Neurocomputing 130 (2014) 36–43

Contents lists available at ScienceDirect

Neurocomputing

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

A unified community detection algorithm in complex network

Kan Li [a,*], Yin Pang [a,b]

[a] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[b] Beijing Institute of Tracking and Telecommunication Technology, Beijing 100094, China

ARTICLE INFO

Article history:
Received 19 June 2012
Received in revised form
31 October 2012
Accepted 30 November 2012
Available online 1 October 2013

Keywords:
Community detection
Vertices similarity probability
Type of network structure
Common neighbor index

ABSTRACT

In the previous methods of community detection, unipartite networks and bipartite networks are dealt with separately, so the type of network should be known in advance. This paper presents a vertices similarity probability (VSP) model to find community structure without the priori knowledge of the type of complex network structure. As vertices in the same community have similar properties, the VSP model uses vertices similarity to find community structure which is a unified algorithm and can be used in any network without knowing the type of network structure. As "Common neighbor index" has been proved to be an effective index for vertices similarity, it is used to measure the vertices similarity probability. Then, we give the method to determine the number of communities using matrix perturbation theory. We apply the model to find community structure in real-world networks and artificial networks. The experimental results show that the VSP model is applicable to both unipartite networks and bipartite networks, and is able to find the community structure successfully without using the type of network structure.

1. Introduction

Attention has been devoted to the computational analysis of complex networks [1–4], as part of the recent surge of research on large, complex networks. Although complex networks, such as social networks and biological networks, have different properties. They have a common feature, namely "community structure". Communities, also known as clusters or modules, are groups of vertices which could share common properties and/or have similar roles within the graph [5]. Finding community structure and clustering vertices in the complex network is the key to learning a complex network topology, to understanding complex network functions, to founding hidden mode, to link prediction, and to evolution detection. Many practical application results have gotten: Spirin et al. revealed the relationship between protein function and interactions inherent [6,7]; Flake et al. find the internal relations of hyperlink and the main page [8,9]; Freeman et al. identify the social organizations to evolve over time [10,11] and so on.

Generally, there are two types of network structure, unipartite networks and bipartite networks. Unipartite networks assume that connections between the vertices in the same community are dense, and between the communities are sparse, such as Social network [12], biochemical network [13] and information network [14]. Most of the community detection methods are made for unipartite networks, and try to find communities with sparse edges among them. However, some real networks are bipartite. Vertices of a bipartite network can be partitioned into two disjoint sets such that no two vertices within the same set are adjacent [21]. In another word, edges in a bipartite network joining only vertices of different communities, such as shopping networks [15], people attending events network [16], plant-animal mutualistic networks [17], scientific publication networks [18], etc. Methods for bipartite networks try to find communities with sparse edges inside them. Among the methods for community detection, the type of the network must be known, because the properties of networks are different in different types of the network structure. The usual approach taken to detect communities in bipartite networks is to construct a unipartite projection network of one part of the network, and then find communities in that projection using methods for unipartite networks. In 2007, Barer proposed a method that can detect communities in bipartite networks directly without the projection [21]. Generally speaking, people should first know the type of the network, unipartite or bipartite, then chose the corresponding method to find communities.

In the famous modularity matrix method [19,20] proposed by Newman et al. communities are found by the spectral of modularity matrix. If the type of the network structure is known, modularity optimization is able to find community structure in both unipartite and bipartite networks by the maximum or minimum eigenvalue and the eigenvetor separately. Barer propose a BRIM [21] method which does community detection in bipartite

* Corresponding author.
E-mail addresses: likan@bit.edu.cn (K. Li), pangyin@bit.edu.cn (Y. Pang).

0925-2312/$ - see front matter
http://dx.doi.org/10.1016/j.neucom.2012.11.054

networks and can determine the number of communities at the same time. Furthermore, Barber and Clark use the label-propagation algorithm (LPA) for community detection [22] too. However, [21,22] cannot be used without knowing the type of network.

In hierarchy networks, hierarchical clustering is adopted frequently for community detection. Vertices are grouped into communities that further subdivide into smaller communities, and so forth, as in [19]. Clauset, Moore and Newman propose HRG [23] using the maximum likelihood estimation to forecast the probability of connections between vertices when the organization of vertices are known. In unipartite networks, most edges are in the same level, and in bipartite networks, edges are between levels. Hierarchical methods perform remarkably in clear hierarchy network, but not so impressive under contrary circumstance. Moreover, a hierarchical method always has high computational complexity.

Different from traditional concept which divide network by principle of "inside connection dense outside sparse", stochastic block models assume that the probability that two vertices are connected depends on the blocks to which they belong [24], such as the method proposed in 2009 by Roger Guimera and Marta Sales-Pardo based on HRG. However, the assumption that vertices in same blocks have same connection probability is not accurate. Recently, Karrer and Newman [25] also proposed a stochastic block model which considers the variation in vertex degree. This stochastic block model solves the heterogeneous vertex degrees problem and got a better result than other previous researches without degree correction. It can be used in both types of networks, but different types of networks should be dealt with separately none the less.

