Eur. Phys. J. B **71**, 623–630 (2009)
DOI: 10.1140/epjb/e2009-00335-8

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Predicting missing links via local information

Tao Zhou[1,2,3,a], Linyuan Lü[1,2], and Yi-Cheng Zhang[1,2,3]

[1] Research Center for Complex System Science, University of Shanghai for Science and Technology, 200093 Shanghai, P.R. China
[2] Department of Physics, University of Fribourg, Chemin du Musée 3, 1700 Fribourg, Switzerland
[3] Department of Modern Physics, University of Science and Technology of China, 230026 Hefei Anhui, P.R. China

Received 5 January 2009 / Received in final form 1st June 2009
Published online 10 October 2009

Abstract. Missing link prediction in networks is of both theoretical interest and practical significance in modern science. In this paper, we empirically investigate a simple framework of link prediction on the basis of node similarity. We compare nine well-known local similarity measures on six real networks. The results indicate that the simplest measure, namely Common Neighbours, has the best overall performance, and the Adamic-Adar index performs second best. A new similarity measure, motivated by the resource allocation process taking place on networks, is proposed and shown to have higher prediction accuracy than common neighbours. It is found that many links are assigned the same scores if only the information of the nearest neighbours is used. We therefore design another new measure exploiting information on the next nearest neighbours, which can remarkably enhance the prediction accuracy.

PACS. 89.75.-k Complex systems – 05.65.+b Self-organized systems

1 Introduction

Many social, biological, and information systems can be properly described as networks with nodes representing individuals or organizations and edges representing the interactions among them. The study of complex networks has attracted increasing attention and become a common focus of many branches of science. Many efforts have been made to understand the evolution of networks [1,2], the relations between topologies and functions [3,4], and the network characteristics [5]. Very recently, a fresh question has arisen [6], that is, how to predict missing links of networks? For some networks, especially biological networks such as protein-protein interaction networks, metabolic networks and food webs, the discovery of links (i.e., interactions) is costly in the laboratory or the field, and thus the current knowledge of those networks is substantially incomplete [7,8]. Instead of blindly checking all possible interactions, prediction based on the interactions already known and focusing on those links most likely to exist can sharply reduce the experimental costs if the predictions are accurate enough. For some others like the web-based friendship networks [9,10], very likely but not yet existent links can be suggested to the relevant users as recommendations of promising friendships, which can help users in finding new friends and thus enhance their loyalties to web sites.

The majority of previous works on missing link prediction have used some external information besides the network topology [11]. Graven et al. [12] predicted the semantic relationships of the world wide web with the help of web content. Popescul and Ungar [13] designed a regression model to predict citations made in scientific literature based not only on the citation graph, but also on authorship, journal information and content. Taskar et al. [14] applied the relational Markov network algorithm to predict missing links in a network of web pages and a social network, in which the well-defined attributes of each node are exploited. O'Madadhain et al. [15] constructed local conditional probability models for link prediction, based on both structural features and nodes' attributes. The usage of external information can somewhat enhance the algorithmic accuracy, however the content and attribute information are generally not available and thus the applications of the above algorithms are strongly limited. Goldberg and Roth [16] exploited the neighbourhood cohesiveness property of small-world networks to assess confidence for individual protein-protein interactions. Liben-Nowell and Kleinberg [17] empirically investigated the similarity-based prediction algorithms for large scientific collaboration networks. Clauset et al. [18] designed a prediction algorithm based on the inherent hierarchical organization of social and biological networks.

The above-mentioned works are successful in dealing with specific networks, however thus far a comprehensive picture of the dependence of algorithmic performance on network topology is lacking. The reason is twofold: (i) the works from engineering and biological communities have not yet caught up with the current state of development

[a] e-mail: zhutou@ustc.edu

in characterizing the topologies of complex networks [5], while (ii) the physics community has not paid enough attention to the link prediction problem. Accordingly, dozens of important issues are still insufficiently explored. For example, one may be concerned with how to choose a suitable algorithm given some structural descriptions of a network, such as the small-world phenomenon [19], degree heterogeneity [20], mixing pattern [21], community structure [22], and so on. From the opposite viewpoint, comparison of the performances of some prediction algorithms may reveal some of the structural information of the networks. It is just like the community structure has a significant effect on the network synchronizability [23], while the synchronizing process can be used to reveal the underlying community structure [24]. In addition, the algorithms based only on local information are generally fast but of lower accuracy, while the ones making use of knowledge of global topology are of higher accuracy yet higher computational complexity [17]. Can we find a good trade-off that provides high quality predictions while requiring light computation?

