PRAMANA — journal of physics

Vol. 82, No. 3
March 2014
pp. 571–583

Predicting the growth of new links by new preferential attachment similarity indices

KE HU[1], JU XIANG[2], XIAO-KE XU[3,*], HUI-JIA LI[4], WAN-CHUN YANG[5] and YI TANG[1]

[1]Hunan Key Laboratory for Micro-Nano Energy Materials and Devices, and Laboratory for Quantum Engineering and Micro-Nano Energy Technology, Xiangtan University, Xiangtan 411105, China
[2]Department of Basic Sciences, The First Aeronautical Institute of the Air Force, Xinyang 464000, China
[3]College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116605, China
[4]School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100080, China
[5]College of Information Engineering, Xiangtan University, Xiangtan 411105, China
*Corresponding author. E-mail: xiaokeeie@gmail.com

MS received 17 June 2013; revised 10 September 2013; accepted 21 October 2013
DOI: 10.1007/s12043-013-0634-0; ***e*Publication:** 7 March 2014

Abstract. By revisiting the preferential attachment (PA) mechanism for generating a classical scale-free network, we propose a class of novel preferential attachment similarity indices for predicting future links in evolving networks. Extensive experiments on 14 real-life networks show that these new indices can provide more accurate prediction than the traditional one. Due to the improved prediction accuracy and low computational complexity, these proposed preferential attachment indices can be helpful for providing both instructions for mining unknown links and new insights to understand the underlying mechanisms that drive the network evolution.

Keywords. Link prediction; preferential attachment; network evolution.

PACS Nos 89.75.Fb; 89.20.Ff; 89.75.Hc

1. Introduction

Link prediction, aiming at estimating the likelihood of the existence of a link between two agents based on observed links and the attributes of agents [1,2], is a rich subject with a wide scope of applications [3–7]. For example, link prediction can provide significant

instruction for mining missing interactions in real-life networks even without complete information [8–10]. Usually, the identification of missing links in experiment requires examining all the possible connections. This is both time-consuming and expensive, and sometimes it is unacceptable in practice. In this case, obtaining the prioritization of missing links by link prediction algorithms is helpful to largely reduce both the time consumption and experimental cost [3].

Link prediction algorithms are beneficial for understanding the evolution mechanism of real networks, for predicting links accurately in a network can provide important clues or evidences about the underlying rule that drives its evolution [7]. With the help of link prediction algorithms, one may construct a proper evaluation system to evaluate the evolving mechanism for given complex networks [6]. Starting from these facts, link prediction algorithms have been extended to broader applications. Recently, prediction algorithms have been successfully applied to the classification problem in partially labelled networks [11,12] as well as to the identification of the spurious links resulting from inaccurate information in the data [5].

In the past few years, a large number of methods have been developed for predicting unknown links in complex networks [13–21]. Generally, these methods are designed respectively based on two types of information: the external information besides the network topology such as the unit attributes [13–16] and the information from the network structure [17–21]. Because of the difficulty in getting external information, the information of network topology is usually preferable. Among all the topology-based methods, one of the simplest and effective algorithms is the one based on the preferential attachment (PA) index [17]. It is motivated by the popular PA mechanism in the famous evolving scale-free network model [22]. Thus, it is helpful for understanding underlying mechanism of network evolution, and also for predicting the growth of new links, without predicting the spurious links.

As is well known, the classical PA expresses that the interaction probability between new nodes and old ones is decided by the product form of the degree of related nodes, i.e., the pairwise interaction between nodes i and j is proportional to $k_i k_j$ (degree-product form). The legitimate mechanism of network evolution can provide significant clues for designing a proper predictive algorithm. Benefitted from the long-held and partially proved assumption of degree product form, Zhou *et al* [17] proposed the degree-product preferential attachment (DPPA) index to perform link prediction in complex networks. Although the DPPA index is rooted in the popular PA mechanism of network evolution, it cannot give optimal prediction for most experimental networks [17]. Moreover, the framework of link prediction need to not only explore interactions between the newly added node and the old ones [22], but also describe the interaction between two old nodes [23–25]. In some real networks, the formation of new links between two old nodes may not follow the degree product form accurately [26–29].

In this paper, we find that the traditional degree product form is not the most suitable one for the link prediction by analysing the PA mechanism comprehensively and systematically. Therefore, we develop a class of new PA indices. Extensive experiments on 14 real networks demonstrate that the new PA indices can give more accurate link prediction than the traditional DPPA index in most sample networks.