In some cases, researchers have no priori knowledge of the network structure. For example, when we know the interaction of vertex in the protein network, we may have no knowledge of the type of the network. And moreover, when we get a network which consists of people's relationships in schools, the type of network may not be sure. This is because that if links are between students only, the network will be a unipartite network; or if links are between students and teachers, the network will be a bipartite one. An effective method which can be used for finding community structure in both unipartite and bipartite networks requires the knowledge of the network topology only is needed. Although the properties of "edges" in the two types of networks are different, vertices in the same communities should be similar because vertices in same communities have similar properties. In this paper, we develop a unified VSP model which is based on the vertices similarity. Therefore, the VSP model can be used in any type of networks as long as we put similar vertices in same communities. The VSP model gets ideal result both by theoretical proof and experimental analysis.

The paper is organized as follows. In Section 2, we prove that vertices similarity theory is suitable for finding community structure. In Section 3, we present the VSP model and the method to group network into two communities. In Section 4, we give the method to determine the number of communities in the network. Finally, we make the experiment in both unipartite and bipartite network. Compared with Newman's modularity and other methods, the VSP model is an accurate unified model which can find community structure without prior knowledge of type of the network structure. In Section 5, we give our conclusions.

2. Vertex similarity in finding community structure

The concept of community informs that vertices in the same community should share common properties no matter in unipartite or bipartite network. It means that vertices in the same community should be similar, although edges in different type of the network structures are connected in different ways. Therefore, we change our focus from "edges" to "vertices" for finding communities.

Vertex similarity is widely studied by researchers in complex network. It is sometimes called structural similarity, to distinguish it from social similarity, textual similarity, or other similarity types. It is a basic premise of research on networks that the structure of a network reflects real information about the vertices the network connects, so it is reasonable that meaningful structural similarity measures might exist [26]. In general, if two vertices have a number of common neighbors, we believe that these two vertices are similar. In community detection, we assume that two similar vertices have similar properties and should be grouped in the same community.

Let Γ_x be the neighborhood of vertex x in a network, i.e., the set of vertices that are directly connected to x via an edge. Then $|\Gamma_x \cap \Gamma_y|$ is the number of common neighbors of x and y. Common neighbor index, Salton index, Jaccard index, Sorenson index, LHN (Leicht–Holme–Newman) index, and Adamic–Adar index [27–31] are five famous methods for vertex similarity. Many researchers have analyzed and compared these methods. Liben-Nowell [32] and Zhou Tao [33] proved that the simplest measurement "common neighbor index" performs surprisingly well. We use "common neighbor index" to measure the vertex similarity in our VSP model.

Definition 1. For two vertices x and y, if there is a vertex z to be the neighbor of x and y at the same time, we call x and y a pair, denoted as $pair(x,y)$. z is called the common neighbor of $pair(x,y)$.

Since vertices which are in the same community have similar properties, we assume that vertices in the same community are similar vertices. The more similar the vertices inside a community are the more common neighbors they have. The number of common neighbors N_{ij} of vertices i and j is given by

$$N_{ij} = |\Gamma_i \cap \Gamma_j|, \quad \text{and} \quad N_{ii} = 0.$$

The sum of common neighbors with vertices in same communities N_{in} is given by

$$N_{in} = \sum_{\substack{i,j \in \text{ same} \\ \text{community}}} N_{ij}.$$

And the sum of common neighbors with vertices in different communities N_{out} is given by

$$N_{out} = \sum_{\substack{i,j \notin \text{ same} \\ \text{community}}} N_{ij}.$$

Therefore, the task of maximizing the number of common neighbors in the same community is to get $\max(N_{in})$ or to get $\min(N_{out})$. The sum of common neighbors in the network R is given by

$$R = \frac{1}{2} \sum_{i,j \in n} N_{ij}.$$

We define the adjacency matrix A to be the symmetric matrix with elements A_{ij}. If there is an edge joining vertices i and j, $A_{ij} = 1$; if no, $A_{ij} = 0$. Define a_i as ith vector of A, so as A can be rewritten as $A = [a_1, a_2, \ldots, a_n]$. If and only if $A_{ik}A_{kj} = 1$, the vertex k is a common neighbor of vertices i and j. Therefore N_{ij} can be rewritten as

$$N_{ij} = \sum_k A_{ik}A_{kj} = a_i \cdot a_j,$$

when i and j are two different vertices. As $a_i \cdot a_j = k_i$, matrix N is

$$N = A^T A - \Lambda_k,$$