In this paper, we empirically investigate a simple framework of link prediction on the basis of node similarity. Although the framework is simple, it opens a rich space for exploration since the design of similarity measures is challenging and can be related to very complicated physical dynamics and mathematical theory, such as random walks [25] and the counting problem of spanning trees [26]. Here we concentrate on local-information-based similarities. We compare nine well-known local measures on six real networks, and the results indicate that the simplest measure, namely *common neighbours*, has the best overall performance, which is in accordance with the empirical results reported in reference [17]. Motivated by the resource allocation process in transportation networks, we next propose a new similarity measure, which performs noticeably better than common neighbours, while requiring no more information and computational time. Furthermore, it is found that many links get the same scores under local similarity measures, just like the degeneracy of energy levels. We therefore design a new measure using the information of the next nearest neighbours, which can break the "degeneracy of states" and thus remarkably enhance the algorithmic accuracy. Finally, we outline some future interests in this direction.

2 Method

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, $x, y \in V$, every algorithm referred to in this paper assigns a score s_{xy}. This score can be viewed as a measure of similarity between nodes x and y, and hereinafter we do not distinguish *similarity* and *score*. All the nonexistent links are sorted in decreasing order according to their scores, and the links at the top are most likely to exist.

To test the algorithm's accuracy, the observed links, E, are randomly divided into two parts: the training set, E^T, is treated as known information, while the probe set, E^P, is used for testing and no information in this set is allowed to be used for prediction. Clearly, $E = E^T \cup E^P$ and $E^T \cap E^P = \varnothing$. In this paper, the training set always contains 90% of links, and the remaining 10% of links constitute the probe set. We use a standard metric, the area under the receiver operating characteristic (ROC) curve [27], to quantify the accuracy of the prediction algorithms. In the present case, this metric 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 $U \setminus E$, where U denotes the universal set). In the implementation, among n independent comparisons, if there are n' occurrences of the missing link having a higher score and n'' occurrences of the missing link and nonexistent link having the same score, we define the *accuracy* as:

$$\mathtt{AUC} = \frac{n' + 0.5n''}{n}. \tag{1}$$

If all the scores are generated from an independent and identical distribution, the accuracy should be about 0.5. Therefore, the degree to which the accuracy exceeds 0.5 indicates how much better the algorithm performs than pure chance.

3 Data

In this paper, we consider six representative networks drawn from disparate fields: (i) PPI – A protein-protein interaction network containing 2617 proteins and 11855 interactions [28]. Although this network is not well connected (it contains 92 components), most of the nodes belong to the giant component, whose size is 2375. (ii) NS – A network of coauthorships between scientists who are themselves publishing on the topic of networks [29]. The network contains 1589 scientists, 128 of which are isolated. Here we do not consider those isolated nodes. The connectivity of NS is not good, in fact NS consists 268 connected components, and the size of the largest connected component is only 379. (iii) Grid – An electrical power grid of the western US [19], with nodes representing generators, transformers and substations, and edges corresponding to the high voltage transmission lines between them. (iv) PB – A network of the US political blogs [30]. The original links are directed, however here we treat them as undirected. (v) INT – The router-level topology of the Internet, as collected by the *Rocketfuel Project* [31]. (vi) USAir – The network of the US air transportation system, which contains 332 airports and 2126 airlines [32].

