Physica A 443 (2016) 537–542

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Link prediction based on local information considering preferential attachment

Shan Zeng

School of Information Engineering, China University of Geosciences, Beijing 100083, People's Republic of China

HIGHLIGHTS

- A new index is presented to estimate the likelihood of the existence of links.
- The index is compared with five well-known indices on six real networks.
- The index is competitive with LP and Katz, and more accurate than CN and PA.

ARTICLE INFO

Article history:
Received 24 February 2015
Received in revised form 25 May 2015
Available online 13 October 2015

Keywords:
Link prediction
Complex networks
Similarity index
Node similarity

ABSTRACT

Link prediction in complex networks has attracted much attention in many fields. In this paper, a common neighbors plus preferential attachment index is presented to estimate the likelihood of the existence of a link between two nodes based on local information of the nearest neighbors. Numerical experiments on six real networks demonstrated the high effectiveness and efficiency of the new index compared with five well-known and widely accepted indices: the common neighbors, resource allocation index, preferential attachment index, local path index and Katz index. The new index provides competitively accurate prediction with local path index and Katz index while has less computational complexity and is more accurate than the other two indices.

1. Introduction

A wide range of systems from nature to society can be described as complex networks with nodes representing cells, individuals and so on, and links representing the interactions between them [1,2]. The study of complex networks has received much attention, for example, the evolution of networks [3,4], the relations between topologies and functions [5,6] and the characteristic of networks [7] have be studied. Due to the large size and the complexity of the interactions, the current knowledge of these networks such as some biological networks is insufficient [8–10]. It is impossible to check all the interactions, since the existence of links in those biological networks must be discovered in the field and/or in the laboratory, which is time-consuming. Accurate prediction algorithms, based on the known interactions, which can find out the links most likely to exist are required to reduce the experimental costs.

Prediction algorithm could be applied to other systems such as social network where data missing problem occurs [11] and could solve the problem well. Yin et al. proposed an attribute-augmented social network model [12], which is called Social-Attribute Network (SAN) and extended with several leading link prediction algorithms. And these algorithms are compared using large-scale Google+ dataset [13]. Another experimental comparison of existing link prediction techniques is performed using real social network data for the testing [14]. Besides, the prediction algorithm could be used to predict the links which would appear in the near future. New links may appear to show new interaction in the network [15], for

E-mail address: zengshan@cugb.edu.cn.

http://dx.doi.org/10.1016/j.physa.2015.10.016
0378-4371/

example, for networks like friendship networks in web society [16,17], new links presenting promising friendships can be recommended to the relevant users, which is valuable for users and could enhance their loyalties to the web sites.

The simplest framework of link prediction algorithms is the similarity-based method and a survey on them is given [18]. Node similarity, which is defined just by the essential attributes of nodes, means that two nodes are similar if they have many common features [19]. Structural similarity, another group of similarity indices, is based only on the structure of the network and can be further classified as local similarity indices such as the common neighbors (CN) [20], quasi-local indices such as Local path index (LP) [21,22], and global similarity indices such as Katz index [23].

In this paper, a so-called common neighbors plus preferential attachment index characterized to the local similarity indices is proposed. Numerical simulations on six real networks demonstrate that this similarity index is both highly effective and efficient. Its prediction accuracy is competitive with global similarity index katz index while only using local information. So, when the network is huge, this index could be used to predict the missing link and link in the future successfully.

2. Problem description and some existing method

Consider an undirected network $G(V, E)$ without multiple links or self-connections allowed, where V denotes the set of nodes and E denotes the set of links. The set of non-existent links is $U - E$, where U is the universal set. Set U contains all $(|V| \times (|V| - 1))/2$ possible links, where $|V|$ is the number of elements in set V. Assume there are some missing links (or future links) in the set $U - E$. The task of link prediction is finding out these links. To solve the problem, each node pair, $i, j \in V$ is assigned a score s_{ij} which is representing the measure of the existence probability of links between i and j. A descending ordered list of all non-observed links is provided according to their scores and the links at the top are most likely to exist.

