Physica A 512 (2018) 902–912

Contents lists available at ScienceDirect

Physica A

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

An intermediary probability model for link prediction

Xuejun Zhang [a,b,c], Wenbo Pang [a,b,c], Yongxiang Xia [d,*]

[a] School of Electronic and Information Engineering, Beihang University, Beijing 100191, China
[b] Beijing Key Laboratory for Network-based Cooperative Air Traffic Management, Beijing 100191, China
[c] Beijing Laboratory for General Aviation Technology, Beijing 100191, China
[d] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

HIGHLIGHTS

- A novel intermediary probability (IMP) model for link prediction is proposed.
- New algorithms based on the IMP model achieve better accuracy.
- A new method is proposed to analyze features of link prediction indexes.

ARTICLE INFO

Article history:
Received 16 October 2017
Received in revised form 10 July 2018
Available online xxxx

Keywords:
Complex networks
Link prediction
Intermediary probability model
K-shell decomposition

ABSTRACT

Among the numerous link prediction algorithms in complex networks, similarity-based algorithms play an important role due to promising accuracy and low computational complexity. Apart from the classical CN-based indexes, several interdisciplinary methods provide new ideas to this problem and achieve improvements in some aspects. In this article, we propose a new model from the perspective of an intermediary process and introduce indexes under the framework, which show better performance for precision. Combined with k-shell decomposition, our deeper analysis gives a reasonable explanation and presents an insight on classical and proposed algorithms, which can further contribute to the understanding of link prediction problem.

1. Introduction

Link prediction, a problem focusing on estimating the likelihood of existence of link between two nodes based on available information [1,2], has attracted widespread attentions from several scientific communities. It is valuable and applicable in many disparate areas for its quite simple and general definition. For example, the mining of missing links contributes to friendship recommendation in online social networks, commodity recommendation for online shopping [3], and interaction mining for biological networks such as food webs, protein networks [4] and metabolic networks [5]. On the other hand, the prediction of future links provides a possible evaluation standard for numerous network evolution models [6,7]. Apart from widespread applications, some important issues concerning the link prediction problem itself are investigated by network researchers. Zhang et al. considered the noise in network data and analyzed the robustness of link prediction algorithms [8]. Lv et al. attempted to measure the predictability of networks through perturbation of adjacency matrix, and the structural consistency proposed in the paper made it possible to monitor the sudden changes in network's evolving mechanisms [9]. These applications and studies from different aspects reveal the significant value and sustained

* Corresponding author.
E-mail address: xiayx@zju.edu.cn (Y. Xia).

https://doi.org/10.1016/j.physa.2018.08.068
0378-4371/

development of link prediction research, which is identical to network sciences as its widespread influence in diverse fields like physics [10–12], mathematics [13–15], social science [16], biological science [17], etc.

Various link prediction methods have been proposed in the past decades. From the perspective of computer science, Markov [18] and machine learning [19,20] methods were applied for network analysis and link prediction as an aspect of data mining, while most of them require attribution and information of nodes [1]. Scientists in complex networks, instead, concentrated on the network structure and put forward algorithms with broader adaptability. The mainstream was based on the assumption that two nodes with more common neighbors (two-order paths) have higher chance to form a link [2]. To measure the structural similarity of two nodes, Common Neighbor (CN) [21] is the simplest index, while Adamic-Adar (AA) [22] and Resource Allocation (RA) [23] indexes vary by adjusting the weights of different common neighbors, and algorithms like Local Path (LP) [23] and Katz [24] take the high-order paths into account to distinguish score of links further. Apart from summing up different contributions directly, random walk [25] process, Bayes theory model [26] and information-theoretic model [27–29] were introduced tactfully to calculate similarity and achieved improvements in accuracy. Maximum-likelihood methods are another important way of prediction. The hierarchical structure model estimates the likelihood of links through dendrogram and suits networks with hierarchical structure well [30]. The stochastic block model divides nodes into groups and the probability of connection is decided by the groups nodes lie in [31]. Pan et al. applied predefined structural Hamiltonian to maximize the likelihood of observed network and score links according to the probability of adding the link to the observed network [32]. These likelihood-based methods require more computational complexity but enrich the understanding of networks from different perspectives.

Most of the similarity-based methods calculate score of likelihood by gathering weights of independent instances of one or several structural features. The mechanism of CN tends to rank the links lying in the dense part of networks higher. Even if that AA and RA punish the weights of high-degree nodes (usually consistent with nodes in the dense part) and attain better performance, the adjustments of weights could be arbitrary and vary with different networks, and finding the best is unpractical.

In this paper, we consider a more practical process in which links between nodes form as a result of an intermediary effect. Take the common neighbor feature as an example. Like most previous algorithms, each common neighbor node is thought to be relevant to the formation of target link. However, we think that the effect on the formation may be either promoting or inhibiting, which distinguishes our model from the classical. To measure the positive or negative effect quantificationally, we introduce the intermediary probability, which might be estimated by network features or node attributions. Furthermore, we adopt the assumption that different common neighbors work independently, and, according to probability theory, the probability of target link's existence is deducible. The illustration above shows a typical example of the intermediary process. Under this model, we acquire several new indexes by applying different network features. Experiments in real-world networks show that these algorithms have better accuracy of prediction, especially the precision metric. Moreover, our study on an example network combining the k-shell decomposition shows that, by using algorithms of our model, links from different parts of networks tends to get more balanced chance to reach a higher ranking, which is achieved by the normalized probabilistic score.

This article is organized as follows. In the next section, we first introduce the link prediction and several algorithms briefly, then present our intermediary probability model and apply different network structural features in the model. Four new indexes are obtained there. Their performance in experiments is presented in Section 3. In Section 4, we study a case network utilizing the k-shell decomposition. At last, we summarize and discuss our work in Section 5.

2. Model

2.1. Problem and previous methods

The link prediction problem aims at predicting missing links based on observed links. To validate algorithms for this problem, all existent links of an undirected simple network $G(V, E)$, where V is the set of nodes and E is the set of links, are randomly divided into two parts: the training set E^T and the probe set E^P. Obviously, $E^T \cup E^P = E$ and $E^T \cap E^P = \Phi$. The prediction algorithms take the training set E^T as known information and calculate a score for all unknown links $U - E^T$, where U represents all $(|V \parallel V - 1|)/2$ links of the fully connected network. The links with higher scores are considered to have higher likelihood to exist, and vice versa.

A good performance of prediction means that links of probe set E^P generally have higher scores than non-existent links. To evaluate and compare the performances of different algorithms precisely, two metrics, AUC (area under the receiver operating characteristic curve) and precision, are widely used. AUC is a metric which focuses on overall ranking result and can be interpreted as the probability that a randomly chosen missing link (link belongs to E^P) obtains a higher score than a randomly chosen non-existent link (link belongs to $U-E$). In practice, AUC is usually calculated through n times independent comparisons as follows:

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

where n' (n'') represents the number of times that a randomly chosen missing link has a higher (equal) score comparing with a randomly chosen non-existent link. The AUC of a random prediction is equal to 0.5 because the score of missing links and