Table 1 summarizes the basic topological features of these networks. Brief definitions of the monitored topological measures can be found in the table caption. For more details, please see the review articles [1–5]. We here give a few remarks on the numbers which may be unexpected

Table 1. The basic topological features of six example networks. N and M are the total numbers of nodes and links, respectively. N_c denotes the size of the giant component. For example, the entry 2375/92 in the first line means that the network has 92 components and the giant component consists of 2375 nodes. e is the network efficiency [33], defined as $e = \frac{2}{N(N-1)} \sum_{x,y \in V, x \neq y} d_{xy}^{-1}$, where d_{xy} is the shortest distance between x and y, and $d_{xy} = +\infty$ if x and y are in two different components. C and r are the clustering coefficient [19] and assortative coefficient [21], respectively. Nodes with degree 1 are excluded from the calculation of clustering coefficient. H is the degree heterogeneity, defined as $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$, where $\langle k \rangle$ denotes the average degree.

Nets	N	M	N_c	e	C	r	H
PPI	2617	11855	2375/92	0.180	0.387	0.461	3.73
NS	1461	2742	379/268	0.016	0.878	0.462	1.85
Grid	4941	6594	4941/1	0.063	0.107	0.003	1.45
PB	1224	19090	1222/2	0.397	0.361	–0.079	3.13
INT	5022	6258	5022/1	0.167	0.033	–0.138	5.50
USAir	332	2126	332/1	0.406	0.749	–0.208	3.46

for some readers: (i) it is well known that in the protein-protein interaction networks, links between highly connected proteins are systematically suppressed, while those between highly-connected and weakly-connected pairs are favoured [34]. That is to say, the assortative coefficient should be negative for PPI (for example, as reported in reference [21], the Yeast PPI network has an assortative coefficient –0.156), however, in the present network, the assortative coefficient is very positive – 0.461. This is because the data set used here [28] is determined from functional interactions and not from physical interactions. More detailed discussion can be found in reference [35]. (ii) The extremely large clustering coefficient of NS is due to the specific construction rule of collaboration networks, namely that all the participants in an act are fully connected. Relevant discussion can be found in *Appendix B* of reference [36].

4 Comparison of nine similarity measures based on local information

In this section, we compare the prediction accuracies of nine similarity measures. All these measures are based on the local structural information contained in the testing set. We first give a brief introduction of each measure as follows.

(i) *Common Neighbours* – 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 many common neighbours. The simplest measure of this neighbourhood overlap is the directed count, namely

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

(ii) *Salton Index* – The Salton index [37] is defined as

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

where $k(x) = |\Gamma(x)|$ 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 [38] over a hundred years ago, and is defined as

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

(iv) *Sørensen Index* – This index is used mainly for ecological community data [39], and is defined as

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

(v) *Hub Promoted Index* – This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks [40], and is defined as

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

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

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

(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 [41]. It is defined as

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

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* [42].

(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 $k(x)$ [20]. A similar mechanism can also lead to scale-free networks without growth [43], where at each time step, an old link is removed and a new link is generated. The probability of this new link connecting x and y is proportional to $k(x) \times k(y)$. 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 [44], as well as having been used to quantify the functional significance of links subject to various network-based dynamics, such as percolation [45], synchronization [46] and transportation [47]. 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 [48], and is defined as:

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

We present the algorithmic accuracies for the six example networks in Table 2, with those entries corresponding to the highest accuracies being emphasized in black. To our surprise, the simplest measure, common neighbours, performs the best. This result is in accordance with the one reported in reference [17] for social collaboration networks. After CN, the Adamic-Adar index performs the next best since its accuracies are always close to the best one, while others, such as the Jaccard index, Sørensen index and HDI, perform far worse in the cases of PB and USAir.

Note that the first seven measures, from CN to LHN, only differ in the denominators. If all the nodes have pretty much the same degree, corresponding to a very small H, then the difference among those measures becomes insignificant. In addition, for a given network, if its clustering coefficient is very small, whether two nodes have common neighbours plays the most important role, while the denominator is less important. In a word, significant difference among those seven measures can be found only if the investigated network simultaneously has large clustering coefficient and large degree heterogeneity, such as PPI, PB and USAir. As shown in Table 2, the performances of those seven algorithms on PB and USAir are clearly different, but for PPI, they are more or less the same. A possible reason is that PPI is a very assortative network (i.e., $r = 0.461$), and thus two nodes of a link tend to have similar degrees, which reduces the difference in denominators.