To test the algorithms' accuracy, the observed links set E, is randomly divided into two parts: the training set, E^T, is provided as known links information, while the probe set, E^P, treated as unknown links information, is used for testing but not for prediction. Obviously, $E^T \cup E^P = E$ and $E^T \cap E^P = \emptyset$. A standard metric is used to quantify the accuracy of the prediction algorithms: area under the receiver operating characteristic curve (AUC) [24]. AUC can be defined as the probability that a randomly chosen missing link (a link in E^T) is given a higher score than a randomly chosen non-existent link (a link in $U-E$). In the algorithmic implementation, a missing link and a non-existent link are randomly chosen and their scores are compared at each time. If there are n' occurrences of the missing link having a higher score and n'' occurrences of the missing link having the same score as the non-existent link among the n independent comparisons, the *AUC* is defined as $\frac{n'+0.5n''}{n}$. The AUC value should be about 0.5, if all the scores are generated from an independent and identical distribution. Thus, the degree to which the value exceeds 0.5 indicates how much better the algorithm is than random choice.

Similarity-based algorithms are widely accepted to solve the problem and a brief introduction of five similarity indices among them are given: Common Neighbors (CN), Resource Allocation Index (RA), Preferential Attachment index (PA), Local Path index (LP) and Katz index. Their definitions and relevant motivations are introduced as follows:

(1) Common neighbors, also called structural equivalence in Ref. [20], means that two nodes i and j are more likely to form a link in the future if they have many common neighbors. For a node i, let $\Gamma(i)$ denote the set of neighbors of i. The simplest measure of the neighborhood index is the directed count

$$s_{ij} = |\Gamma(i) \cap \Gamma(j)|. \tag{1}$$

The similarity matrix S can be written by $S = A^2$, where A is the adjacency matrix, in which $A_{ij} = 1$ if node i and node j are directly connected and $A_{ij} = 0$ otherwise. Some complex measures are proposed in later work, such as Adamic–Adar Index [25]. Due to the insufficient information, though CN consumes little time and performs relatively well among many local indices, its accuracy cannot catch up with the measures based on global or quasi-local information.

(2) Resource allocation index (RA) is motivated by the resource allocation dynamics on complex networks Ref. [21]. Considering a pair of nodes i and j, which are not directly connected, the node i can send some resource to node j, with their common neighbors playing the role of transmitters. Assume each transmitter has a unit of resource, and will equally distribute it to all its neighbors. The similarity between i and j can be defined as the amount of resource j received from i, which is

$$s_{ij} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k_z}. \tag{2}$$

(3) Preferential attachment index can be used in both growing and ungrowing networks. This mechanism is used to generate evolving scale-free networks, where the probability that a new link is connected to the node i is proportional to the degree $k(i)$ of the node [26]. While a new link is generated instead of an old link in a scale-free networks without growth using similar mechanism [27]. The probability of a new link connecting node i to node j is proportional to $k(i) \times k(j)$. The corresponding similarity index can be defined as

$$s_{ij} = k(i) \times k(j), \tag{3}$$

which can be also written as $s_{ij} = |\Gamma(i)| \times |\Gamma(j)|$ and has already been widely used to quantify the functional significance of links subject to various network-based dynamics [28–30].

(4) Local path index, a typical quasi-local index, uses the information involving the next nearest neighbors besides common neighbors [22]. It is defined as

$$S = A^2 + \epsilon A^3, \tag{4}$$

where ϵ is a free parameter. When nodes i and j are not directly connected, $(A^3)_{i,j}$ equals to the number of different paths with length 3 connecting them. Obviously, this measure degenerates to CN when $\epsilon = 0$.

(5) Katz index, a typical global similarity index, based on the ensemble of all paths, is exponentially damped by length to give the short paths more weights and directly sums them up. It is defined as

$$s_{ij} = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{ij}^{\langle l \rangle}|, \tag{5}$$

where $paths_{ij}^{\langle l \rangle}$ is the set of all paths with length l connecting nodes i and j, and β is a free parameter representing the weights of the paths. The similarity matrix S can be written as $S = (I - \beta A)^{-1} - I$, where β must be lower than the reciprocal of the maximum of the eigenvalues of matrix A to ensure the convergence of Eq. (5).