2. Datasets and method

2.1 *Datasets*

In this paper, 14 networks that are drawn from different fields are considered. These networks are simply described as follows:

(1) Collaboration network in computational geometry (CCG) [30]: It is an author-collaboration network in which nodes represent authors, and two authors are linked with a link if they wrote a common work (paper, book, etc.). Multiple links between two nodes represent multiple joint works they wrote. In this paper, we consider only unweighted networks, and thus multiple links between a pair of authors are replaced by a single link.
(2) Internet (INT) [31]: The Internet can be divided into subnetworks that are under separate administrative authorities. Here, we consider the Internet as an autonomous system (the interdomain level) where each domain is represented by a single node and each link is an interdomain interconnection.
(3) USAir [32]: The network of the US air transportation system in which nodes and links represent airports and airlines respectively.
(4) Protein–protein interaction network (PPI) [33]: A protein–protein interaction network in budding yeast with nodes representing proteins and links representing the interactions among proteins.
(5) FoodWeb (FW) [34]: A network of foodweb in Florida Bay during wet season. Each species is represented as a node of the network, and a link is placed between two species whenever one of them feeds on the other.
(6) NetScience (NS) [35]: A network of coauthorships between scientists who are themselves publishing on the topic of networks.
(7) The network of common adjective and noun adjacencies (CAN) [35]: It is the network of common adjective and noun adjacencies for the novel David Copperfield by Charles Dickens. Nodes represent the most commonly occurring adjectives and nouns in the book. Links connect any pair of words that occur in adjacent position in the book.
(8) World Wide Web (WWW) [36]: A large directed graph whose nodes are documents from University of Notre Dame (domain nd.edu) and whose edges are links (URLs) that point from one document to another.
(9) Gnutella peer-to-peer network (GP) [37]: A sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. It is a directed network and nodes represent hosts in the Gnutella network topology and links represent connections between the Gnutella hosts.
(10) Enron e-mail network (EE) [38]: Enron e-mail communication network covers all the e-mail communications within a dataset of around half million e-mails. Nodes of the network are e-mail addresses and if an address x sends at least one e-mail to address y, the graph generates a directed link from x to y. Note that non-Enron e-mail addresses act as sinks and sources in the network as we only observe their communication with the Enron e-mail addresses.

(11) Epinions social network (ES) [39]: This is a who-trust-whom online social network of the general consumer review site – Epinions.com. Members of the site can decide whether to 'trust' each other. All the trust relationships interact and form the Web of Trust which is then combined with review ratings to determine which reviews are shown to the user.
(12) Slashdot social network (SS) [40]: Slashdot is a technology-related news website known for its specific user community. The website features primarily user-submitted and editor-evaluated current technology-oriented news. In 2002 Slashdot introduced the Slashdot Zoo feature which allows users to tag each other as friends or foes. The network contains friend/foe links between the users of Slashdot.
(13) E-mail (EM) [41]: The e-mail network studied here is the email network of University at Rovirai Virgili in Tarragona, Spain, and is built by considering each e-mail address as a node and linking two nodes if there is an e-mail communication between them.
(14) Power grid (PG) [42]: An electrical power grid of the western US, with nodes representing generators, transformers and substations, and links corresponding to the high voltage transmission lines between them.

It should be mentioned that some of the above networks consist of many separated components but the sizes of the largest connected component are still very large relative to the whole networks (see the fourth column in table 1). So we only consider the giant connected component in all the networks. In addition, some networks mentioned above are directed and/or weighted. In this study, we focus on predicting links in undirected and unweighted networks, i.e., we do not consider the effects of direction and weights on link prediction. Thus, all the networks are treated as undirected and unweighted. The basic topological statistics of the 14 real-world networks are summarized in table 1.

2.2 *Preferential attachment similarity indices*

The preferential attachment similarity is very useful for link prediction because (1) it requires the least information since it only depends on the degrees of relative nodes and (2) the preferential attachment similarity is originated from the popular PA evolving mechanism of generating the scale-free network [22]. The traditional PA similarity index, i.e., the DPPA index, can be defined by the degree-product form as follows:

$$S_{xy} = k_x \times k_y, \tag{1}$$

where k_x and k_y are the degrees of nodes x and y respectively. This index has been used to quantify the functional significance of links subject to various network-based dynamics, such as percolation [43], synchronization [44] and transportation [45]. Applying the PA mechanism to link prediction, one may naturally think of the above degree-product form, but the DPPA index is poor in link prediction compared to other node-similarity indices [17–21]. So we argue that the DPPA index is not very reasonable for the link prediction in real networks, although it perhaps can represent the interaction between two nodes in some cases.