The preferential attachment index has the worst overall performance. However, we are interested in it for it requires the least information. One may intuitively think that PA will give good predictions for assortative networks, while performing badly for disassortative networks. However, no obvious correlation between assortative coefficient and algorithmic accuracy based on PA can be found from our numerical results. The reason is twofold. Firstly, links between pairs of high-degree nodes contribute positively to the assortative coefficient and are assigned high scores by PA, while links between pairs of low-degree nodes also contribute positively to the assortative coefficient but are disfavored by PA. Actually, the assortative coefficient is an integrated measure involving many ingredients, and there is no simple relation between this measure and the performance of PA. Secondly, the assortative coefficient itself is very sensitive to the degree sequence, and a network of higher degree heterogeneity tends to be disassortative [49]. Therefore, this single parameter cannot reflect the detailed linking patterns of networks. Clearly, if the high-degree nodes are very densely connected to each other, and the low-degree nodes are rarely connected to each other, PA will perform relatively well. The former relates to the so-called *rich-club phenomenon* [50], and we have checked that PB and USAir clearly exhibit the rich-club phenomenon with respect to their randomized versions (we followed the method proposed by Colizza et al. [51], who have already demonstrated the presence of rich-club phenomenon in the air transportation network). In addition, in USAir, more than 40% of nodes are very small local airports, with degrees no larger than 3. A local airport usually connects to a nearby central airport and very few hubs, with direct links between two local airports rarely found. This topological feature is also favoured by PA. As shown in Table 2, PA gives relatively good predictions for PB and USAir, in accordance with the above

Table 2. Accuracies of algorithms, measured by the area under the ROC curve. Each number is obtained by averaging over 10 implementations with independently random partitions of testing set and probe set. The abbreviations, CN, Salton, Jaccard, Sørensen, HPI, HDI, LHN, PA, and AA, stand for Common Neighbours, Salton Index, Jaccard Index, Sørensen Index, Hub Promoted Index, Hub Depressed Index, Leicht-Holme-Newman Index, Preferential Attachment and Adamic-Adar Index, respectively. The entries corresponding to the highest accuracies among these nine measures are emphasized in black. RA and LP are abbreviations for the Resource Allocation Index and Local Path Index, proposed in Sections 5 and 6 respectively. The parameter for LP, ϵ, is fixed as 10^{-3}.

Measures	PPI	NS	Grid	PB	INT	USAir
CN	0.889	0.933	0.590	0.925	0.559	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	0.933	0.590	0.882	0.559	0.901
Sørensen	0.888	0.933	0.590	0.881	0.559	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	0.933	0.590	0.877	0.559	0.895
LHN	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	0.590	0.922	0.559	0.925
RA	0.890	0.933	0.590	0.931	0.559	0.955
LP	0.939	0.938	0.639	0.936	0.632	0.900

discussion. Note that all the other eight measures will automatically assign zero score to a pair of nodes located in different components. Therefore, PA performs badly when the network consists many components. This is the very reason why PA gives very bad predictions for NS, although NS clearly exhibits the rich-club phenomenon. We also note that PA performs even worse than pure chance for the Internet at router level and the power grid. In these two networks, the nodes have well-defined positions and the links are physical lines. Actually, geography plays a significant role and links with very long geographical distances are rare (the empirical analysis of the spatial dependence of links in the Internet can be found in reference [52], and the absence of clustering-degree correlation in the router-level Internet and power grid can be considered as an indicator of a strong geographical constraint [53]). PA can not take into account the effect of geographical localization at all. As local centers, the high-degree nodes have longer geographical distances to each other than average. Correspondingly, they also have a lower probability of directly connecting to each other. Actually, these two networks exhibit the anti-rich-club phenomenon, that is, the link density among very-high-degree nodes is even lower than the randomized versions. This anti-rich-club effect leads to the bad performance of PA. In contrast, although USAir has well-defined geographical positions of nodes, its links are not physical. Empirical data has demonstrated that the air transportation networks show an inverse relation between clustering coefficient and degree [54], and the number of airline flights is not sensitive to the geographical distance within a range of about 2000 kilometers [55]. As a final remark, comparing equations (8) and (9), LHN is, to some extent, inverse to PA. Therefore when PA performs badly, LHN will give relatively good predictions, and vice versa.