3. The new method

Preferential attachment index requires less information than all the others mentioned in this context, since it does not require information on the neighborhood of each node except the degree of each node. As a result, it has the least computational complexity but relatively poor accuracy. Katz index needs all the paths resulting good accuracy but most computational complexity among the indices introduced in the last section, so it in-practical in large network. Though common neighbors has low computational complexity, due to the insufficient information, its accuracy cannot catch up with the measures based on quasi-local information such as LP index or global information such as katz index. The probability that two node pairs are assigned the same score is high. Taking INT [31] as an example, there are more than 10^7 two node pairs, most (99.59%) of which are assigned zero score by CN. For all the node pairs having non-zero scores, 91.11% are assigned score 1, and 4.48% are assigned score 2 [21]. LP index uses information involving the next nearest neighbors to make the scores more distinguishable and therefore could improve the accuracy with some computational complexity improved. Actually, CN involves the information of the number of node pairs' common neighbors, but ignores the number of their own neighbors which PA exactly considers. That is to say, $|\Gamma(i)|$ and $|\Gamma(j)|$ which are equal to the degrees of nodes i and j could be considered as the extra information besides common neighbors.

CN performs relatively good among many local indices and widely accepted, while LP using more information improves the accuracy based on it. On the other hand, preferential attachment mechanism indicates a new link is connected to the node i is proportional to $k(i)$ and a "rich-gets-richer" phenomenon that can be easily detected in real networks [26]. To consider the advantages of these three indices, a new index considering local information (the common neighbors and their own neighbors of the node pairs) and preferential attachment is proposed, named *Common Neighbors plus Preferential Attachment* (CN+PA). It is defined as

$$s_{ij} = |\Gamma(i) \cap \Gamma(j)| + \epsilon \frac{|\Gamma(i)| \times |\Gamma(j)|}{\frac{\sum_{z \in V} |\Gamma(z)|}{|V|}}. \tag{6}$$

Obviously, $\frac{\sum_{z \in V} |\Gamma(z)|}{|V|}$ equals to the average degree $\langle k \rangle$ of the network which is calculated only once for a network in implementing the algorithm. So the additional item could also be written as $\epsilon \frac{k(i) \times k(j)}{\langle k \rangle}$, whose computational complexity is similar to PA index and much lower than ϵA^3. This measure degenerates to CN when $\epsilon = 0$. This new index only uses information involving the nearest neighbors which CN uses, and makes the scores more distinguishable to improve the accuracy. The information in neighbors of the node pairs is only used to distinguish node pairs with the same number of common neighbors but different node degrees, therefore ϵ should be a very small number close to zero supposed in Ref. [21]. So in the algorithm realization, the guide is followed and ϵ is set as a small number.

4. Data and analysis

To test the accuracy of the algorithm proposed in this paper, its accuracy measured by AUC is compared with other five similarity indices: Common Neighbors (CN), Resource Allocation Index (RA), Preferential Attachment (PA), Local path index (LP), and katz index, using six representative networks drawn from disparate fields for testing. PPI is a protein–protein interaction network [32] and the giant component contains 2375 proteins and 11693 interactions. NS, a weighted coauthorship network of scientists working on network theory and experiment, contains 1589 scientists [33], and the size of the largest connected component contains only 379 nodes and 914 links. The weighted links are treated as unweighted ones. Grid, a network representing the topology of the western states power grid of the United States [34], with nodes representing generators, transformers and substations, and links corresponding to the high-voltage transmission