Table 1. The basic topological statistics of 14 experimental 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 3621/2091 in the first line means that the network has 2091 components and the giant component consists of 3621 nodes. M_C is the number of links belonging to the giant component. e is the network efficiency, defined by $e = (1/N(N-1))\sum_{x\neq y\in E}(1/d_{xy})$, where d_{xy} is the shortest path length between x and y and $d_{xy} = +\infty$ if x and y are in two different components. C is the clustering coefficient. r is the Pearson correlation coefficient of degrees: $r = \frac{M^{-1}\sum_s j_s k_s - [M^{-1}\sum_s \frac{1}{2}(j_s+k_s)]^2}{M^{-1}\sum_s \frac{1}{2}(j_s^2+k_s^2) - [M^{-1}\sum_s \frac{1}{2}(j_s+k_s)]^2}$, where j_s, k_s are the degrees of the nodes at the ends of the sth edge, with $s = 1, \ldots, M$. H is the degree heterogeneity, defined as $H = \langle k^2\rangle/\langle k\rangle^2$, where $\langle k\rangle$ denotes the average degree.

Networks	N	M	N_C	M_C	e	C	r	H
CCG	7343	11898	3621/2091	9461	0.051	0.408	0.243	4.706
INT	22963	48436	22963/1	48436	0.276	0.230	−0.198	61.978
USAir	332	2126	332/1	2126	0.406	0.625	−0.208	3.464
PPI	2361	7182	2224/161	6609	0.218	0.291	0.059	2.763
FW	128	2137	128/1	2106	0.624	0.335	−0.104	1.231
NS	1461	2742	379/268	914	0.016	0.694	0.462	1.849
CAN	112	425	112/1	425	0.442	0.173	−0.129	1.815
WWW	325729	1090108	325729/1	1090108	0.154	0.235	−0.053	41.934
GP	62586	147892	62561/12	147878	0.174	0.006	−0.093	2.455
EE	36692	183831	33696/1065	180811	0.221	0.497	−0.111	13.980
ES	75879	405740	75877/2	405739	0.245	0.138	−0.041	17.194
SS	82140	500481	82140/1	500481	0.257	0.059	−0.073	12.153
EM	1133	5451	1133/1	5451	0.300	0.220	0.078	1.942
PG	4941	6594	4941/1	6594	0.063	0.107	0.004	1.450

Generally, missing links can be divided into two types: unobserved links and future links, while the PA index focusses on prediction of the growth of new links (future links) in evolving networks. Therefore, to improve the accuracy of link prediction, we need to revisit the PA mechanism. As is well known, the growth of new links can happen in two possible ways: one is that the link is added between a newly added node and old ones [22] and the other is that the link is added between two old nodes. Here we give two types of new PA indices corresponding to the two types of links.

(1) *Links between newly added nodes and old ones*: A new node x is added along with several new links, each of which is randomly attached to an existing node y by the degree preferential attachment probability,

$$\Pi_{x\to y}^{\text{new}-\text{old}} = \frac{k_y}{\sum_j k_j}. \tag{2}$$

Clearly, this probability is proportional to the degree of the existing node y, independent of the degree of the newly added node x. Note that the degree preferential attachment probability is a conditional probability that states the link formation

from certain nodes (the newly added nodes) to old nodes in evolving network model, while not the preferential attachment probability of link formation between any pair of nodes. However, the DPPA index is designed to describe the similarity between any pair of nodes in link prediction, and thus it does not correspond to the preferential mechanism completely. In addition, as the degrees of the newly added nodes are generally smaller than the old nodes, the new nodes and the old nodes can be distinguished by their degrees at the moment of link formation. Based on the above analysis, a complete counterpart to the traditional PA mechanism and a more reasonable form of the PA index in this case can be designed as

$$S_{xy} = \max(k_x, k_y). \tag{3}$$

For convenience, we name this index as the high-degree node determine preferential attachment (HDPA) index.

(2) *Links between old nodes*: In addition to the above new link addition from new nodes, another way cannot be neglected: a large number of new links can appear between old nodes as the network evolves. Such internal links are often also subject to preferential attachment. Following the above link formation from new nodes to old nodes, a link between two old nodes x and y may be constructed by three ways: (a) preferential attachment from node x to node y, (b) preferential attachment from node y to node x or (c) both. The probabilities of link formation corresponding to the three cases are

$$\Pi_{x\to y} = \frac{k_y}{\sum_j k_j}, \qquad \Pi_{y\to x} = \frac{k_x}{\sum_j k_j} \qquad \text{and} \qquad \Pi_{x\leftrightarrow y} = \frac{k_x k_y}{\sum_j k_j \sum_j k_j},$$

respectively. Then the corresponding similarity scores for link prediction can be written as

$$S_{xy}^{x\to y} = k_y, \quad S_{yx}^{y\to x} = k_x \quad \text{and} \quad S_{xy}^{x\leftrightarrow y} = k_x k_y.$$

