lead to an even higher accuracy. The dependence of accuracy on the proportion of training set, labeled by p,

[2] $L = D - A$, where D is the degree matrix with $D_{ij} = \delta_{ij}k_i$.

http://doc.rero.ch

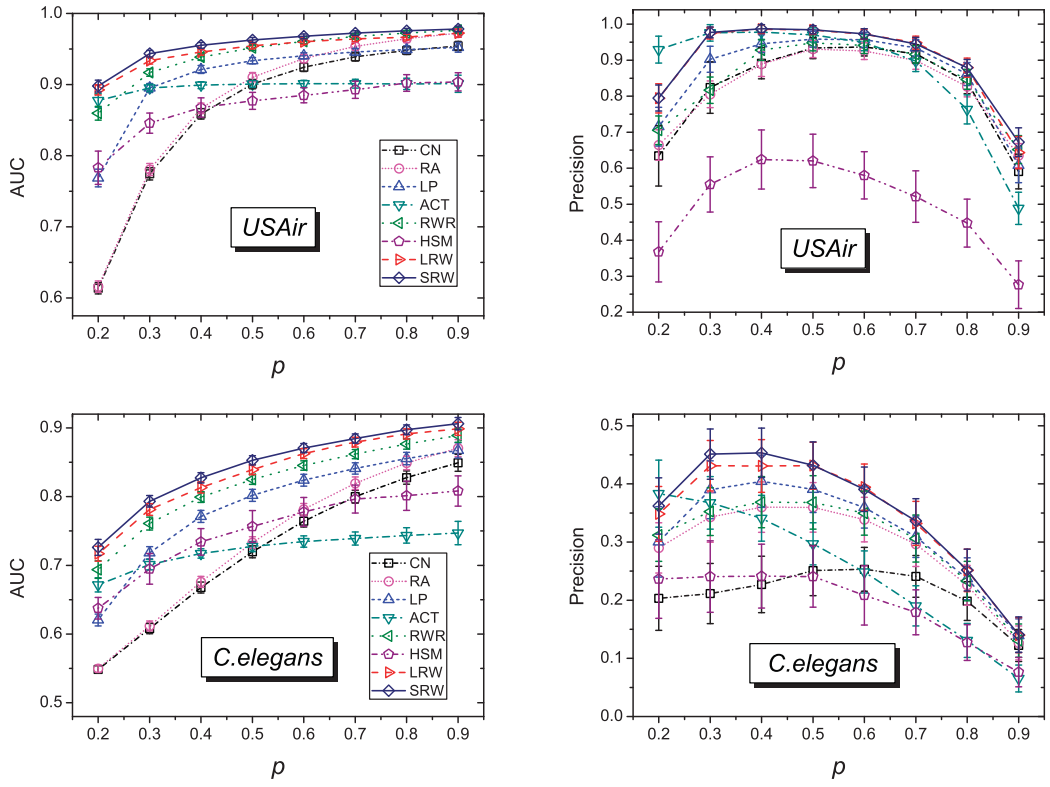

Fig. 1: (Color online) Dependence of AUC and Precision on the size of training set, denoted by p, in USAir and *C. elegans*. Each number is obtained by averaging over 1000 implementations with independently random divisions of the training set and probe set. For HSM we generate 5000 samples of dendrograms for each implementation.

in USAir network and *C. elegans* network[3] is shown in fig. 1. The results indicate that the advantages of LRW index and SRW index are not sensitive to the density of the network.

Interestingly, when predicting by the LRW index, as shown in fig. 2, we find a positive correlation between the optimal step and the average shortest distance. For example, $\langle d\rangle$ of USAir and *C. elegans* are very small, no more than 3, their optimal steps are also small: 2 and 3, respectively, in the case of $p = 0.9$. However, in the power grid with $\langle d\rangle \approx 16$, its AUC keeps increasing at the beginning and reaches a near optimum at step 16, where one more step leads to only 0.2% improvement. We also find that with the decreasing of p, the optimal step increases. This is because the removal of links to the probe set will increase $\langle d\rangle$, as shown in the insets of fig. 2. This result provides a possible way to choose the walking step, as a free parameter, before predicting.