5 Similarity measure based on resource allocation

Except PA, all the other measures introduced in the last section are neighbourhood-based. Although they are simple and mathematically elegant, they are not tightly related to any physical processes. In this section, motivated by the resource allocation process taking place in networks [56], we propose a new similarity measure, which has overall higher accuracy than all the measures mentioned in Section 4.

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. In the simplest case, we assume that each transmitter has a unit of resource, and will equally distribute it between all its neighbours. The similarity between x and y can be defined as the amount of resource y received from x, which is:

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

Clearly, this measure is symmetric, namely $s_{xy} = s_{yx}$.

The algorithmic accuracies on the six example networks are presented in Table 2, with RA the abbreviation for Resource Allocation. Compared with all the nine measures introduced in Section 4, RA performs the best, especially for the networks (i.e., PB and USAir) with large clustering coefficient, high degree heterogeneity and absence of a strongly assortative linking pattern. It is observed that RA exhibits particularly good performance on USAir. The reason may be that the resource allocation process was originally proposed to explain the nonlinear correlation between transportation capacity and connectivity of each airport [54,57,58].

Note that, although resulting from different motivations [48,56], the Adamic-Adar index and resource allocation index have a very similar form. Indeed, they both suppress the contributions of common neighbours with high degrees. The difference between $\frac{1}{\log k(z)}$ and $\frac{1}{k(z)}$ (see Eqs. (10) and (11)) is insignificant if the degree, $k(z)$, is small, while it is great if $k(z)$ is large. Therefore, when the average degree is very small, the prediction results of AA and RA are very close, while for the networks of high average degree, such as PB and USAir, the results are clearly different and the RA measure performs better, which implies that AA's penalty for high-degree common neighbours is insufficient.

RA can be extended to the asymmetric case. Assuming a unit of resource is located in x, which will be equally send to all x's neighbours, each of which will equally distribute the receives resource one step further to all its neighbours. The amount of resource a node y received can be considered as the importance of y in x's view, denoted as

$$s_{xy} = \frac{1}{k(x)} \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}. \tag{12}$$

In this case, $s_{xy} \neq s_{yx}$. This idea has already found its applications in a personalized recommendation algorithm of bipartite user-object networks [59,60].

6 Improving algorithmic accuracy by breaking the degeneracy of states

The neighbourhood-based measures require only information on the nearest neighbours, and therefore have very low computational complexity. However the information usually seems insufficient and the probability that two node pairs are assigned the same score is high. That is to say, the neighbourhood-based similarity measures are less distinguishable from each other. If we consider the score assigned to a node pair as its energy, then many node pairs crowd into very few energy levels. Taking INT as an example, there are more than 10^7 node pairs, 99.59% of which are assigned zero score by CN. For all the node pairs having scores higher than 0, 91.11% are assigned score 1, and 4.48% are assigned score 2. Using a little bit more information involving the next nearest neighbours may break the "degeneracy of the states" and make the scores more distinguishable. Denote by A the adjacency

matrix, where $A_{xy} = 1$ if x and y are directly connected, and $A_{xy} = 0$ otherwise. Obviously, $(A^2)_{xy}$ is the number of common neighbours of nodes x and y, which is also equal to the number of different paths with length 2 connecting x and y. And if x and y are not directly connected (this is the case we are interested in), $(A^3)_{xy}$ is equal to the number of different paths with length 3 connecting x and y. The information contained in A^3 can be used to break the degeneracy of the states, and thus we define a new measure as

$$S = A^2 + \epsilon A^3, \tag{13}$$

where S denotes the similarity matrix and ϵ is a free parameter. We call it the *Local Path* (LP) index since it makes use of the information on local paths with lengths 2 and 3. Clearly, LP reduces to CN when $\epsilon = 0$. Here, the information in A^3 is only used to break the degeneracy of the states, therefore ϵ should be a very small number close to zero (of course, given a network, one can tune ϵ to find its optimal value corresponding to the highest accuracy, however this optimal value is different for different networks, and a parameter-dependent measure is less practical in dealing with huge-size networks since the tuning process may take much time). In the real implementation, we directly count the number of different paths with length 3, which is much faster than the matrix multiplication, and thus equation (13) is also based on local calculation.