In order to make their scores comparable, we rescale the first two indices as

$$S_{xy}^{x\to y} = k_y^2 \qquad \text{and} \qquad S_{yx}^{y\to x} = k_x^2,$$

since the rescaling does not change the order of the similarity scores of links. Then we integrate them into one by simple weighted summarization:

$$S_{xy} = ak_x^2 + bk_y^2 + ck_x k_y, \tag{4}$$

where a, b and c are three adjustable parameters for controlling the relative contributions of the three indices to the integrated one. Consider the symmetry of link formation from node x to y and from y to x and without lack of generality, we can set $a = b = \varepsilon$ and $c = 1$, and obtain general internal-links preferential attachment (GILPA) index:

$$S_{xy} = \varepsilon k_x^2 + \varepsilon k_y^2 + k_x k_y. \tag{5}$$

Although one can tune ε to find its optimal value corresponding to the highest accuracy for any given network, the optimal value of ε is different for different networks. Furthermore, a parameter-dependent measure is less practical in dealing

with huge-size networks since the tuning process may take much time. Here we only consider several special values of ε. Clearly, the index reduces to the traditional DPPA index when $\varepsilon = 0$. Particularly, when $\varepsilon = 0.5$, we rescale the GILPA, and then obtain a very simple similarity measure:

$$S_{xy} = k_x + k_y. \tag{6}$$

We name it as the degree-summation preferential attachment (DSPA) index. Another limiting case is $\varepsilon \to +\infty$, which corresponds to a degree-squared-summation preferential attachment (DSSPA) index:

$$S_{xy} = k_x^2 + k_y^2. \tag{7}$$

In the above discussion, we have divided the link formations into two types according to the rules of their addition and present several new similarity indices to predict these links. These indices are clearly different from the DPPA index. This is because the PA probabilities that they depend on (which is not the degree-product form) are different from that between two old nodes in random networks with given degree distribution.

2.3 *Evaluation metrics*

Let $G(V, E)$ be a simple undirected and unweighted network, which is described by the sets of nodes V and links E. Multiple links and self-connections are excluded from E. Every algorithm referred in this paper will be assigned a similarity matrix $\mathbf{S}$ whose real entry S_{xy} expresses how similar the node x is to node y: we say that S_{xy} is their similarity score. For each pair of nodes x and y ($x, y \in V$), $S_{xy} = S_{yx}$ since the networks are undirected. All the nonexistent links are sorted in a decreasing order according to their similarity scores, and the links at the top are the most likely to exist.

To quantify the prediction accuracy, the set of the observed links E is randomly divided into two parts: the training set E^{T} and the probe set E^{P}. The training set is treated as known information, while the probe set will be predicted and no information in this set is allowed to be used for prediction. Clearly, $E = E^{\mathrm{T}} \cup E^{\mathrm{P}}$ and $\phi = E^{\mathrm{T}} \cap E^{\mathrm{P}}$. In this study, the training set always contains 90% of links and naturally the remaining 10% of the links constitute the probe set. The prediction quality is then evaluated by the standard metrics: the area under the receiver operating characteristic curve (AUC) [46]. In the present case, the AUC can be interpreted as the probability that a randomly chosen missing link (a link in E^{P}) is given a higher similarity score than a randomly chosen nonexistent link (a link in $U - E^{\mathrm{T}}$, 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

$$\mathrm{AUC} = \frac{n' + 0.5n''}{n}. \tag{8}$$

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

3. Results

We give the numerical results of these similarity indices in the 14 sample networks in table 2. Due to the different topological properties of these networks, it is not unexpected that the PA indices (DPPA, HDPA, DSSPA, DSPA and GILPA) give distinct link-prediction accuracies. In particular, the performances of these indices strongly depend on the degree heterogeneity. If all the nodes in a given network have pretty much the same degree (a very small H), then the PA indices will give relatively bad predictions. On the contrary, the larger is the degree heterogeneity, the higher is the prediction accuracy, which can be seen in figure 1a. This indicates that the degree heterogeneity has an important implication to the PA mechanism, or more specifically, the scale-free property implies preferential attachment [22,47]. In addition, one may intuitively think that these PA indices will give good predictions for assortative networks (i.e., $r > 0$), while performing badly for disassortative networks (i.e., $r < 0$). However, no obvious and direct correlation between assortative coefficient and algorithmic accuracy based on these PA indices can be found from our numerical results. Moreover, from figure 1b, one may find that the predictive accuracies of these PA indices are even somewhat higher in the disassortative networks than those in the assortative ones. The reason is two-fold. Firstly, links between pairs of high-degree nodes contribute positively to the assortative coefficient and are assigned high scores by these PA indices, while links between pairs of low-degree nodes also contribute positively to the assortative coefficient but are disfavoured by them. Actually, the assortative coefficient is an integrated measure involving many ingredients, and there is no simple relation between this measure and the performance of these PA indices. Secondly, the assortative coefficient itself is very sensitive to the degree sequence, and a network of higher degree heterogeneity tends to be disassortative. Our results presented in figure 1a indicate that these PA indices seem favourable for