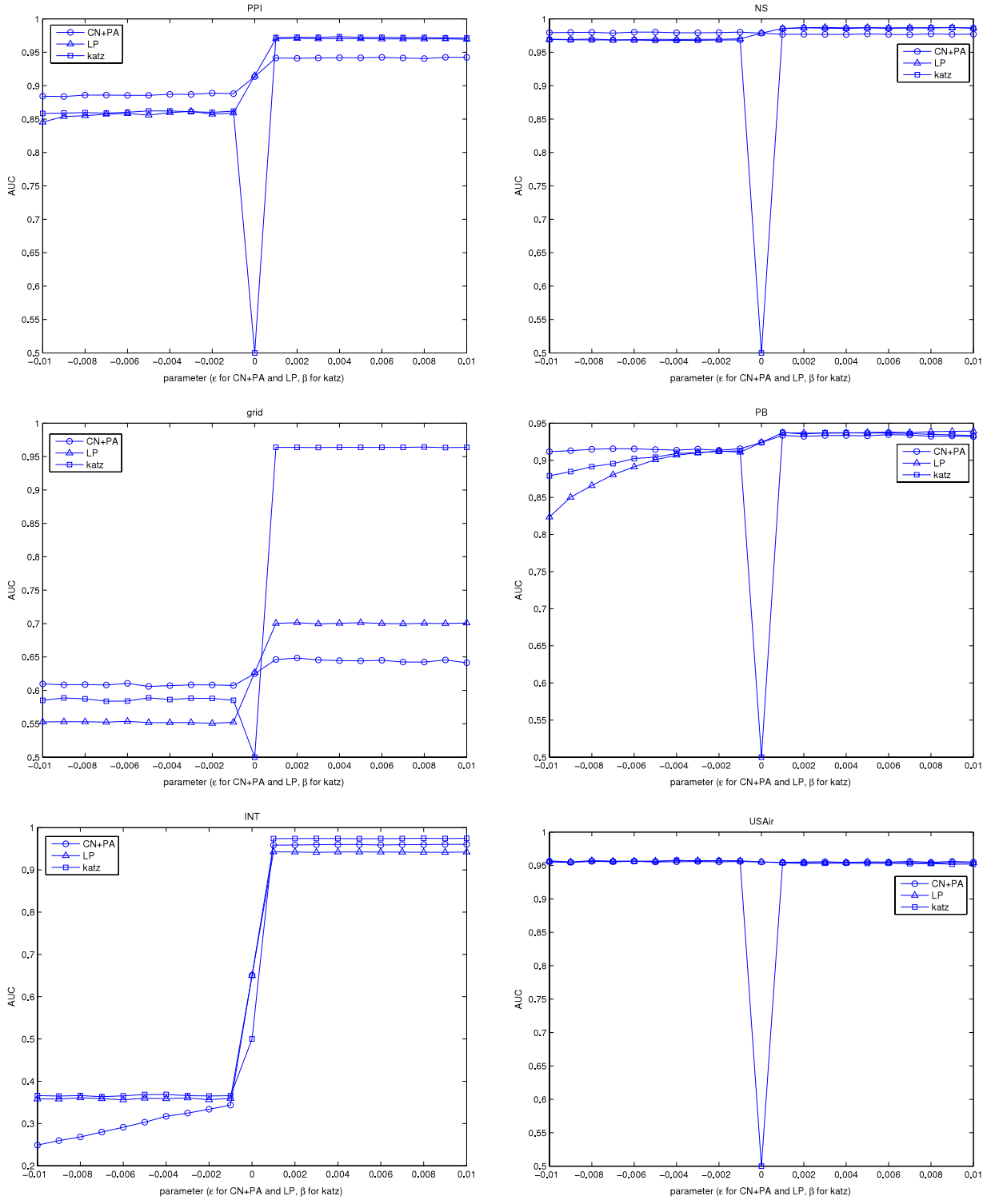

Fig. 1. (Color online) AUC vs the parameter (ϵ for CN+PA and LP, and β for katz) for three similarity indices: CN+PA (circles), LP (triangles), and Katz index (squares). Each data point is obtained by 10 independent realizations.

lines between them, contains 4941 nodes and 6594 links. PB is a directed network of the US political blogs [35] and the giant component contains 1222 nodes and 16714 links. The directed links are treated as undirected ones. INT, collected by the Rocketfuel Project [31], the router-level topology of the Internet, contains 5022 nodes and 6258 edges. USAir, the network

Table 1
Accuracies of the six similarity indices, are measured by AUC value. Each number is obtained by 100 realizations with independently random partitions of testing set and probe set. In the algorithm realization, the training set always contains 90% of the links, and the remaining 10% of links constitute the probe set. For CN+PA, LP and katz indices, the AUC values are corresponding to the optimal parameter. CN+PA*, LP* and katz* denote the CN+PA, LP and katz indices with parameters ϵ and β fixed as 0.01 (-0.01 for USAir in LP and CN+PA both, and -0.01 for NS in CN+PA only), respectively. The entries corresponding to the highest accuracies among these measures are emphasized in black. The very small difference between the optimal case and the case with $\epsilon = 0.01$ suggests that in the real application, one can directly set ϵ as a very small number instead of finding out its optimum.