The algorithm's accuracies on the six example networks are presented in Table 2, where this measure is denoted by LP and the parameter is fixed at $\epsilon = 10^{-3}$. It is pleasing to see that the accuracy, except for USAir, can be largely enhanced by LP. In USAir, the large-degree nodes are densely connected and share many common neighbours. Some links among high-degree nodes are removed into the probe set. Even without the contribution of ϵA^3, those links are assigned very high scores, and thus the additional item, ϵA^3, changes their relative positions little. Consider two small local airports, x and y, which are connected to their local central airports, x' and y'. Of course, many hubs are common neighbours of x' and y', and x' and y' may be directly connected. If the link (x, x') is removed, the similarities between x and other nodes are all zero for both CN and LP. If (x, x') exists, by LP, the similarities $s_{xy'}$ (by x-x'-hub-y'), s_{xy} (by x-x'-y'-y), and s_{xh} where h represents a hub node (by x-x'-hub-h and/or x-x'-y'-h) are positive due to the contributions of paths with length 3. There are many links connecting small local airports and local centers, some of which are removed, while the others are kept in the testing set. According to the above discussion, the removed links have a lower score than the nonexistent links due to the additional term ϵA^3. In a word, the very specific structure of USAir (the hierarchical organization consisting of hubs, local centers and small local airports) makes the LP worse than the simple CN. In this specific case, we can break the degeneracy of the states in the opposite direction by setting ϵ equal to -10^{-3}, which leads to an accuracy 0.945, higher than that of CN, 0.937.

7 Conclusion and discussion

In this paper, we have empirically compared some link prediction algorithms based on node similarities. All the similarity measures discussed here, including the two newly proposed ones, can be obtained by local calculations. Numerical results on the nine well-known measures indicate that: (i) the simplest measure, common neighbours, performs best, with the Adamic-Adar index second; (ii) significant difference between these measures, excluding the Adamic-Adar index and the preferential attachment, can be observed only if the monitored network possesses a large clustering coefficient, high degree heterogeneity, and the absence of a strongly assortative linking pattern; (iii) the preferential attachment index performs relatively well if the monitored network displays the rich-club phenomenon.

We have proposed a new measure, RA, motivated by the resource allocation process, which is equivalent to the one-step random walk starting from the common neighbours. This measure has a similar form to the Adamic-Adar index, but performs better, especially for the networks with high average degree. We make the prediction, whose validity requires further evidence from more empirical results, that RA is particularly suitable for link prediction in transportation networks. We strongly recommend this measure for relevant applications and theoretical analyses, not only for its good performance, but also for its simplicity and elegance.

Furthermore, we have found that many links are assigned identical scores based on the local measures using the information on the nearest neighbours only. Exploitation of some additional information on the next nearest neighbours can therefore break the degeneracy of the states and enhance the algorithmic accuracy. In real applications, the algorithms based on global calculations may be less efficient for they require long time and/or huge memory, while the algorithms only exploiting very local information may be less effective due to their low accuracies. A properly designed algorithm can provide a good tradeoff just like the LP index presented in this paper. Indeed, it has been shown recently that the LP index provides competitively accurate predictions compared with the indices making use of global information [61]. A similar idea has also been adopted in the study of network-based traffic dynamics, where the information on the next nearest neighbours can sharply enhance the traffic efficiency compared with the case in which only the information on the nearest neighbours is known [62].

Although the framework adopted here is very simple, it opens a rich space for investigation since in principle, all algorithms can be embedded into this framework differing only in the similarity measures. Besides the ones discussed in this paper, a number of similarity measures are based on global structural information, such as the average commute time of a random walk [25], the number of spanning trees embedding a given node pair [26], the pseudoinverse of the Laplacian matrix [63], and so on. Some other similarity measures are even more complicated, depending on parameters. These include the Katz index [64]

and its variant [41], the transferring similarity [65], the PageRank index [66], and so on. These measures may give better predictions than the local ones, however the calculation of such measures, including determination of the optimal parameters for specific networks, is of high complexity and thus unfeasible for huge-size networks. In any case, we currently lack a systematic comparison and a clear understanding of the performance of these measures, which are set as our future goals.