Table 2. Accuracies of algorithms, measured by AUC. Each number is obtained by averaging over 10 implementations with independently random partitions of testing set and probe set.

Networks	DPPA	HDPA	DSSPA	DSPA	GILPA ($\varepsilon_{\text{optimal}}$)
CCG	0.7222	0.7681	0.7684	0.7707	0.7720 ($\varepsilon = 0.67$)
INT	0.7722	0.9560	0.9531	0.9481	0.9531 ($\varepsilon > 10.0$)
USAir	0.8789	0.8537	0.8670	0.8800	0.8861 ($\varepsilon = 0.28$)
PPI	0.7808	0.8130	0.8202	0.8285	0.8302 ($\varepsilon = 0.23$)
FW	0.6912	0.6572	0.6733	0.6913	0.6913 ($\varepsilon = 0.50$)
NS	0.6368	0.6458	0.6499	0.6556	0.6574 ($\varepsilon = 0.42$)
CAN	0.7392	0.7113	0.7120	0.7548	0.7548 ($\varepsilon = 0.50$)
WWW	0.8167	0.9157	0.9218	0.9252	0.9262 ($\varepsilon = 0.24$)
GP	0.7171	0.7903	0.8116	0.8244	0.8269 ($\varepsilon = 0.25$)
EE	0.8792	0.9033	0.9089	0.9142	0.9246 ($\varepsilon = 0.02$)
ES	0.8849	0.9333	0.9344	0.9368	0.9374 ($\varepsilon = 0.21$)
SS	0.9007	0.9234	0.9283	0.9305	0.9360 ($\varepsilon = 0.07$)
EM	0.7820	0.7447	0.7601	0.7775	0.7872 ($\varepsilon = 0.04$)
PG	0.4421	0.5428	0.5242	0.5028	0.5242 ($\varepsilon > 10.0$)

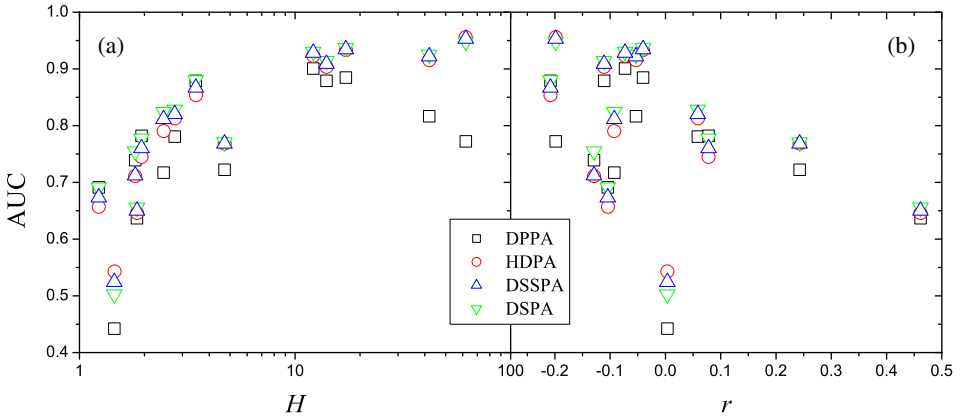

Figure 1. AUC accuracies of the PA indices vs. (**a**) the degree heterogeneity H, (**b**) the Pearson correlation coefficient.

disassortative networks. Therefore, in the disassortative networks, the accuracies of these PA indices may be higher than those in the assortative ones.

