Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

Q1 Link prediction with node clustering coefficient

Q2 Zhihao Wu [a,*], Youfang Lin [a], Jing Wang [a], Steve Gregory [b]

[a] Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, People's Republic of China

[b] Department of Computer Science, University of Bristol, Bristol, BS8 1UB, United Kingdom

HIGHLIGHTS

- Node clustering coefficient is used to propose a novel similarity index.
- An illustration of hierarchy of local similarity indices is given for the first time.
- The proposed CCLP index is parameter free.

ARTICLE INFO

Article history:
Received 27 August 2015
Received in revised form 29 December 2015
Available online xxxx

Keywords:
Link prediction
Complex networks
Clustering coefficient

ABSTRACT

Predicting missing links in incomplete complex networks efficiently and accurately is still a challenging problem. The recently proposed Cannistrai–Alanis–Ravai (CAR) index shows the power of local link/triangle information in improving link-prediction accuracy. Inspired by the idea of employing local link/triangle information, we propose a new similarity index with more local structure information. In our method, local link/triangle structure information can be conveyed by clustering coefficient of common-neighbors directly. The reason why clustering coefficient has good effectiveness in estimating the contribution of a common-neighbor is that it employs links existing between neighbors of a common-neighbor and these links have the same structural position with the candidate link to this common-neighbor. In our experiments, three estimators: precision, AUP and AUC are used to evaluate the accuracy of link prediction algorithms. Experimental results on ten tested networks drawn from various fields show that our new index is more effective in predicting missing links than CAR index, especially for networks with low correlation between number of common-neighbors and number of links between common-neighbors.

1. Introduction

Complex network has shown its significant power in modeling and analyzing a wide range of complex systems, such as social, biological and information systems, and the study of complex networks has attracted increasing attention and becomes a popular tool in many different branches of science [1–5]. Prediction is one of the key problems in various research and application fields. Link prediction in complex networks aims at estimating the likelihood of the existence of a link between two nodes, and it has many applications in different fields. For example, predicting whether two users know each other can be used to recommend new friends in Social Networking Sites, and in the field of biology, accurate prediction of protein–protein interaction has great value to sharply reduce the experimental costs. Some researchers also

* Corresponding author.
E-mail address: zhwu@bjtu.edu.cn (Z. Wu).

http://dx.doi.org/10.1016/j.physa.2016.01.038
0378-4371/

applied the link prediction algorithms in partially labeled networks for prediction of protein functions or research areas of scientific publication [6,7]. In addition, the study of link prediction is closely related to the problem of network evolving mechanisms [8,9]. Qianming Zhang and Tao Zhou et al. employed link prediction methods to evaluate network models and attained better results than some classical models [8]. Recently, through measuring multiple evolution mechanisms of complex networks, they found the evolution of most networks is affected by both popularity and clustering at the same time, but with quite different weights [9].

Many link prediction methods have been proposed under different backgrounds in recent years [10,11]. In this paper, we only focus on similarity-based methods using topology structural information. The basic assumption for this kind of link prediction methods is that two nodes are more likely to have a link if they are similar to each other. Therefore, the key problem is to define proper similarity measures between nodes. Some methods combine many factors to define the similarity between nodes, such as attributes of nodes and links and structural information. One group of similarity indices is based solely on the network structure. The simplest one is PA index [12], which is defined as the product of degrees of two seed nodes. Common-Neighbor (CN) [13] counts the number of common-neighbors and Jaccard index (JC) [14] is a normalization of CN. To get better resolution, Adamic–Adar (AA) [15] and Resource Allocation (RA)[16] are defined by employing the degree information of common-neighbors. Recently, a new index, called Cannistrai–Alanis–Ravai (CAR) [17], is proposed by Cannistraci et al. Their main point is that link information of common-neighbors is useful but still noisy. They find level-2 links, i.e. links between common-neighbors, are more valuable and can be used to improve most classical Node-Neighborhood-based similarity indices. The above methods are all local measures, and to pursue higher prediction precision some global and quasi-local methods are also proposed, such as Katz [18], SimRank [19], Hitting Time [20], Average Commute Time [21], Local Path [22], Transferring Similarity [23], Matrix Forest Index [24] and so on. Obviously, considering more information and features in prediction methods may cause more time and space costs.

Besides, there are also some more complex models and methods to solve the link prediction problem. Clauset et al. proposed an algorithm based on the hierarchical network structure, which gives good predictions for the networks with hierarchical structures [25,26]. Guimera et al. solved this problem using stochastic block model [27]. Recently, Linyuan Lü et al. proposed a concept of structural consistency, which could reflect the inherent link predictability of a network, and they also proposed a structural perturbation method for link prediction, which is more accurate and robust than the state-of-the art method [28]. Although the above methods can attain better results than most Node-Neighborhood-based methods, they are hard to be applied to large networks.

Heretofore, efficient link prediction is still a big challenge. In our opinion, local methods are still good candidates for solving link prediction problem in large networks. Some results have shown that community/cluster structures can help improve the performance of link prediction [29,17]. Some researchers directly combine the communities detected by various community detection algorithms with some similarity indices, and show that cluster information can improve link prediction algorithms a lot in some cases [30,31]. This kind of methods relies on the community detection algorithms, but there are lots of different algorithms. How to choose a proper algorithm is still not very clear.

In this paper, we present a new similarity index, called CCLP (Clustering Coefficient for Link Prediction), which employs more local link/triangle structure information than CAR index, but costs less computational time. The key idea of our method is to exploit the value of links between other neighbors of common-neighbors, except seed nodes and common-neighbor nodes, and they can be efficiently conveyed by using clustering coefficient of common-neighbors. Some related literatures also suggest that clustering coefficient has some relations with the problem of link prediction problem [32,33]. The experimental results on 10 networks from five various fields show that our new method performs better than CAR index on networks with not very high LCP_{-corr} and is more efficient.

2. Methods

2.1. Definition

Considering an unweighted undirected simple network $G(V, E)$, where V is the set of nodes and E is the set of links. For each pair of nodes, $x, y \in V$, we assign a score to the pair of seed nodes. All the nonexistent links are sorted in decreasing order according to their scores, and the links in the top are most likely to exist. The common-used framework always sets the similarity to the score, so the higher score means the higher similarity, and vice versa. The definitions of similarity indices mentioned in this paper are as follows.

Preferential attachment (PA).

$$PA(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| \tag{1}$$

where $\Gamma(x)$ denotes the set of neighbors of node x and $|A|$ is the number of elements in set A.

Common neighbor (CN).

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)|. \tag{2}$$

Jaccard (JC).

$$JC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}. \tag{3}$$

Please cite this article in press as: Z. Wu, et al., Link prediction with node clustering coefficient, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.01.038