Empirical analysis of more real networks as well as more known and newly proposed similarity measures is very valuable for building up knowledge and experience, and we can expect a clear picture of this issue to be completed by the putting together of many fragments from respective empirical studies. However, the empirical results may not be clear at all since many unknown and uncontrollable ingredients are always mixed together in real networks. An alternative route is to build artificial network models with controllable topological features, and to compare the prediction algorithms on these models (see Ref. [61] for the comparison of link prediction algorithms on modeled networks with controllable density and noise strength).

This work is partially supported by the Swiss National Science Foundation (Project 205120-113842) and Physics of Risk through project C05.0148. T.Z. acknowledges the National Natural Science Foundation of China (Grant Nos. 10635040, 60744003 and 10905052). L.L. acknowledges the National Natural Science Foundation of China under Grant No. 60973069.

References

1. R. Albert, A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
2. S.N. Dorogovtsev, J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002)
3. M.E.J. Newman, SIAM Rev. **45**, 167 (2003)
4. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, Phys. Rep. **424**, 175 (2006)
5. L.d.F. Costa, F.A. Rodrigues, G. Travieso, P.R.U. Boas, Adv. Phys. **56**, 167 (2007)
6. S. Redner, Nature **453**, 47 (2008)
7. N.D. Martinez, B.A. Hawkins, H.A. Dawah, B.P. Feifarek, Ecology **80**, 1044 (1999)
8. E. Sprinzak, S. Sattath, H. Margalit, J. Mol. Biol. **327**, 919 (2003)
9. A. Grabowski, N. Kruszewska, R.A. Kosiński, Phys. Rev. E **78**, 066110 (2008)
10. H.-B. Hu, X.-F. Wang, Europhys. Lett. **86**, 18003 (2009)
11. L. Getoor, C.P. Diehl, Link Mining: A Survey, in *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2005)
12. M. Graven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, Artificial Intelligence **118**, 69 (2000)
13. A. Popescul, L. Ungar, Statistical relational larning for link prediction, in *Workshop on Learning Statistical Models from Relational Data* (ACM Press, New York, 2003), pp. 81–90
14. B. Taskar, M.-F. Wong, P. Abbeel, D. Koller, Link prediction in relational data, in *Proceeding of Neural Information Processing Systems* (MIT Press, Cambridge, 2003), pp. 659–666
15. J. O'Madadhain, J. Hutchins, P. Smyth, Prediction and ranking algorithms for even-based network data, In *Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2005)
16. D.S. Goldberg, F.P. Roth, Proc. Natl. Acad. Sci. U.S.A. **100**, 4372 (2003)
17. D. Liben-Nowell, J. Kleinberg, J. Am. Soc. Inform. Sci. Technol. **58**, 1019 (2007)
18. A. Clauset, C. Moore, M.E.J. Newman, Nature **453**, 98 (2008)
19. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
20. A.-L. Barabási, R. Albert, Science **286**, 509 (1999)
21. M.E.J. Newman, Phys. Rev. Lett. **89**, 208701 (2002)
22. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002)
23. T. Zhou, M. Zhao, G.-R. Chen, G. Yan, B.-H. Wang, Phys. Lett. A **368**, 431 (2007)
24. A. Arenas, A. Díaz-Guilera, C.J. Pérez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006)
25. F. Gobel, A. Jagers, Stochastic Processes and Their Applications **2**, 311 (1974)
26. P. Chebotarev, E. Shamis, Automation and Remote Control **58**, 1505 (1997)
27. J.A. Hanely, B.J. McNeil, Radiology **143**, 29 (1982)
28. C. Von Merging, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Nature **417**, 399 (2002)
29. M.E.J. Newman, Phys. Rev. E **74**, 036104 (2006)
30. R. Ackland, *Mapping the US political blogosphere: Are conservative bloggers more prominent*, *Presentation to BlogTalk Downunder* (Sydney, 2005), available at http://incsub.org/blogtalk/images/robertackland.pdf
31. N. Spring, R. Mahajan, D. Wetherall, T. Anderson, IEEE/ACM Trans. Networking **12**, 2 (2004)
32. V. Batageli, A. Mrvar, *Pajek Datasets*, available at http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm
33. V. Latora, M. Marchiori, Phys. Rev. Lett. **87**, 198701 (2001)
34. S. Maslov, K. Sneppen, Science **296**, 910 (2002)
35. J. Schmith, N. Lemke, J.C.M. Mombach, P. Benelli, C.K. Barcellos, G.B. Bedin, Physica A **349**, 675 (2005)
36. T. Zhou, B.-H. Wang, Y.-D. Jin, D.-R. He, P.-P. Zhang, Y. He, B.-B. Su, K. Chen, Z.-Z. Zhang, J.-G. Liu, Int. J. Mod. Phys. C **18**, 297 (2007)
37. G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval* (MuGraw-Hill, Auckland, 1983)
38. P. Jaccard, Bulletin de la Societe Vaudoise des Sciences Naturelles **37**, 547 (1901)
39. T. Sørensen, Biol. Skr. **5**, 1 (1948)
40. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Science **297**, 1553 (2002)
41. E.A. Leicht, P. Holme, M.E.J. Newman, Phys. Rev. E **73**, 026120 (2006)