When comparing the accuracies of different PA indices in the same network, one can find that the DSPA and DSSPA indices clearly outperform the DPPA index. For example, the accuracies of DSPA are much higher than DPPA index (about 22% in the networks of INT, WWW and GP). In addition, the HDPA index is very simple and depends on the degree of single node, but it is still effective and even better than the DPPA index sometimes. According to the definitions of these PA indices, significant difference about similarity scores does not lie in the links between pairs of high-degree nodes since these links are assigned high scores by all these indices. However, the DPPA index usually endows the assortative links (i.e., the links between pairs of the same degree level) with relatively high scores, while the HDPA, DSPA and DSSPA indices can give competitive scores for both the disassortive links (i.e., links between high-degree nodes and low-degree nodes) and the assortative links between high-degree nodes. For example, we assume that the two nodes connected by an assortative link L_{ij} have the same degrees of $k_i = k_j = 20$; and the two nodes connected by a disassortative link $L_{i'j'}$ have very different degrees, $k_{i'} = 2$ and $k_{j'} = 100$. Obviously, the score of L_{ij} is higher than that of $L_{i'j'}$ in the DPPA index, while in the HDPA, DSPA and DSSPA indices, it is lower than that of $L_{i'j'}$. Since many real networks possess scale-free property, there are a large number of links between high-degree nodes and other nodes. Especially in some networks with the strong heterogeneity and disassortative correlations, the probability that a link exists between high-degree nodes and other nodes may be larger than the probability between pairs of medium-degree nodes. In order to demonstrate this point, we define the degree–degree joint density as

$$d_{k,k'} = \frac{L'_{k,k'}}{L_{k,k'}}, \tag{9}$$

where $L'_{k,k'}$ and $L_{k,k'}$ denote the number of existing links and all possible links between nodes with degree k and k', respectively. If $k \neq k'$, $L_{k,k'} = N_k N_{k'}$; if $k = k'$,

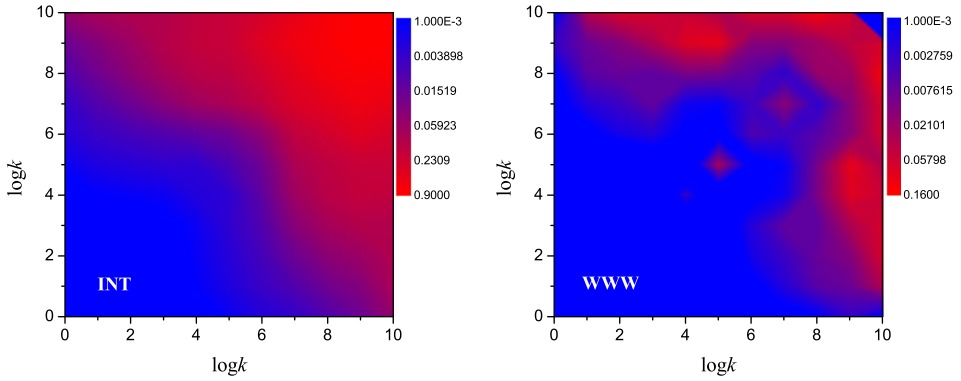

Figure 2. Degree–degree joint density between nodes with different degree levels in two sample networks, INT and WWW.

$L_{k,k'} = N_k(N_k - 1)/2$, where N_k is the number of nodes with degree k. Obviously, the degree–degree joint density reflects the probability that nodes with degree k and degree k' are connected in real-world networks. Now let us look at the degree–degree joint densities for two real networks with both the strong degree heterogeneity and the disassortative degree correlations, INT and WWW, which is presented in figure 2. If we roughly divide the densities into two regions, the assortative link region and the disassortative one, one can find that the density of the assortative link region, except for the region on the high degrees, is even lower than that of the disassortative link region. This indicates that the probability that a link exists between high-degree nodes and medium-degree nodes is larger than the probability between pairs of medium-degree nodes. Thus, for link prediction, the HDPA, DSPA and DSSPA indices are more reasonably designed by defining the disassortative links with high similarity scores than the traditional DPPA index. In short, if the investigated network simultaneously has large degree heterogeneity and disassortative correlation, such as INT, WWW, GP, EE and ES, both the HDPA, DSPA and DSSPA indices perform better than the DPPA index. Moreover, the effect of heterogeneity is relatively more remarkable than the disassortative correlation. In general, higher is the heterogeneity of a network, higher is the accuracies of the HDPA, DSPA and DSSPA indices. For instance, the INT has extremely large degree heterogeneity, and so the performances of the HDPA, DSPA and DSSPA algorithms (AUC is about 0.95) are remarkably better than that of the DPPA one (AUC is about 0.77).

From the perspective of the rules of network evolution, the results in table 2 are also intuitively reasonable. If a mechanism of link formation can properly model the link formation in real complex networks, one can construct a kind of link-prediction algorithm in the networks; conversely a high-accuracy link-prediction algorithm in networks may suggest a possible mechanism of dominating the evolution of the networks [6]. As we see in table 2, the HDPA index has good performance and can give higher accuracies than the DPPA in some networks. The effectiveness of the HDPA index gives an indirect evidence that the networks may be organized under the preferential attachment mechanism and the HDPA can capture the preferential attachment mechanism in these networks better than the DPPA.