AUC	PPI	NS	Grid	PB	INT	USAir
CN	0.9156	0.9766	0.6244	0.924	0.6518	0.9547
RA	0.9171	0.9829	0.6239	0.9269	0.6525	**0.9721**
PA	0.8655	0.6549	0.5785	0.9096	0.9547	0.9124
CN+PA	0.9432	0.9802	0.6464	0.9345	0.9603	0.9565
CN+PA*	0.9432	0.9774	0.6388	0.9328	0.9595	0.9558
LP	0.9708	**0.9871**	0.7014	**0.9393**	0.9455	0.9574
LP*	0.9701	0.985	0.6974	0.939	0.945	0.9551
katz	**0.9733**	0.9870	**0.9642**	0.9375	**0.9776**	0.9579
katz*	**0.9733**	0.9854	0.9631	0.9363	**0.9776**	0.9527

of US air transportation system, contains 332 airports and 2126 airlines [36]. Only the giant components of these network are considered and the training set remains a connected network when the links are moved to the probe set.

Link prediction algorithms are applied on these six real networks, and accuracies of the six similarity indices measured by AUC values are shown in Table 1, with those entries corresponding to the highest values emphasized by black. Clearly, the CN+PA index always performs better than the CN and PA index, especially, for INT compared with CN, the AUC is sharply improved from 0.6518 to 0.9547 and for NS compared with PA, the AUC is sharply improved from 0.6549 to 0.9774. The CN+PA index gives competitively accurate predictions as the LP index and katz index (except grid). The average topological distance of grid is 15.87 which is much larger than the other five example networks, and when a link is removed, it is usually hard to find with the local information of the two endpoints. Therefore, the CN, RA, PA, CN+PA and LP indices, considering the local or quasi-local information, fail to refind the correlation between two directly connected nodes if the link is removed. More detailed explanation could be found in Ref. [22].

The AUC value vs the parameter (ϵ for CN+PA and LP, and β for katz) for CN+PA, LP and Katz indices is given in Fig. 1. Clearly, the prediction accuracy for CN+PA is not sensitive to the parameter ϵ when ϵ is small. In addition, the optimal values of for USAir and NS in CN+PA are negative. In NS, the large-degree nodes share many common neighbors and the links among large-degree nodes are assigned high scores. Therefore, the additional item $\epsilon \frac{k(i)\times k(j)}{\langle k\rangle}$ changes little of their relative positions. Considering two small-degree nodes S_m and S_n, which are scientists in the same group connect to each other and a scientist S_x in the same group is the common neighbor of them. That is to say, nodes S_m, S_n and S_x form a triangle. Of course, S_x may connected to S_y who has collaboration with more authors which means the node S_y has larger degree than S_m and S_n. If the link $S_m - S_n$ is removed to the probe set, the scores of links $S_m - S_n$, $S_m - S_y$, $S_n - S_y$ are the same by CN index since they have the same number of common neighbors. But the degree of S_y is larger than those of S_m and S_n, the existent but removed link $S_m - S_n$ have lower score than the non-existent links $S_m - S_y$ and $S_n - S_y$ due to the additional item $\epsilon \frac{k(i)\times k(j)}{\langle k\rangle}$. Since the clustering coefficient of NS network is large, the pattern structure of triangle is dense in NS. When these small-degree-node links are removed, the removed links have lower scores than the non-existent links due to the additional item. Therefore, the large clustering coefficient of NS makes the CN+PA index with positive ϵ worse than the simple CN corresponding to $\epsilon = 0$, which is also the reason why negative ϵ performs even better. And the reason why negative ϵ performs even better in USAir is similar.

5. Conclusion

In this paper, a common neighbors plus preferential attachment index is presented to estimate the likelihood of the existence of a link between two nodes. From the numerical results, it is clearly that, this new index provides more accurate prediction than the CN and PA, and competitively accurate prediction with LP and katz index. From the algorithm definitions mentioned in Sections 2 and 3, LIPA needs less information than LP and katz index, so the computational complexity should be less than that for them. Therefore it is more competitive with large network.