Besides high accuracy, the low computation complexity is another important concern in the design of algorithms. Generally speaking, the global indices have a higher complexity than the local indices. As we known, the time complexity in calculating the inverse or pseudoinverse of an $N \times N$ matrix is $O(N^3)$, while the time complexity of n-step LRW (or SRW) is approximately $O(N\langle k\rangle^n)$. Since in most networks $\langle k\rangle$ is much smaller than N, LRW and SRW run much faster than ACT and RWR. This advantage is prominent especially in the huge-size (*i.e.* large N) and sparse (*i.e.* small $\langle k\rangle$) networks. For example, LRW for power grid is thousands of times faster than ACT, even when $n = 10$. In HSM, the process to sample a dendrogram asks for $O(N^2)$ steps of the Markov chain [1], and in the worse case, it takes exponential time [40]. Each step consumes a certain time to do some random selections. In addition, to predict the missing links, a large number of dendrograms are acquired. In this letter, we sample 5000 dendrograms for each implementation. Therefore, the time complexity of HSM is relatively high. It can handle networks with up to a few thousand nodes in a reasonable time, while LRW and SRW are able to handle networks with tens of thousands of nodes. Note that, although ACT, RWR and HSM have a higher time complexity, they provide much more information beyond link prediction. For example, the HSM algorithm can be used to uncover the hierarchical organization of real networks.

Conclusion. – In this letter, we proposed two similarity indices for link prediction based on local random walk. We compared our methods with six well-known methods on five real networks. The results show that our methods

[3]In order to ensure the training set is connected, the edges should be no less than $N-1$. Therefore, only USAir and *C. elegans* can be successfully divided to a connected training set containing only 20% of the known links. This is also the reason why fig. 1 only shows these two networks.

http://doc.rero.ch

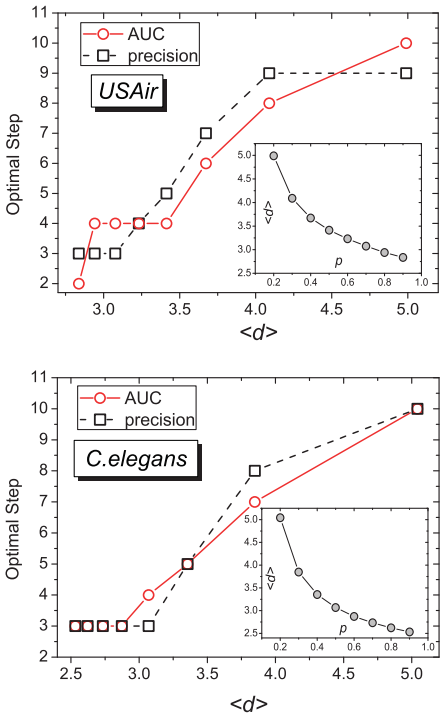

Fig. 2: (Color online) A positive correlation between the average shortest distance, $\langle d \rangle$, and the optimal step of the LRW method. The eight points from left to right correspond to the cases with p from 90% to 20%, respectively. The insets show the dependence of $\langle d \rangle$ on the size of the training set.

can give remarkably better prediction than the three local similarity indices. When comparing with the three global methods, ours can give slightly better prediction with a lower computational complexity.

* * *

We acknowledge V. Batageli and A. Mrvar for Pajek Datasets, A. Clauset for the HSM code, T. Zhou and J.-G. Liu for their helpful suggestions. This work is partially supported by the Swiss National Science Foundation (200020-121848), the Future and Emerging Technologies programmes of the European Commission FP7-COSI-ICT (QLectives Project, 231200) and the National Natural Science Foundation of China (60973069).

