International Journal of Modern Physics C

PREDICTION OF MISSING LINKS AND RECONSTRUCTION OF COMPLEX NETWORKS*

CHENG-JUN ZHANG[†]
School of Computer & Software, Nanjing University of Information Science & Technology,
No.219, Ningliu Road
Nanjing, Jiangsu 210044/Zone, P. R. China[‡]
zhangcj5@gmail.com

AN ZENG
School of Systems Science, Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District
Beijing, 100875/Zone, P. R. China
zengan1127@gmail.com

Received Day Month Year
Revised Day Month Year

Predicting missing links in complex networks is of great significance from both theoretical and practical point of view, which not only helps us understand the evolution of real system but also relates to many applications in social, biological and online systems. In this paper, we study the features of different simple link prediction methods, revealing that they may lead to the distortion of networks' structural and dynamical properties. Moreover, we find that high prediction accuracy is not definitely corresponding to a high performance in recovering the network properties when using link prediction methods to reconstruct networks. Our work highlight the importance of considering the feedback effect of the link prediction methods on network properties when designing the algorithms.

Keywords: link prediction; complex networks; missing links; network properties

PACS Nos.: 11.25.Hf, 123.1K

1. Introduction

Many social, biological and information systems are naturally described by networks where nodes represent individuals, proteins, genes, computers, web pages, and so on, and links denote the relations or interactions between nodes. As such, network

*For the title, try not to use more than 3 lines. Typeset the title in 10 pt roman, uppercase and boldface.
[†]Typeset names in 8 pt roman, uppercase. Use the footnote to indicate the present or permanent address of the author.
[‡]State completely without abbreviations, the affiliation and mailing address, including country. Typeset in 8 pt italic.

analysis has been applied to many fields including biology, ecology, technology, and sociology [1]. However, it is not for sure that the network data obtained is reliable and complete. For example, biological networks that are inferred from experiments or social networks that result from spontaneous human activity may lose some crucial information, resulting in missing links in networks [2,?].

The problem of identifying missing interactions, known as *link prediction*, consists of estimating the likelihood of the existence of a link between two nodes according to the observed links and node's attributes [4]. Link prediction has already attracted much attention from disparate research communities due to its broad applicability. For instance, in many biological networks (such as food webs, protein-protein interactions, and metabolic networks), the discovery of interactions is often difficult and expensive, hence, accurate predictions can reduce the experimental costs and speed the pace of uncovering the truth [5,6]. Applications in social networks include the prediction of the actors co-starring in acts [7] and of the collaborations in coauthorship networks [8], the detection of the underground relationships between terrorists [5], and many others. In addition, the process of recommending items to users can be considered as a link prediction problem in a user-item bipartite graph [9] so that similarity-based link prediction techniques have been applied to personalized recommendations [10]. Moreover, the link prediction approach can be used to solve the classification problem in partially labeled networks, such as predicting protein functions [11], detecting anomalous email [12], distinguishing the research areas of scientific publications [13], and finding out the fraud and legitimate users in cell phone networks [14]. For a review of the field, see Ref. [15].

However, the previous work in link prediction overwhelmingly focus on improving the prediction accuracy [15]. Actually, it is also necessary to study the consequence if the network is reconstructed based on the link prediction methods (i.e., if the identified "missing links" are added to the networks). If some unexpected links are incorrectly identified as missing links and are added to the network, the system's structure and function may be altered significantly or may even be compromised. Similar problems have been studied when removing spurious links from networks [16], but never been systematically investigated in link predictions. In this paper, we consider more than 10 different kinds of link prediction algorithms. We find that they may lead to the distortion of networks' structural and dynamical properties. Moreover, we find that high prediction accuracy is not definitely corresponding to a high performance in recovering the network properties when using link prediction methods to reconstruct networks. Our work suggests that a well-performed link prediction algorithm should not only enjoy high accuracy but also be effective in preserve the network functionalities when the network is reconstructed based on it.

