Physica A 454 (2016) 24–33

Contents lists available at ScienceDirect

Physica A

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

The degree-related clustering coefficient and its application to link prediction

Yangyang Liu, Chengli Zhao, Xiaojie Wang, Qiangjuan Huang, Xue Zhang, Dongyun Yi*

College of Science, National University of Defense Technology, Changsha, Hunan, China

HIGHLIGHTS

- We propose the degree-related clustering coefficient to estimate the cluster ability of nodes.
- The degree-related clustering coefficient exhibits a high robustness in estimating the cluster ability of nodes when the observed bias of links is considered.
- The DCP algorithm is proposed to achieve high accuracy and robustness on link prediction.
- The discussion about parameter setting improves our algorithm's efficiency.
- The stability analysis shows that our index is stable when measuring node similarity.

ARTICLE INFO

Article history:
Received 23 September 2015
Received in revised form 18 January 2016
Available online 3 March 2016

Keywords:
Link prediction
Complex networks
Networks evolution
Clustering coefficient

ABSTRACT

Link prediction plays a significant role in explaining the evolution of networks. However it is still a challenging problem that has been addressed only with topological information in recent years. Based on the belief that network nodes with a great number of common neighbors are more likely to be connected, many similarity indices have achieved considerable accuracy and efficiency. Motivated by the natural assumption that the effect of missing links on the estimation of a node's clustering ability could be related to node degree, in this paper, we propose a *degree-related clustering coefficient* index to quantify the clustering ability of nodes. Unlike the classical clustering coefficient, our new coefficient is highly robust when the observed bias of links is considered. Furthermore, we propose a *degree-related clustering ability path* (DCP) index, which applies the proposed coefficient to the link prediction problem. Experiments on 12 real-world networks show that our proposed method is highly accurate and robust compared with four common-neighbor-based similarity indices (Common Neighbors(CN), Adamic-Adar(AA), Resource Allocation(RA), and Preferential Attachment(PA)), and the recently introduced *clustering ability* (CA) index.

1. Introduction

Complexity is a common phenomenon in both nature and human society. In recent years, researchers have found that interactions among systems can be clearly and concisely expressed through networks, in which individuals are represented by nodes, and relationships between individuals are represented by edges. Research on complex networks has produced

[*] Corresponding author.
E-mail address: 895355449@qq.com (D. Yi).

http://dx.doi.org/10.1016/j.physa.2016.02.014
0378-4371/© 2016 Elsevier B.V. All rights reserved.

many important achievements, and received sustained attention [1–4]. After more than ten years of development, network science has become a significant interdisciplinary field, involving information technology [5], social sciences [6], biology [7], physics [8], epidemiology [9–11] and other disciplines. Consequently, many new branches have been studied; one of the important branches focuses on understanding the evolution process of networks, as well as the impact of network topology structures on function [12–14]. Among the various topics, link prediction has been viewed as the fundamental issue [15]. Link prediction focuses on techniques for estimating the likelihood of nonexistent links by analyzing the observed links; it is also a process of information retrieval [16]. Link prediction is of significant applied value in practical use. First, it is difficult to obtain a complete network structure. For example, in research on protein–protein networks [17], correlations can be identified through experiments. Nevertheless, the high cost and complexity make it almost impossible to build a complete network frame. This prompts us to consider methods to exploit the known links to predict the potential links. Furthermore, link prediction can also be applied in the recommended system [18] and the reconstruction of traffic networks [19].

Link prediction has been extensively studied by numerous scholars, and some efficient algorithms have already been created. The basic principle of most conventional algorithms is based on assigning a "similarity index" to a pair of unconnected network nodes, and calculating the likelihood that the pair is connected, according to the index. Generally, similarity indices can be divided into three classes. The first class calculates a similarity index using the global link information in networks [5,20]; this class includes the renowned *Katz* index, which sums all paths connecting the pair of nodes in a weighted manner [21]. The second class uses local information to calculate the indices, which reduces the complexity of the calculation. For example, the well-regraded Common Neighbors (CN) [22] method simply counts the common neighbors of the pairs, and provides satisfactory results. Several improvements on the CN index have mainly involved the type of penalization used large-degree nodes (e.g. Salton Index [16], Resource Allocation Index [23], Hub Depress Index [24], Adamic-Adar Index [8], etc.). Although the local information approaches greatly reduce computational complexity, prediction accuracy also decreases. Thus, we seek a tradeoff between prediction accuracy and computational complexity, which is the main goal of the approaches in the third class. The third class of similarity metrics is based on quasi-local structures; these metrics include the local path index, described in Refs. [23,25]. In addition, some researchers have dedicated themselves to estimating the probability of nonobserved links with sophisticated models. Among these methods, the Local Naïve Bayes model (LNB) [26], the Hierarchical Structure Model [27] and Stochastic Block Model [28] are good representatives. By applying probability modeling to the observed network structure, prediction problems are transformed into model parameter optimization problems. However, from the viewpoint of practical applications, the largest challenge is that these methods are very time consuming. In some cases, the probabilistic approaches' prediction accuracy was insufficient compared with that of the similarity approaches.

Recently, some scholars attempted to combine more information about the feature of networks and obtained high performance in the specific networks. The community structure information of the networks was considered to provide insights for link prediction, and received a good performance in the hierarchical structure networks [29]. A novel link prediction index called Neighbor Set Information (NSI) was designed through measuring the contributions of different structural features to link prediction [30]. By using both the temporal information and topological information, a method was proposed for link prediction in time-varying networks [31]. Beyond that, rather than defining predictability as the possible maximum precision of a prediction algorithm [32], Lü et al. characterized predictability as the inherent measurement of network topology, namely its so-called "structural consistency". Based on this concept, a new method of structural perturbation was proposed for link prediction; this method was proved to be more accurate and robust than the state-of-the-art methods. The structural consistency index solved a problem involve the selection of algorithms, and provided significant insights into the mechanisms of networks. However, defining a structural consistency index through a first-order random perturbation matrix introduces a certain randomness. Further, the computational complexity is somewhat high. Wu et al. proposed a more efficient index called *clustering ability* (CA) based on clustering coefficient. It is defined in Eq. (1)

$$S_{xy}^{CA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \overline{C(k_z)}. \tag{1}$$

In the preceding equation, $\overline{C(k_z)}$ is the average clustering coefficient of nodes with a degree equal to k_z. The clustering coefficient of a node is defined in Eq. (2):

$$C_i = \frac{2t_i}{k_i(k_i - 1)}. \tag{2}$$

Here, t_i is the number of triangles containing node i, and k_i is the degree of node i.

The CA index outperformed state-of-the-art common-neighbor based similarity indices in terms of precision, especially on sparse networks with low average clustering coefficients [33]. However, the mean-field method reduces the distinction between node similarity indices to an unacceptable degree, leading to poor performance on some networks. Research efforts must be undertaken to make these approaches more practical.

The clustering coefficient indicates the degree to which a node's neighbors are clustered. Thus, a pair of nodes which have a high clustering coefficient common neighbor is more likely to be linked, because nodes with a high clustering coefficient haves the ability to cluster their neighbors together. Unfortunately, because the networks we obtained always