REFERENCES

[1] Clauset A., Moore C. and Newman M. E. J., *Nature*, **453** (2008) 98.
[2] Liben-Nowell D. and Kleinberg J., *J. Am. Soc. Inf. Sci. Technol.*, **58** (2007) 1019.
[3] Getoor L. and Diehl C. P., in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York) 2005.
[4] Martinez N. D., Hawkins B. A., Dawah H. A. and Feifarek B. P., *Ecology*, **80** (1999) 1044.
[5] Sprinzak E., Sattath S. and Margalit H., *J. Mol. Biol.*, **327** (2003) 919.
[6] Gallagher B., Tong H., Eliassi-Rad T. and Falousos C., in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York) 2008.
[7] Leicht E. A., Holme P. and Newman M. E., *Phys. Rev. E*, **73** (2006) 026120.
[8] Sun D., Zhou T., Liu J.-G., Liu R.-R., Jia C.-X. and Wang B.-H., *Phys. Rev. E*, **80** (2009) 017101.
[9] Liljeros F., Edling C. R., Amaral L. A. N., Stanley H. E. and Åberg Y., *Nature*, **411** (2001) 907.
[10] i Cancho R. F. and Solé R. V., *Proc. R. Soc. London, Ser. B*, **268** (2001) 2261.
[11] White D. R. and Reitz K. P., *Soc. Netw.*, **5** (1983) 193.
[12] Lorrain F. and White H. C., *J. Math. Sociol.*, **1** (1971) 49.
[13] Jaccard P., *Bull. Soc. Vaudoise Sci. Nat.*, **37** (1901) 547.
[14] Adamic L. A. and Adar E., *Soc. Netw.*, **25** (2003) 211.
[15] Barabási A.-L. and Albert R., *Science*, **286** (1999) 509.
[16] Katz L., *Psychmetrika*, **18** (1953) 39.
[17] Gobel F. and Jagers A., *Stoch. Processes Appl.*, **2** (1974) 311.
[18] Fouss F., Pirotte A., Renders J.-M. and Saerens M., *IEEE Trans. Knowl. Data Eng.*, **19** (2007) 355.
[19] Brin S. and Page L., *Comput. Netw. ISDN Syst.*, **30** (1998) 107.
[20] Jeh G. and Widom J., in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York) 2002.
[21] Blondel V. D., Gajardo A., Heymans M., Senellart P. and Dooren P. V., *SIAM Rev.*, **46** (2004) 647.
[22] Zhou T., Lü L. and Zhang Y.-C., *Eur. Phys. J. B*, **71** (2009) 623.
[23] Lü L., Jin C.-H. and Zhou T., *Phys. Rev. E*, **80** (2009) 046122.
[24] Lü L. and Zhou T., *EPL*, **89** (2010) 18001.
[25] Granovetter M., *Am. J. Sociol.*, **78** (1973) 1360.
[26] Kemeny J. G. and Snell J. L., *Finite Markov Chains* (Springer-Verlag) 1976.
[27] Norris J., *Markov Chains* (Cambridge University Press) 1997.
[28] Zhou T., Jiang L.-L., Su R.-Q. and Zhang Y.-C., *EPL*, **81** (2008) 58004.
[29] Bonacich P., *Am. J. Sociol.*, **92** (1987) 1170.
[30] Noh J. D. and Rieger H., *Phys. Rev. Lett.*, **92** (2004) 118701.
[31] Wilcoxon F., *Biom. Bull.*, **1** (1945) 80.
[32] Mann H. B. and Whitney D. R., *Ann. Math. Stat.*, **18** (1947) 50.
[33] Hanely J. A. and McNeil B. J., *Radiology*, **143** (1982) 29.
[34] Herlocker J. L., Konstann J. A., Terveen K. and Riedl J. T., *ACM Trans. Inf. Syst.*, **22** (2004) 5.
[35] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.
[36] Newman M. E. J., *Phys. Rev. Lett.*, **89** (2006) 208701.
[37] Newman M. E. J., *Phys. Rev. E.*, **74** (2002) 036104.
[38] von Merging C., Krause R., Snel B., Cornell M., Oliver S. G., Fields S. and Bork P., *Nature*, **417** (2002) 399.
[39] Klein D. J. and Randic M., *J. Math. Chem.*, **12** (1993) 81.
[40] Mossel E. and Vigoda E., *Science*, **309** (2005) 2207.