2. Methods

In this section, we describe our procedure to study the features and to evaluate the performance of a link prediction algorithm. We make use of five empirical undirected

Table 1. Features of empirical networks: number of nodes (N) and edges (E), average degree ($\langle k \rangle$), average shortest path length ($\langle d \rangle$), clustering coefficient (C), degree assortativity (r), degree heterogeneity ($H = \langle k^2 \rangle / \langle k \rangle^2$) and traffic congestability (B_{max})

	N	E	$\langle k \rangle$	$\langle d \rangle$	C	r	H	B_{max}
CE	297	2148	14.46	2.46	0.308	-0.163	1.801	$2.65 \cdot 10^4$
Email	1133	5451	9.62	3.61	0.220	0.078	1.942	$5.06 \cdot 10^4$
SC	379	914	4.82	4.93	0.798	-0.082	1.663	$5.66 \cdot 10^4$
PB	1222	16717	27.36	2.51	0.360	-0.221	2.970	$1.46 \cdot 10^5$
USAir	332	2126	12.81	2.46	0.749	-0.208	3.464	$2.28 \cdot 10^4$

networks: the *Caenorhabditis elegans* (CE) neural network [22], an email (Email) network [23], a scientists' coauthorship (SC) network [24], the U.S. political blogs' (PB) network [25], and the U.S. air transportation (USAir) network [27].We only consider the GC of these real networks. Some properties of these systems are reported in Table I. All of these networks are widely used in the literature as model systems, hence, we assume that they are true networks, which we denote as A^t . We then randomly remove a fraction f of links from these true networks to obtain observed networks, which we denote as A^o, and evaluate the ability of the link prediction algorithm to recover the features of the true networks.

To quantify the accuracy of the algorithm in identifying the missing links we use the standard metric of the area under the receiver operating characteristic curve (AUC) [28]. Since the algorithm returns an ordered list of links (or equivalently gives each nonexisting link a score to quantify its potential to be a true link), the AUC represents the probability that an nonexisting but true link has higher score than other nonexisting links. To obtain the value of the AUC, we randomly pick a true link and a false link among the nonexisting links in the observed network A^o and compare their scores. If, among n independent comparisons, the true link has higher score than the false link n' times and equal score n'' times, the AUC value is:

$$\text{AUC} = \frac{n' + n''/2}{n} \tag{1}$$

Note that if links were ranked at random, the AUC value would be equal to 0.5.

Actually, high accuracy is not sufficient for a link prediction algorithm: when the link prediction algorithms are used to reconstruct the network, the structural and dynamical properties of the network may change dramatically. A well-performed link prediction algorithm should not only enjoy high accuracy but also be effective in preserve the network functionalities when the network is reconstructed based on it. To study the robustness of the algorithm in this respect, we added the fraction f' of the top-ranked links to the observed network to obtain the "reconstructed" network, which we denote as A^r. We then compare the structure and functionality of true and reconstructed networks. If the accuracy of the link prediction algorithm is high enough, the topology properties of the reconstructed network are supposed to be almost the same with that of the true network. In the topology properties, we will consider degree heterogeneity H, clustering coefficient C [29] and average shortest path length d. In network functionalities, we will consider traffic congestability T (i.e. the maximum betweenness centrality in the network) [30], spreading ability E (i.e., the maximum eigenvalue of the adjacency matrix of the network) [31] and synchronizability R (the ratio of the maximum and second smallest eigenvalue of the Laplacian matrix of the network) [32].

