Physica A 470 (2017) 1–11

Contents lists available at ScienceDirect

Physica A

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

Link prediction in complex networks based on an information allocation index

Panpan Pei, Bo Liu *, Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province, 710071, China

HIGHLIGHTS

- A generalized information-theoretic model has been proposed.
- The concept of information allocation has been proposed and introduced to link prediction.
- A new index with better overall performance has been designed.

ARTICLE INFO

Article history:
Received 27 June 2016
Received in revised form 27 October 2016
Available online 29 November 2016

Keywords:
Complex networks
Link prediction
Information theory
Resource allocation

ABSTRACT

An important issue in link prediction of complex networks is to make full use of different kinds of available information simultaneously. To tackle this issue, recently, an information-theoretic model has been proposed and a novel Neighbor Set Information Index (NSI) has been designed. Motivated by this work, we proposed a more general information-theoretic model by further distinguishing the contributions from different variables of the available features. Then, by introducing the resource allocation process into the model, we designed a new index based on neighbor sets with a virtual information allocation process: Neighbor Set Information Allocation Index(NSIA). Experimental studies on real world networks from disparate fields indicate that NSIA performs well compared with NSI as well as other typical proximity indices.

1. Introduction

Link prediction in complex networks, which aims at estimating the existence likelihood of a link based on available information [1–3], has received extensive attention from researchers in disparate scientific fields in recent years due to its theoretical and practical significance [4–6]. Link prediction can find applications in many areas. For example, in biological networks such as food webs, protein–protein interaction networks and metabolic networks, the existence of a new link has to be determined by experiments, which can be very costly if we blindly check all possible interactions. However, the costs can be sharply reduced if we focus on the most likely existing interactions provided by accurate prediction [7,8]. In online social networks, link prediction can help users to find new friends by recommending promising friendships to them [9,10]. Besides, it also plays an important role in evaluating network evolving mechanisms [11]. Although many models have been proposed to mimic the evolution of real world networks [12–15], it is usually very difficult or even impossible to fairly evaluate which one is the best. In this situation, link prediction can provide the likelihood of the currently observed network driven by each model, thus makes it possible to quantitatively compare their goodness [11].

* Corresponding author.
E-mail address: liub@xidian.edu.cn (B. Liu).

http://dx.doi.org/10.1016/j.physa.2016.11.069
0378-4371/

Over the past decades, various kinds of link prediction algorithms have been proposed. Among them, the similarity-based algorithms provide the simplest framework, which assumes that the connection likelihood of a node pair is determined by the value of their similarity, and the pairs of more similar nodes are more likely to be connected. In such algorithms, the central issue is how to define an appropriate similarity between two nodes, which can be based either on node attributes or on network structure. Since node attributes are often difficult to obtain in practice, research activities mainly focus on structure based similarity and various algorithms have been proposed. Among all the similarity indices, Common Neighbor Index(CN) [16] is undoubtedly the most simple and less time-consuming one with a comparatively high prediction accuracy [17]. And many CN-based indices [17–23] with different normalization of the number of common neighbors also obtained competitive performances in various cases. Comparatively, the Preferential Attachment Index (PA) [24] needs the least information while has a bad overall performance. These are local indices that only depend on information from nearest neighbors. Besides, there are global indices [22,25–29] and semi-local ones [17,30,31] that requires non-local structural information of the networks. Comparatively, the accuracy and computational complexity are lower in the local indices, and higher in the global ones, with the semi-local ones providing a tradeoff between them. Besides the similarity based algorithms, algorithms based on maximum-likelihood methods or probabilistic models have also been proposed. For example, the hierarchical structure model [32,33] can be employed to predict missing links in hierarchical networks, while stochastic block model [34–37] may be a good choice for networks with community features.

A fundamental issue in link prediction is the so-called link predictability problem, i.e., to which extent the links in a network can be predicted. In [38], Lü et al. investigated this problem based on a perturbation method and proposed a structural consistency index to measure link predictability in a network. Link predictability can be used to evaluate link prediction algorithms and monitor sudden changes in network evolving processes. In principal, link predictability is determined by the available information of the networks and sets the upper bounds for the performance of all link prediction algorithms. Thus, we can develop a nice link prediction algorithm only if we can make effective use of the available information. However, in practice, various available features often reflect different aspects of the networks and how to combine their contributions is a big issue. To combine the contributions from different features, the first step is to put them into a unified framework. From the information theoretic point of view [39,40], link prediction aims at quantifying the likelihood of the connection event of two nodes, which can be measured by the self-information of the event. Thus, the contributions of different features can be measured by the amount of their values of information, which is additive and thus can be directly combined. Based on this idea, Zhu et al. recently proposed a novel Mutual Information Index (MI) [41] based on the mutual information between the node pair and their common neighbors. Later, they extended it into a more general information-theoretic model to incorporate arbitrarily more available features and designed a Neighbor Set Information Index(NSI) [42] based on the information of the neighbor sets of the node pair which outperformed the MI and other typical proximity indices in experimental studies.

Compared to traditional link prediction methods, one advantage of the information-theoretic model is that it easily integrates multiple features of the networks into one framework and provides more flexibility on various kinds of networks. For example, the MI index (Fig. 1 of [41]) showed properties either similar with PA or with CN for different kinds of node pairs. Thus, it provides a promising framework for developing better link prediction algorithms. In this paper, we will further develop the information theoretic model into a more general framework. As indicated by previous studies, a major obstacle for higher prediction accuracy in local indices such as CN, is the so-called "degeneracy of states", which means that many node pairs are assigned the same score and thus results in a low discriminative resolution. The reason is that all the nodes in the common neighbor set contributed the same to the connection likelihood despite of their differences in other aspects. To tackle this issue, some CN-based indices such as AA and RA try to distinguish the contributions from different nodes in the common neighbor set. Particularly, the RA index is motivated by the resource allocation process taking place on networks and outperforms CN and AA in numerical studies. Moreover, by introducing the resource allocation process into the local path index (LP), Bai, Hu and Tang [43] proposed a RALP index with an improved performance in both weighted and unweighted networks. This suggests that further distinguish the contributions of different variables can improve the performance of link prediction algorithms, and the resource allocation process may have an important influence on the connection event of two nodes. Motivated by these observations, we derived a more general information theoretic-model that further distinguishes the contributions from different variables. Furthermore, by mimicking the resource allocation process, we introduced a virtual "information allocation process" and designed a new Neighbor Set Information Allocation index (NSIA). Experimental studies on fourteen real-world networks including two large scale networks indicate that NSIA outperforms NSI as well as some other typical similarity indices in most cases.

The article is organized as follows. In the next section, we state the link prediction problem and some typical existing indices. In Section 3, we propose a more general information theoretic model and designed a Neighbor Set Information Allocation index. Data description and experimental results are presented in Section 4. At last, we summarize our work in Section 5.

2. Problem statement and related methods

Given an undirected network $G(V, E)$, where V and E are the sets of nodes and links, respectively (Here we only consider simple graphs without multiple links and self loops). Let U denote the universal set of all possible links. Then $U - E$ is the set of nonexistent/missing links. The goal of link prediction is to find out the missing links from the set $U - E$. A general