Physica A 417 (2015) 76–85

Contents lists available at ScienceDirect

Physica A

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

Prediction of missing links based on multi-resolution community division

Jingyi Ding*, Licheng Jiao, Jianshe Wu, Yunting Hou, Yutao Qi

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xian 710071, China

HIGHLIGHTS

- First contribution, a new mechanism based on multi-resolution community division is provided.
- Second contribution, a frequency statistical model is proposed.
- Third contribution, it is very easy to understand and the time and space complexity is much lower.

ARTICLE INFO

Article history:
Received 4 November 2013
Received in revised form 20 August 2014
Available online 22 September 2014

Keywords:
Modularity density
Community detection
Multi-resolution
Link prediction

ABSTRACT

The investigation of link prediction in networks is an important issue in many disciplines. The research of prediction algorithms which required short time but high accuracy is still a challenging task. Most of the existing algorithms are based on the topological information of the networks, including the local or global similarity indices. It is found that the hierarchical organization and community structure information may indeed provide insights for link prediction. In this paper, we propose a simple link prediction method, which fully explore the community structure information of the networks. Firstly, the community structure of the networks under different resolutions is extracted. Then, a simple frequency statistical model is applied to calculate how many times that a pair of nodes divided into the same community under different resolutions. Finally, the probability of the missing links is calculated. The performance of our algorithm is demonstrated by comparing with other seven well-known methods on two kinds of networks in different scales. The results indicate that our approach not only has a good performance on the accuracy, but also has a lower time complexity than any other algorithms which are based on hierarchical structure of the network.

1. Introduction

Many complex social, biological, and information systems are described by networks, where nodes represent individuals, and links indicate the relationships or interactions between nodes. The links include friendships among people, connections in an on-line social network, physical interactions among proteins or genes, etc. The problem of predicting missing links of a network (or the future structure of the network) from the observed structure is called link prediction, which is one of the important tasks of link mining in the data mining field [1,2]. Link prediction helps us not only to understand the evolution mechanism of the complex networks theoretically [3–5], but also to solve very important issues in real-world applications.

* Corresponding author. Tel.: +86 15929737040.
E-mail addresses: jyding87@163.com, dingfeimn@163.com, dingjingyi@iiip.xidian.edu.cn (J. Ding).

http://dx.doi.org/10.1016/j.physa.2014.09.005
0378-4371/

According to the different information used for prediction, the link prediction problem falls into three kinds: (i) link prediction based on topological information; (ii) link prediction based on nodes information; (iii) link prediction based on admixture information. The first type of link prediction considers only adjacency matrices. Based on the known parts of the adjacency matrices, the unknown parts are predicted by using, for example, matrix factorization techniques [6,7]. The second type of link prediction exploits nodes information such as feature vectors of nodes or similarity values among nodes. The third type of link prediction considers not only the topology information of the network, but also the feature vectors of the nodes. In this paper, we just discuss the first type, i.e. link prediction based on the known parts of the adjacency matrices.

Most of the existing algorithms which are based on topological information of the networks only consider local or global similarity indices, such as Common Neighbors (CN) [8], Jaccard Coefficient (JC) [9], Preferential Attachment (PA) [8], Shortest Paths (SP) [10], Adar–Adamic (AA) [11] and Resource-Allocation (RA) [12]. Because the Resource-Allocation index of original is not good enough in stability, it was improved in Ref. [1]. In this paper, we just compared with this new one. These similarity indices defined as Table 1, where $\Gamma(x)$ denote the set of neighbors of x, k_z denote the degree of the node z.

However, these works do not make full use of the structural characteristics of networks, such as the hierarchical organization information and community structure information [13,14], which may indeed provide insights for link prediction. Recently, Clauset, Moore and Newman proposed a maximum likelihood method for link prediction [15]. They calculate the possibility of missing connections by sampling an ensemble of dendrograms which was given by the same network. This approach considered the hierarchical organization information of the network, so it would have better performance in the networks which have obvious hierarchical organization. However, the runtime required increase exponentially as the number of vertices increases. Later, another representative work about likelihood method on link prediction that related to block model was proposed by Guimerà and Sales-Pardo [16]. However, the whole process is very time consuming and its impossible to sum over all partitions even in a small network. In this paper, we propose a link prediction method based on the community structure. Instead of the accurate classification result obtained by a general community detection algorithm, the proposed method just needs the results obtained under different resolutions. The time complexity of the method is $O(n^2)$.

There are three main steps in our algorithm. The first step of our work is multi-resolution community division. In order to avoid the problem of resolution limit, and to get enough statistical samples, we use the method based on the optimization of modularity density to detect communities of the network. The second step is to calculate the frequency of each pair of vertices divided into the same community. We assume that the community divisions under different resolutions have different impact factors for prediction, so we give different weighted value to them. Finally, we estimate the probability of each pair of nodes according to the conversion formula between frequency and probability. If a pair of vertices has a greater probability, they are more likely to be connected by a missing link. These pairs are regarded as the most likely candidates for missing connections.

There are several advantages in our algorithm. First, it is easy to understand and implement. Moreover, the method is extremely fast. Most of the running time in our method is focus on the first step. The complexity is linear when the data is sparse. This is because the gains in modularity density are easy to compute and the number of communities decreases rapidly after just a few passes.

The contributions of this work are:

(1) A new mechanism of link prediction based on multi-resolution community division is provided. This is the first time to predict missing connections using multi-resolution community division information with high accuracy.
(2) A frequency statistical model is proposed, which can obtain the frequency of links between nodes effectively. Those with the highest frequencies are the most likely to be connected.
(3) Experimental results demonstrate that the proposed method for link prediction can be used in the networks with thousands of nodes. With much less computational time and space, the prediction qualities are competitive to those of the exact ones.

The rest of the paper is organized as follows. Section 2 describes the problem formulation. Section 3.1 introduces the approach that extracting the community structure of networks. Section 3.2 explains the frequency statistical model and defines a new similarity metric: CommPre (CP). The experiments and analyzes are provided in Section 4. Finally, Section 5 concludes the paper with some discussion and promising future work.

2. Problem formulation

Consider an undirected un-weighted network $G(V, E)$, where V is the set of vertices and E is the set of edges. Multiple links and self-connections are not allowed. For each pair of nodes, $x, y \in V$. We associate a probability p_{xy} with each pair of nodes.

To test the algorithm's accuracy, we remove a subset of connections as the probe set E^P. And then we attempt to predict missing links based on the remaining set E^T. Clearly, the observed links $E = E^T \cup E^P$ and $E^T \cap E^P = \phi$. In this paper, we adopt the AUC statistic which is equivalent to the area under the receiver-operating characteristic (ROC) curve to evaluate the quality of different link predicted methods [17]. In the present circumstances, the AUC is an important performance measure that relates the sensitivity (true positive rate) and specificity (true negative rate) of a classifier [18]. Formally, the measurement AUC can be defined as

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