3. Link prediction algorithms

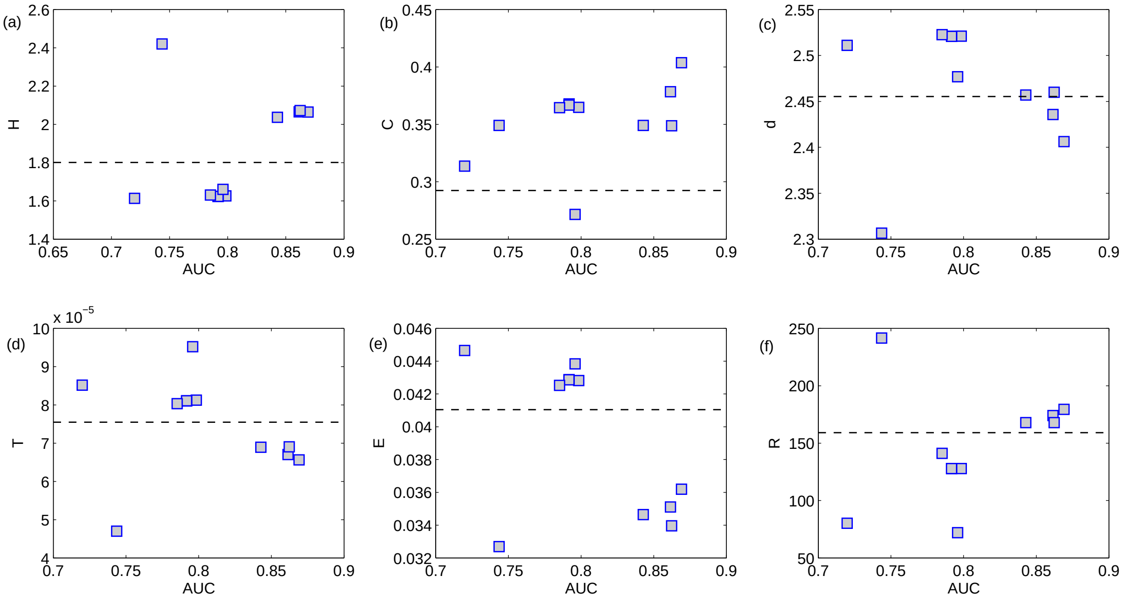

Fig. 1. AUC and the network properties of the reconstructed networks through different link prediction methods in CE network. Note that each point in this figure denotes a link prediction algorithm. The horizontal dash line is the network property of the true network.

In this section, we describe some representative link prediction methods. These algorithms assign to each nonexisting link in A^o a score (denoted as S_{ij} for the nonexisting link between nodes i and j) which quantifies the likelihood of its true existence and allows for link ranking.

(i) *Common Neighbours*-For a node x, let $\tau(x)$ 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 many common neighbours. The simplest measure of this neighbourhood overlap is the directed count, namely

$$s_{xy} = |\tau(x) \cap \tau(y)|. \tag{2}$$

(ii) *Salton Index*-The Salton index [33] is defined as

$$s_{xy} = \frac{|\tau(x) \cap \tau(y)|}{\sqrt{k(x) \times k(y)}} \tag{3}$$

where $k(x) = |\tau(x)|$ denotes the degree of x. The Salton index is also called the cosine similarity in the literature.

(iii) *Jaccard Index*-This index was proposed by Jaccard [34] over a hundred years ago, and is defined as

$$s_{xy} = \frac{|\tau(x) \cap \tau(y)|}{|\tau(x) \cup \tau(y)|} \tag{4}$$

(iv) *Sorensen Index*-This index is used mainly for ecological community data [35], and is defined as

$$s_{xy} = \frac{2 \times |\tau(x) \cap \tau(y)|}{k(x) + k(y)} \tag{5}$$

(v) *Hub Promoted Index*-This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks [36], and is defined as

$$s_{xy} = \frac{|\tau(x) \cap \tau(y)|}{min k(x), k(y)} \tag{6}$$

Under this measure, the links adjacent to hubs (here, the term ”hub“ represents a node with very large degree) are likely to be assigned high scores since the denominator is determined by the lower degree only.

(vi) *Hub Depressed Index*-Analogously to the above index, we consider a measure with the opposite effect on hubs for comparison, defined as

$$s_{xy} = \frac{|\tau(x) \cap \tau(y)|}{max k(x), k(y)} \tag{7}$$

(vii) *Leicht-Holme-Newman Index*-This index assigns high similarity to node pairs that have many common neighbours compared not to the possible maximum, but to the expected number of such neighbours [37]. It is defined as

$$s_{xy} = \frac{|\tau(x) \cap \tau(y)|}{k(x) \times k(y)} \tag{8}$$

where the denominator, $k(x) \times k(y)$, is proportional to the expected number of common neighbours of nodes x and y in the configuration model [38].

(viii) *Preferential Attachment*-The mechanism of preferential attachment can be used to generate evolving scale-free networks (i.e., networks with power-law degree distributions), where the probability that a new link is connected to the node x is proportional to k(x) [39]. Motivated by this mechanism, a corresponding similarity index can be defined as

$$s_{xy} = k(x) \times k(y), \tag{9}$$

which has already been suggested as a proximity measure [40], as well as having been used to quantify the functional significance of links subject to various network-based dynamics, such as percolation [41], synchronization [42] and transportation [43]. Note that this index requires less information than all the others, namely it does not require information on the neighbourhood of each node. As a consequence, it also has the least computational complexity.