According to the analysis in §2.2, the DSPA index can be viewed as a special case of the general similarity index for internal links by eq. (5), and the DPPA and DSSPA indices are its two limiting cases. Among the three indices (DPPA, DSSPA and DSPA), the DSPA index shows the best performance on most of the sample networks. Thus, we think the DSPA index can more completely reflect the mechanism of internal link formation in these networks than the DPPA and DSSPA indices. In addition, it should be pointed out that the degree-summation form (DSPA) may be the simplest and efficient form for the general similarity index (GILPA) given by eq. (5), but it is not the highest-accuracy one among these PA indices. Although better link prediction accuracy may be obtained by searching the optimal value of the parameter ε in GILPA (see the last column in table 2), this paper does not aim at highlighting the parameter-dependent index. Instead, we attempted to uncover the simplest mechanism of internal link formation. Moreover, the results also show that the additional computational efforts of searching the optimal ε values do not remarkably improve the accuracies of the DSPA index in all sample networks. It indicates that the DSPA index perhaps provides a reasonable mechanism for internal link formation in the networks.

According to our assumption, we should apply the HDPA index for the links between new nodes and old nodes, and apply the DSPA index for the internal links between old nodes. Generally, besides the link formation from new nodes to old ones, many networks have the link formation from old nodes to new ones and between old nodes. It is usually difficult to distinguish the two types of links in most real networks. Thus, we do not know which index should be better to perform the link prediction in the networks. In this case, the DSPA may be a good choice. Because the degrees of new nodes are often smaller than that of old nodes, the DSPA can partially reflect the behaviour of the HDPA in capturing the link formation from new nodes to old ones. The DSPA can be considered as a trade-off between two types of link formations in networks.

4. Conclusions

In summary, based on the popular network evolution mechanism, i.e., the PA mechanism, we have developed a class of new PA similarity indices to predict the growth of new links. By applying them to 14 real networks, we have shown that the proposed indices can provide more accurate predictions than the traditional DPPA index, especially in the networks with large degree heterogeneity and disassortative degree correlation. Moreover, the computational complexity of these indices is almost the same as or lower than the DPPA index. So they can provide more competitively efficient prediction.

In addition, owing to the strong correlation between the algorithm of link prediction and the mechanism of network evolution [6], this work is also helpful to understand the mechanism of network evolution, especially for the formation mechanism of the internal links. In principle, the rules of the additions of links can be considered as a kind of link prediction algorithm, and so it is easy to build a bridge between the link prediction and the network evolving model [6]. A proper link prediction method (such as DSPA in this study) may give new insights to some underlying mechanisms that drive the network evolution.

Finally, we also note that for the problem of link prediction in weighted networks, the link weight has been recently considered and significant improvements of link-prediction accuracies are presented [19]. By replacing the node's degree with its strength, these PA indices can be also easily extended to the corresponding weighted versions. We hope that the introduction of strengths into these PA algorithms can further improve the link prediction performance of weighted networks, which will be investigated in our future works.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos 6104104, 11147121 and 61104143), the Scientific Research Fund of Education Department of Hunan Province (Grant No. 11B128) and Natural Science Foundation of Hunan Province (Grant No. 13JJ4045). The authors are also grateful for the comments and suggestions from the anonymous referee.