42. M. Molloy, B. Reed, Random Structure Algorithms **6**, 161 (1995)
43. Y.-B. Xie, T. Zhou, B.-H. Wang, Physica A **387**, 1683 (2008)
44. Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries* (ACM Press, New York, 2005)
45. P. Holme, B.J. Kim, C.N. Yoon, S.K. Han, Phys. Rev. E **65**, 056109 (2002)
46. C.-Y. Yin, W.-X. Wang, G.-R. Chen, B.-H. Wang, Phys. Rev. E **74**, 047102 (2006)
47. G.-Q. Zhang, D. Wang, G.-J. Li, Phys. Rev. E **76**, 017101 (2007)
48. L.A. Adamic, E. Adar, Social Networks **25**, 211 (2003)
49. S. Zhou, R.J. Mondragón, New J. Phys. **9**, 173 (2007)
50. S. Zhou, R.J. Mondragón, IEEE Commun. Lett. **8**, 180 (2004)
51. V. Colizza, A. Flammini, M.A. Serrano, A. Vespignani, Nat. Phys. **2**, 110 (2006)
52. S.-H. Yook, A.-L. Barabási, H. Jeong, Proc. Natl. Acad. Sci. U.S.A. **99**, 13382 (2002)
53. E. Ravasz, A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003)
54. H.-K. Liu, T. Zhou, Acta Physica Sinica **56**, 106 (2007)
55. M.T. Gastner, M.E.J. Newman, Eur. Phys. J. B **49**, 247 (2006)
56. Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Phys. Rev. E **75**, 021102 (2007)
57. W. Li, X. Cai, Phys. Rev. E **69**, 046106 (2004)
58. A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, Proc. Natl. Acad. Sci. U.S.A. **101**, 3747 (2004)
59. T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Phys. Rev. E **76**, 046115 (2007)
60. T. Zhou, L.-L. Jiang, R.-Q. Su, Y.-C. Zhang, Europhys. Lett. **81**, 58004 (2008)
61. L. Lü, C.-H. Jin, T. Zhou, e-print `arXiv: 0905.3558`
62. B. Tadić, S. Thurner, G.J. Rodgers, Phys. Rev. E **69**, 036102 (2004)
63. F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, IEEE Trans. Knowl. Data. Eng. **19**, 355 (2007)
64. L. Katz, Psychmetrika **18**, 39 (1953)
65. D. Sun, T. Zhou, R.-R. Liu, C.-X. Jia, J.-G. Liu, B.-H. Wang, Phys. Rev. E **80**, 017101 (2009)
66. S. Brin, L. Page, Computer Networks and ISDN Systems **30**, 107 (1998)