(ix) *Adamic-Adar Index*-This index refines the simple counting of common neighbours by assigning the less-connected neighbours more weight [44], and is defined as:

$$s_{xy} = \sum_{z\in\tau x\cap\tau y} \frac{1}{logk(z)}. \tag{10}$$

(iix) *Resource Allocation Index*-The similarity between x and y can be defined as the amount of resource y received from x, which is:

$$s_{xy} = \sum_{z\in\tau x\cap\tau y} \frac{1}{k(z)}. \tag{11}$$

(iix) *Local Path Index*-This index is proposed in [45]. It is used to solve the problem of data sparsity. The similarity matrix can be expressed as:

$$S = A^2 + \beta A^3, \tag{12}$$

where β is a tunable parameter.

4. Method Comparison and Conclusion

In this section, we compare the features of the link prediction approaches mentioned above. As an example, we select the CE network and report AUC and the network properties of the reconstructed networks through different link prediction methods. The results are shown in Fig. 1. Clearly, even when the AUC is high, the network properties of the reconstructed networks can be very far from that of the true network. This phenomenon is more obvious in clustering coefficient where a higher AUC results in a farther clustering coefficient to the true network. This is because most of the link predictions are based on the triangle structure.

In order to compare the performance of different link prediction methods in reproducing the network properties of the true network, we define the *difference rate* as $D' = \frac{|D^r - D^t|}{D^t}$ where D^t denotes a certain network property measure of the true network and D^r is the same network property measure of the reconstructed network. For each network property measure, we can calculate D' (the smaller, the better). In order to estimate the overall ability of these link prediction methods in preserving network properties, we calculate $\langle D' \rangle$ by averaging D' over all the six network property measures (i.e., degree heterogeneity, clustering coefficient, average shortest path length, congestability, spreading ability and synchronizability). In table II, we report AUC and $\langle D' \rangle$ of different link prediction methods.

Table 2. The ranking of different link prediction methods in AUC (ranking1) and preserving network properties (ranking2).

	CN	SI	JI	SSI	HPI	HDI	LHN	PA	AA	RA	LP	WLP
Ranking1												
CE	0.843	0.798	0.792	0.792	0.796	0.785	0.72	0.744	0.862	**0.869**	0.862	0.848
Email	0.856	0.856	0.854	0.855	0.853	0.856	0.848	0.806	0.859	0.858	0.922	**0.924**
SC	0.958	0.958	0.957	0.958	0.961	0.956	0.955	0.665	0.963	0.963	0.978	**0.982**
PB	0.925	0.881	0.878	0.881	0.859	0.878	0.764	0.915	0.929	0.929	**0.94**	0.931
USAir	0.963	0.932	0.921	0.92	0.888	0.915	0.779	0.918	0.970	**0.974**	0.96	0.964
Ranking2												
CE	0.104	0.115	0.117	0.116	0.172	0.097	0.152	0.283	0.133	0.154	0.11	**0.071**
Email	0.122	0.155	0.147	0.13	0.097	0.147	0.139	0.516	0.148	0.143	0.147	**0.07**
SC	0.071	0.057	0.058	0.058	0.094	0.056	0.069	0.318	0.033	**0.031**	0.085	0.063
PB	0.144	0.105	0.109	0.109	0.134	0.116	**0.103**	0.301	0.178	0.238	0.146	0.126
USAir	0.055	0.079	0.08	0.08	0.083	0.079	0.077	0.07	0.051	**0.044**	0.056	0.048

As shown in table 2, the performance of different methods in AUC is relatively stable. Generally speaking, the LP and RA methods enjoy the highest AUC. However, network properties will be changed significantly after adding links according to these two methods (see $\langle D' \rangle$).

In order to achieve a high performance in both AUC and $\langle D' \rangle$, we propose a method called weighted local path (WLP). We denote the local path matrix as $P = A^2 + \beta A^3$ and a normalization matrix as $W = \vec{k} * \vec{k}'$ where $\vec{k}$ is the degree vector of nodes. Mathematically, the WLP method can be expressed as

$$S_{xy} = \frac{P_{xy}}{W^{\theta}_{xy}}, \tag{13}$$

where θ is a tunable parameter. When $\theta = 0$, the WLP method degenerates to the original LP method. When $\theta > 0$, the likelihood of two nodes to have a missing link is adjusted by the degree products of these two nodes. By considering the path with length 3, LP method to enjoy high AUC is because it can effectively break the tie of node pairs with 0 similarity. In the WLP method, this effect is still present, so we expect its prediction accuracy to be high.