References

[1] C. Koch, G. Laurent, complexity and the nervous system, Science 284 (1999) 96–98.
[2] S. Wasserman, K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, 1994.
[3] R. Albert, A.-L. Barabasi, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) 47.
[4] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, Adv. Phys. 51 (2002) 1079.
[5] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167.
[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, Complex networks: structure and dynamics, Phys. Rep. 424 (2006) 175.
[7] L.d.F. Costa, F.A. Rodrigues, G. Travieso, P.R.U. Boas, Characterization of complex networks: a survey of measurements, Adv. Phys. 56 (2007) 167.
[8] N.D. Martinez, B.A. Hawkins, H.A. Dawah, B.P. Feifarek, Effects of sampling effort on characterization of food-web structure, Ecology 80 (1999) 1044.
[9] E. Sprinzak, S. Sattath, H. Margalit, How reliable are experimental protein-protein interaction data, J. Mol. Biol. 327 (2003) 919.

[10] H. Yu, et al., High-quality binary protein interaction map of the yeast interactome network, Science 322 (2008) 104.
[11] J.W. Neal, Kracking the missing data problem: applying Krackhardts cognitive social structures to school-based social networks, Soc. Educ. 81 (2008) 140.
[12] Z. Yin, M. Gupta, T. Weninger, J. Han, LINKREC: A unified framework for link recommendation with user attributes and graph structure, in: Proceedings of the 19th International Conference on World Wide Web, ACM, Raleigh, North Carolina, USA, 2010, pp. 1211–1212.
[13] N.Z. Gong, A. Talwalkar, L. Mackey, et al., Joint link prediction and attribute inference using a social-attribute network, ACM Trans. Intell. Syst. Technol. 5 (2014) 27.
[14] D. Sharma, U. Sharma, S. Khatri, An experimental comparison of the link prediction techniques in social networks, Int. J. Model. Optim. 4 (2014) 21.
[15] P. Gupta, S. Sharma, A survey on link prediction problem in social network, Int. J. Sci. Res. Technol. 1 (2015) 43.
[16] A. Grabowski, N. Kruszewska, R.A. Kosinski, Dynamic phenomena and human activity in an artificial society, Phys. Rev. E 78 (2008) 066110.
[17] H.-B. Hu, X.-F. Wang, Disassortative mixing in online social networks, Europhys. Lett. 86 (2009) 18003.
[18] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Phys. A 390 (2011) 1150.
[19] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufman Publishers, San Francisco, CA, 1998, pp. 296–304.
[20] F. Lorrain, H.C. White, Structural equivalence of individual in social networks, J. Math. Sociol. 1 (1971) 49.
[21] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623.
[22] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (2009) 046122.
[23] L. Katz, A new status index derived from sociometric analysis, Psychmetrika 18 (1953) 39.
[24] J.A. Hanely, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29.
[25] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (2003) 211.
[26] A.-L. Barabasi, R. Albert, Emergence of scalingin random networks, Science 286 (1999) 509.
[27] Y.-B. Xie, T. Zhou, B.-H. Wang, Scale-free networks withou tgrowth, Phys. A 387 (2008) 1683.
[28] P. Holme, B.J. Kim, C.N. Yoon, S.K. Han, Attack vulnerability of complex networks, Phys. Rev. E 65 (2002) 056109.
[29] C.-Y. Yin, W.-X. Wang, G.-R. Chen, B.-H. Wang, Decoupling process for better synchronizability on scale-free networks, Phys. Rev. E 74 (2006) 047102.
[30] G.-Q. Zhang, D. Wang, G.-J. Li, Enhancing the transmission efficiency by edge deletion in scale-free networks, Phys. Rev. E 76 (2007) 017101.
[31] N. Spring, R. Mahajan, D. Wetherall, T. Anderson, Measuring ISP topologies with Rocketfuel, IEEE/ACM Trans. Netw. 12 (2004) 2.
[32] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (2002) 399.
[33] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[34] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (1998) 440.
[35] R. Ackland, Mapping the US political blogosphere: are conservative bloggers more prominent, in: Presentation to BlogTalk Downunder, Sydney, 2005; available at http://incsub.org/blogtalk/images/robertackland.pdf.
[36] V. Batageli, A. Mrvar, Pajek Datasets, available at http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm.