Physica A 447 (2016) 508–519

Contents lists available at ScienceDirect

Physica A

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

A general method of community detection by identifying community centers with affinity propagation

Wei-Feng Guo, Shao-Wu Zhang *

Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China

HIGHLIGHTS

- A general method suitable for unweighted, weighted, undirected, directed and signed network.
- Construct the dissimilarity distance matrix with different measures.
- Extract a candidate center set of community with AP algorithm.
- Determine the community by selecting the center subset to maximum the modularity.

ARTICLE INFO

Article history:
Received 28 July 2015
Received in revised form 29 October 2015
Available online 29 December 2015

Keywords:
Dissimilarity distance matrix
Centers
Community
Modularity
Affinity propagation

ABSTRACT

Detection of community structures is beneficial to analyzing the structures and properties of networks. It is of theoretical interest and practical significance in modern science. So far, a large number of algorithms have been proposed to detect community structures in complex networks, but most of them are suitable for a specific network structure. In this paper, a novel method (called CDMIC) is proposed to detect the communities in un-weighted, weighted, un-directed, directed and signed networks by constructing a dissimilarity distance matrix of network and identifying community centers with maximizing modularity. For a given network, we first estimate the distance between all pairs of nodes for constructing the dissimilarity distance matrix of the network. Then, this distance matrix is input to the affinity propagation (AP) algorithm to extract a candidate center set of community. Thirdly, we rank these centers in descending order according to the sum of their availability and responsibility. Finally, we determine the community structure by selecting the center subset from the candidate center set in an incremental manner to make the modularity maximization. On three real-world networks and some synthetic networks, experimental results show that our CDMIC method has higher performance in terms of classification accuracy and normalized mutual information (NMI), and ability to tolerate the resolution limitation.

1. Introduction

Many complex systems in nature and society can be commonly modeled as complex networks or graphs with nodes representing individuals or organizations and edges representing the interactions among these nodes [1]. It has been shown that many real world networks (e.g. social networks, biological networks and the Internet) have a structure of modules or

* Corresponding author.
E-mail address: zhangsw@nwpu.edu.cn (S.-W. Zhang).

http://dx.doi.org/10.1016/j.physa.2015.12.037

communities which are characterized by the sub-graphs of densely connected nodes and sparsely connected with other parts of the networks. Detecting the community structures is beneficial to analyzing the structures and properties of networks [2].

So far, many approaches have been introduced to detect the communities of complex networks, which can be roughly categorized into positive network community detection and signed network community detection according to the characteristics of network structure. Positive network communities are defined as the groups of nodes in which the links in intra-group are dense but the links in inter-groups are sparse [2]. Signed network communities are defined as the groups of nodes in which not only the positive links in intra-group but also the negative links inter-groups are dense [3].

For the methods of detecting the positive network communities, there are two classical algorithms, i.e., spectral bisection [4] and Kernighan–Lin [5]. Spectral bisection algorithm uses the basis of eigenvectors of Laplacian matrix of a graph to find the optimal cuts of the networks, which has been shown to be a NP-complete problem [6]. Kernighan–Lin algorithm divides the networks according to the optimization of the number of intra- and inter-communities edges using a greedy algorithm, which is sensitive to the initial partition. The change in the order of an initial partition may significantly alter the detecting results. To address the issues of the two classical methods [7–11], several algorithms based on optimizing the modularity (represented by Q function) are developed to detect the community structures of complex networks, especially for weighted networks and directed networks [12–15]. However, the existing methods based on modularity maximization suffer from two major limitations. One is that the maximization of Q is an NP-hard problem, though a few of the optimization techniques such as simulated annealing [16] and extreme optimization [10] are introduced to obtain the suboptimal solutions. Another is resolution limit so-called that it cannot detect the communities whose node number is smaller than a predefined threshold [17]. Thus, several algorithms have been proposed to alleviate the resolution limit by redefining the modularity function or adding weights to the edges [18,19].

The methods of detecting positive network community cannot be suitable for detecting the community structure of signed networks with positive links and negative links. Thus, various approaches have been proposed to detect the communities of signed networks by designing the improved modularity function, adopting an agent-based heuristic strategy and other strategies [3,15,20]. Although above-mentioned algorithms are suitable for the undirected networks, they cannot be extended to directed networks.

In this paper, we will introduce a novel method (called CDMIC) to detect the community structures of weighted, un-weighted, directed, un-directed and signed networks. The key strategy of our CDMIC method is to identify the community centers such that the modularity for the network partition is maximization. By using different similarity measures, CDMIC estimates the similarity between pair of nodes in the network, transforming it into dissimilarity by a decreasing function to obtain a distance matrix of the network. Then we use affinity propagation (AP) algorithm to extract the candidate center set of community, ranking these centers in descending order according to the sum of their availability and responsibility. By selecting the center subset from the candidate center set in an incremental manner to partition the network, and calculating their corresponding modularity, we choose the partition of modularity maximization as the final result of community detection. On three real-world networks and some synthetic networks, our CDMIC method shows higher performance in terms of classification accuracy and normalized mutual information (NMI), and also has strong robustness and ability to tolerate the resolution limitation.

This paper is organized as follows. In Section 2, we explain some crucial concepts, and briefly introduce the affinity propagation algorithm. In Section 3, we describe our novel CDICM method for detecting the community structures in detail. Experimental results of some synthetic networks and three real-world networks are shown in Section 4. In Section 5, we discuss the effects of similarity measures, different modularity measures, and dissimilarity distance matrix. Finally, our conclusions are presented in Section 6.

2. Basic concepts

2.1. Similarity measure

Similarity measures play an important role in the research of complex networks. Employing appropriate similarity measure to grasp as much network structure information as possible can help to accurately detect the community structures of a network. In Table 1, we list the ten similarity measures [21], which are often used to measure the similarity between any two nodes or two links in different networks. The measures of Common Neighbors (CN), Salton, Jaccard, Sørenson, Hub Promoted (HP), Hub Depressed (HD), Leicht–Holm–Newman (LHN), Preferential Attachment (PA) and Adamic–Adar (AA) are based on the local structural information (i.e. neighborhood information). In addition, the first seven measures, from CN to LHN, only differ in the denominator. If the investigated network simultaneously has large clustering coefficient and large degree heterogeneity, there are significant differences among those seven measures [21]. PA is a proximity measure and often used to quantify the functional significance of edges subject to various network-based dynamics, which does not require information on the neighborhood of each node [21]. AA refines the simple counting of common neighbors by assigning the less connected neighbors more weight [21,22]. Assuming that each transmitter has a unit of resource, and equally distribute it between all its neighbors, then resource allocation (RA) index can be defined as the amount of resource v_j received from v_i, which works well on the networks with large clustering coefficient, high degree heterogeneity and absence of a strongly assortative linking pattern [21].