Besides examining the existing algorithms, we design this new link prediction algorithm. Specifically, we calculate the prediction score of two nodes by dividing the Local Path between them by their degree product. The basic idea for this new method is that most link prediction algorithms are more likely to give a high prediction score to links between high degree nodes. However, real systems do not always have such strong biased degree preference. By considering the degree products of the two nodes, we can adjust the prediction score between high degree nodes and provide a higher chance for links between low degree nodes to exist. Therefore, the degree adjustment might improve the ability of the method to preserve the network properties when it is used to reconstruct the network.

To test the effectiveness of the WLP method, we again calculate AUC and $\langle D' \rangle$

in Fig. 2. Generally speaking, AUC decreases with θ yet there is a plateau for AUC when θ is small. Interestingly, most of D's decrease with θ when $\theta < 0.3$. After that, these D's start to increase quickly with θ. The results indicates that one can use a small θ to improve the ability of the method to preserve the network properties and the AUC won't be significantly lowered. Specifically, in SC network, AUC can be even improved after introducing the parameter θ.

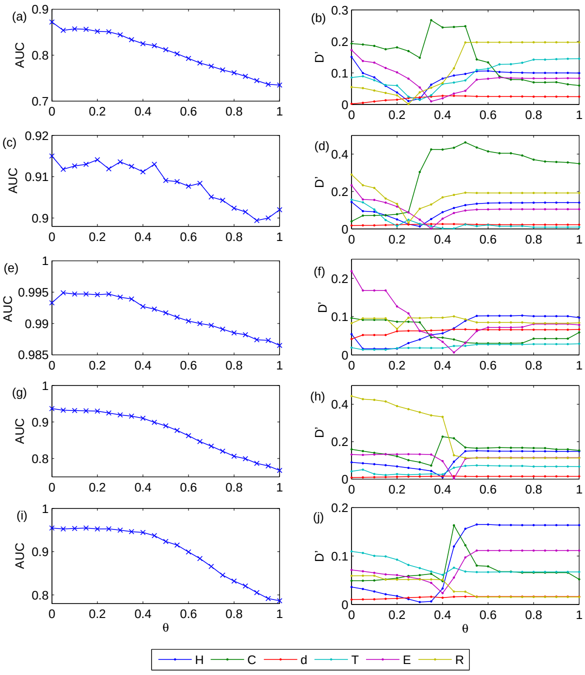

Fig. 2. the dependence of AUC and D' on parameter θ. (a), (b) are the results of CE network. (c), (d) are the results of Email network. (e), (f) are the results of SC network. (g), (h) are the results of PB network. (i) (j) are the results of USAir network.

We set $\theta = 0.2$ in the WLP method and report its AUC and $\langle D' \rangle$ in table II. As one can see that WLP method enjoys almost the same value in AUC as the LP method. Meanwhile, its ability to preserve the network properties are remarkably improved. As we can see, $\langle D' \rangle$ of WLP is not only much smaller than that of LP method, but also close to the optimal value in each network ($\langle D' \rangle$ of WLP is in fact the best in CE and Email networks).

5. Conclusion

In this paper, we investigate the performance of preserving the network properties when the link prediction methods are applied to reconstruct the network (i.e. adding the most likely existing links back to the probe network). We find that many accurate link prediction method may significantly change the network properties. To solve the problem, we propose a weighted local path method. The results show that this method can enjoy quite high accuracy in identifying the missing links. At the same time, its ability to preserve the network properties are remarkably improved compared to some well-known methods.

ACKNOWLEDGEMENTS

The authors would like to acknowledge ???????????????????. The research was supported by The Startup Foundation for Introducing Talent of NUIST. Project No.:2241101301061.