References

[1] L Getoor and C P Diehl, *ACM SIGKDD Explorations Newsletter* **7**, 3 (2005)
[2] L Lü and T Zhou, *Physica A* **390**, 1150 (2011)
[3] A Clauset, C Moore and M E J Newman, *Nature* **453**, 98 (2008)
[4] S Redner, *Nature* **453**, 47 (2008)
[5] R Guimerà and M Sales-Pardo, *Proc. Natl. Acad. Sci.* **106**, 22073 (2009)
[6] W Q Wang, Q M Zhang and T Zhou, *Europhys. Lett.* **98**, 28004 (2012)
[7] H K Liu, L Lü and T Zhou, *Sci. China Ser. G* **41**, 816 (2011)
[8] H Yu, P Braun, M A Yildirim, I Lemmens, K Venkatesan, J Sahalie, T Hirozane-Kishikawa, F Gebreab, N Li, N Simonis, T Hao, J F Rual, A Dricot, A Vazquez, R R Murray, C Simon, L Tardivo, S Tam, N Svrzikapa, C Fan, A S Smet, A Motyl, M E Hudson, J Park, X Xin, M E Cusick, T Moore, C Boone, M Snyder, F P Roth, A L Barabási, J Tavernier, E E Hill and M Vidal, *Science* **322**, 104 (2008)
[9] M P H Stumpf, T Thorne, E de Silva, R Stewart, H J An, M Lappe and C Wiuf, *Proc. Natl. Acad. Sci.* **105**, 6959 (2008)
[10] L A N Amaral, *Proc. Natl. Acad. Sci.* **105**, 6795 (2008)
[11] B Gallagher, H Tong, T Eliassi-Rad and C Faloutsos, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 256 (2008)
[12] Q M Zhang, M S Shang and L Lü, *Int. J. Mod. Phys. C* **21**, 813 (2010)
[13] R R Sarukkai, *Computer Networks* **33**, 377 (2000)
[14] J Zhu, J Hong and J G Hughes, *Proceedings of the First International Conference on Computing in an Imperfect World*, 60 (2002)
[15] A Popescul and L H Ungar, *IJCAI Workshop on Learning Statistical Models from Relational Data*, 109 (2003)
[16] K Yu, W Chu, S Yu, V Tresp and Z Xu, *Adv. Neural Inform. Processing Systems* **19**, 1553 (2007)
[17] T Zhou, L Lü and Y C Zhang, *Eur. Phys. J. B* **71**, 623 (2009)
[18] L Lü, C H Jin and T Zhou, *Phys. Rev. E* **80**, 046122 (2009)
[19] L Lü and T Zhou, *Europhys. Lett.* **89**, 18001 (2010)
[20] W P Liu and L Lü, *Europhys. Lett.* **89**, 58007 (2010)

[21] Z Liu, Q M Zhang, L Lü and T Zhou, *Europhys. Lett.* **96**, 48007 (2011)
[22] A L Barabási and R Albert, *Science* **286**, 509 (1999)
[23] S N Dorogovtsev and J F F Mendes, *Europhys. Lett.* **52**, 33 (2000)
[24] R Albert and A L Barabási, *Phys. Rev. Lett.* **85**, 5234 (2000)
[25] W X Wang, B H Wang, B Hu, G Yan and Q Ou, *Phys. Rev. Lett.* **94**, 188702 (2005)
[26] P L Krapivsky and S Redner, *Computer Networks* **39**, 261 (2002)
[27] V Rosato and F Tiriticco, *Europhys. Lett.* **66**, 471 (2004)
[28] B Tadic, *Physica A* **293**, 273 (2001); **314**, 278 (2002)
[29] A L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert and T Vicsek, *Physica A* **311**, 590 (2002)
[30] V Batagelj and M Zaveršnik, *Pajek Datasets*, available at http://vlado.fmf.uni-lj.si/pub/networks/data/collab/geom.htm
[31] This snapshot was created by M E J Newman from data for July 22, 2006 and is not previously published, available at http://www-personal.umich.edu/~mejn/netdata/
[32] V Batagelj and A Mrvar, *Pajek Datasets*, available at http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm
[33] D B Bu, Y Zhao, L Cai, H Xue, X P Zhu, H C Lu, J F Zhang, S W Sun, L J Ling, N Zhang, G J Li and R S Chen, *Nucleic Acids Research* **31**, 2443 (2003)
[34] C J Melián and J Bascompte, *Ecology* **85**, 352 (2004)
[35] M E J Newman, *Phys. Rev. E* **74**, 036104 (2006)
[36] R Albert, H Jeong and A L Barabási, *Nature* **401**, 130 (1999)
[37] M Ripeanu, I Foster and A Iamnitchi, *IEEE Internet Computing Journal* **2429**, 85 (2002)
[38] J Leskovec, J Kleinberg and C Faloutsos, *Proceedings of the ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, 177 (2005)
[39] M Richardson, R Agrawal and P Domingos, *Proceedings of the Second International Semantic Web Conference*, 351 (2003)
[40] J Leskovec, K Lang, A Dasgupta and M Mahoney, arXiv:cs.DS/0810.1355
[41] R Guimerà, L Danon, A Díaz-Guilera, F Giralt and A Arenas, *Phys. Rev. E* **68**, 065103 (2003)
[42] D J Watts and S H Strogatz, *Nature* **393**, 440 (1998)
[43] P Holme, B J Kim, C N Yoon and S K Han, *Phys. Rev. E* **65**, 056109 (2002)
[44] C Y Yin, W X Wang, G R Chen and B H Wang, *Phys. Rev. E* **74**, 047102 (2006)
[45] G Q Zhang, D Wang and G J Li, *Phys. Rev. E* **76**, 017101 (2007)
[46] J A Hanely and B J McNeil, *Radiology* **143**, 29 (1982)
[47] K A Eriksen and M Hörnquist, *Phys. Rev. E* **65**, 017102 (2001)