Physica A 406 (2014) 244–252

Contents lists available at ScienceDirect

Physica A

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

Potential links by neighbor communities

Zheng Xie [a], Enming Dong [a], Jianping Li [a], Dexing Kong [b,*], Ning Wu [a]

[a] Department of Mathematics and System Science, National University of Defense Technology, Changsha, 410073, China
[b] Department of Mathematics, Zhejiang University, Hangzhou, 310027, China

HIGHLIGHTS

- A link prediction method is proposed based on network community and node neighbors.
- The derivation of this method is based on statistical inference.
- The method is suitable for link prediction on large-scale networks and outperforms some methods based on common neighbors.

ARTICLE INFO

Article history:
Received 12 September 2013
Received in revised form 13 March 2014
Available online 22 March 2014

Keywords:
Neighbors' community
Complex network
Link prediction

ABSTRACT

The probability of two nodes to be linked is related to their similarities in the network. Based on statistical inference, a network-structure similarity index, therefore, is proposed to find the potential links. This index quantifies the effects of the node communities on these links. And an algorithm for the index is also successfully designed. The experiments on several networks with ground-truth groups and temporal attributes reveal that two nodes are likely to be connected if some of their neighbor nodes are in common communities. The results from these experiments with tested networks, several of which cover more than a million nodes, show the reliability of the index and the advantage of its algorithm.

1. Introduction

The fast development of networks provides not only a convenient life to modern people but also some significant problems to researchers [1,2]. Consider two strangers Tom and Mary with no common friends, but some of their friends are possibly in the same social-circle. Are Tom and Mary going to be connected to each other in the future? This is a link prediction problem on social networks. This prediction aims at estimating the likelihood of the relationship between two individuals by the study of the observed links and the properties of the nodes. Based on predicting the potential links, researchers in different areas can efficiently find their cooperators or assistants in citation and coauthorship networks, and common people can find suitable friends in social networks.

In general, the prediction is based on the node similarity. However, an appropriate definition of the node similarity is still a challenge and needs further researches. One of the similarity definitions is that two nodes are considered to be similar if they have many common features, which relies on the essential attributes of the nodes. However, those attributes are usually hidden and researchers have to study the network-structure similarity. A local structural similarity can be expressed by the idea that two nodes are more likely to form a link if they have more common neighbor nodes [3,4]. Some definitions of quasi-local and global structural similarities use the information of path and community of the networks [5,6]. Based on statistical inference, some of the approaches to prediction are more accurate with higher computational complexity. Stochastic Block

* Corresponding author.
E-mail addresses: xiezheng81@nudt.edu.cn (Z. Xie), dkong@sjtu.edu.cn, dkong@zju.edu.cn (D. Kong).

http://dx.doi.org/10.1016/j.physa.2014.03.061
0378-4371/

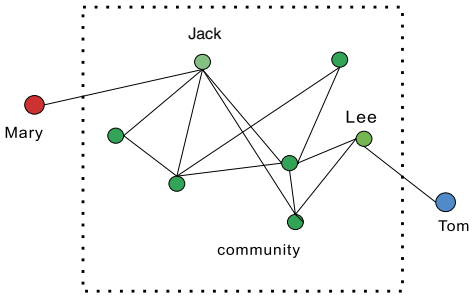

Fig. 1. Illustration of links by neighbor communities. Tom wants to make friends with Mary, who is a stranger to Mary. Lee is one of Tom's friends, and Jack is Mary's friend. And Tom knows that Jack and Lee are in the same community and he wants to acquaint himself with Mary through the community.

Model (BM) [7] and its degree corrected model [8] are two of such approaches, in which the similarity of the nodes is defined by the communities to which they belong, and the similarity of two communities is quantified by the probability that two communities are connected. Hierarchical Random Graph (HRG) [9], a global method, uses the hierarchical structure of networks to predict links. The similarity of the two nodes is described by the fraction of the potential links between the two subtrees of their nearest common ancestor that actually appears in the network.

One of the common approaches for people to make friends is that they would like to get acquainted with the people in or related to their neighbor communities. Take Fig. 1 as an example. To quantify the effect of community in the potential link between Tom and Mary, we propose a network-structure similarity index based on statistical inference. The probability of a pair of nodes to be linked is calculated by summing up the probabilities of all the possible situations in which the two nodes connect to the same community. It is assumed that each node links to a given community is conditionally independent. In other words, given a community C, knowledge of whether node i links to C occurs are assumed to provide no information to the likelihood of node j links to C occurring, and vice versa. Since a community can be regarded as a set of nodes that have similar link patterns to all the other nodes in or out of the community, it is further supposed that the likelihood of a node linking to any other node in one community is the same. And this likelihood is defined as the probability of a link between the node and the community. Based on the observed nodes, links and node communities in the network, those probabilities taken as a convex optimization problem are solved by the Lagrangian multiplier method.

The power-law distribution of the node degrees leads to the sparseness of adjacency matrix. Meanwhile little overlap of communities causes the sparseness of the node belonging coefficient. The above two properties lead to the sparseness of the variants in computing process and that of the final prediction results. The experiments on several networks with ground-truth communities and temporal attributes reveal that two nodes are likely to be connected if some of their neighbor nodes are in common communities. The results from these experiments on social networks, several of which cover more than a million nodes, show the reliability of the index and the advantage of its algorithm.

2. Method

The similarity of two nodes in an undirected unweighted network here is measured by the summation of the probabilities of all the possible situations that the two nodes link to the same community.

Suppose we have an unweighted network of n nodes connected by undirected links, such as communication or coauthorship networks, the network can be represented simply by an adjacency matrix A, in which the element $A_{ij} = 1$ if there is a link between nodes i and j, and $A_{ij} = 0$ if there is not.

Suppose that the n nodes fall into c communities, and node i represents one of the n nodes and g_i is the community to which node i belongs. Suppose that the links and the community memberships can be measured directly or by some effective methods.

This method can be illustrated more simply by defining the following two symbols. (1) If there is a node k in r-community ($k \in C_r$, in which r stands for the tag of the community; $r = 1, \ldots, c$) and node i links to k, then i is said to be adjoined to a r-th community C_r. Denoted this relation as $i \sim C_r$. (2) If two nodes i and j adjoin to a community C_r ($i \sim C_r, j \sim C_r$), the two nodes are said to be adjoined, we define this as $i \sim j$.

Under the premise of the definitions of the symbols, the index for the similarity of the nodes i and j is defined as the probability of $i \sim j$. This index is calculated by summing up the probability of $i \sim C_r$ and $j \sim C_r$ of all the possible situations, thus:

$$P(i \sim j) = \sum_r P(i \sim C_r, j \sim C_r). \tag{1}$$

By the conditional probability formula, we write Eq. (1) as follows:

$$P(i \sim j) = \sum_r P(i \sim C_r, j \sim C_r | C_r) P(C_r), \tag{2}$$

in which $P(C_r)$ is either the fraction of the nodes in the community C_r or the probability that a randomly chosen node is in C_r.