References

1. L. A. N. Amaral and J. M. Ottino, *Eur. Phys. J. B* **38**, 8723 (1987).
2. C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Field and P. Bork, *Nature* **417**, 399 (2002).
3. C. T. Butts, *Soc. Networks* **25**, 103 (2003).
4. L. Getoor and C. P. Diehl, *ACM SIGKDD Explor. Newsl.* **7**, 3 (2005).
5. A. Clauset, C. Moore and M. E. J. Newman, *Nature* **453**, 98 (2008).
6. S. Redner, *Nature* **453**, 47 (2008).
7. J. O'Madadhain, J. Hutchins and P. Smyth, in *Proceedings of SIGKDD* (2005).
8. D. Liben-Nowell and J. Kleinberg, *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019 (2007).
9. J. Kunegis, E. W. D. Luca and S. Albayrak, in *Proceedings of CoRR* (2010).
10. Q.-M. Zhang, M.-S. Shang, W. Zeng, Y. Chen and L. Lü, *Physics Procedia* **3**, 1887 (2010).
11. P. Holme and M. Huss, *J. R. Soc. Interface* **2**, 327 (2005).
12. Z. Huang and D. D. Zeng, in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics* (2006).
13. B. Gallagher, H. Tong, T. Eliassi-Rad and C. Faloutsos, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008).
14. K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A.A. Nanavati and A. Joshi, in *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology* (2008).
15. L. Lü and T. Zhou, *Physica A* **390**, 1150 (2011).
16. A. Zeng and G. Cimini, *Phys. Rev. E* **85**, 036101 (2012).
17. Y. Wang, J. Chu, in *Proceedings of the 20th ACM conference on Hypertext and hypermedia* (2009).
18. D.-H. Kim, J. D. Noh and H. Jeong, *Phys. Rev. E* **70**, 046126 (2004).
19. S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley and S. Havlin, *Nature* **464**, 1025 (2010).
20. R. Guimerà, S. Mossa, A. Turtschi and L. A. N. Amaral, *Proc. Natl. Acad. Sci.* **102**, 7794 (2005).

21. A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno and C. Zhou, *Phys Rep* **469**, 93 (2008).
22. D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
23. R. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, *Phys. Rev. E* **68**, 065103 (2003).
24. M. E. J. Newman, *Phys. Rev. E* **74**, 036104 (2006).
25. R. Ackland, Presentation to BlogTalk Downunder, Sydney, (2005); available at http://incsub.org/blogtalk/images/robertackland.pdf.
26. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, *Nature* (London) **417**, 399 (2002).
27. V. Batageli and A. Mrvar, Pajek Datasets, available at http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm.
28. J. A. Hanely and B. J. McNeil, *Radiology* **143**, 29 (1982).
29. L. F. Costa and F. A. Rodrigues and G. Traviesor and P. R. U. Boas, *Adv. Phys.* 56, 167 (2007)
30. Guimera, R. and Guilera, Diaz A. and Redondo, Vega F. and Cabrales, A. and Arenas, A., *Physical Review Letters.* 89, 248701 (2002)
31. A. N. Bishop and I. Shames, *Europhysics. Lett.* **95**, 18005 (2011).
32. A. Zeng, Y. Hu and Z. Di, *Europhysics. Lett.* **87**, 48002 (2009).
33. G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval* (MuGraw-Hill, Auckland, 1983)
34. P. Jaccard, *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37, 547 (1901).
35. T. Sorensen, *Biol. Skr.* 5, 1 (1948).
36. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabasi, *Science* 297, 1553 (2002)
37. E.A. Leicht, P. Holme, M.E.J. Newman, *Phys. Rev. E* 73, 026120 (2006)
38. M. Molloy, B. Reed, *Random Structure Algorithms* 6, 161 (1995)
39. A.-L. Barabasi, R. Albert, *Science* 286, 509 (1999)
40. Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative ?ltering, In Proceedings of the 5th ACM/IEEECS joint conference on Digital libraries (ACM Press, New York, 2005)
41. P. Holme, B.J. Kim, C.N. Yoon, S.K. Han, *Phys. Rev. E* 65, 056109 (2002)
42. C.-Y. Yin, W.-X. Wang, G.-R. Chen, B.-H. Wang, *Phys. Rev. E* 74, 047102 (2006)
43. G.-Q. Zhang, D. Wang, G.-J. Li, *Phys. Rev. E* 76, 017101 (2007)
44. L.A. Adamic, E. Adar, *Social Networks* 25, 211 (2003)
45. L. Lu, C.-H. Jin, T. Zhou, *Phys. Rev. E* 80, 046122 (2009)