Physica A 442 (2016) 182–192

Contents lists available at ScienceDirect

Physica A

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

Identifying overlapping communities in networks using evolutionary method

Weihua Zhan [a,b,c,*], Jihong Guan [d], Huahui Chen [b], Jun Niu [b], Guang Jin [b]

[a] Department of Control Science and Engineering, Zhejiang University, Hangzhou 310058, PR China
[b] Department of Computer Science, Ningbo University, Ningbo 315211, PR China
[c] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, PR China
[d] College of Electrical Information and Engineering, Tongji University, Shanghai 201804, PR China

HIGHLIGHTS

- Present an encoding scheme for an overlapping partition of a network.
- Present two informativeness measures for a node.
- Present an coevolutionary schema between two segments over the population.

ARTICLE INFO

Article history:
Received 18 January 2015
Received in revised form 20 May 2015
Available online 16 September 2015

Keywords:
Community structure
Evolutionary method
Overlapping communities

ABSTRACT

Community structure is a typical property of real-world networks, and has been recognized as a key to understand the dynamics of the networked systems. In most of the networks overwhelming nodes apparently live in a community while there often exists a few nodes straddling several communities. Hence, an ideal algorithm for community detection is that which can identify the overlapping communities in these networks. We present an evolutionary method for detecting overlapping community structure in the network. To represent an overlapping division of a network, we develop an encoding scheme composed of two segments, the first one represents a disjoint partition and the second one represents an extension of the partition that allows of multiple memberships. We give two measures for the informativeness of a node, and present a coevolutionary scheme between two segments over the population for solving the overlapping partition of the network. Experimental results show this method can give a better solution to a network. It is also revealed that a best overlapping partition of the network might not be rooted from a best disjoint partition.

1. Introduction

As a unified tool for studying various complex systems, networks have attracted tremendous attentions during the last ten years [1–3], with nodes representing the units and edges denoting diverse interactions between these units. In social networks edges often capture various social relations between individuals; in technology networks (such as Internet) an edge may correspond to a physical connection (or communication linkage) between two sites; in information networks an edge usually indicates the flow of information between sites.

* Corresponding author at: Department of Computer Science, Ningbo University, Ningbo 315211, PR China.
E-mail address: zhanweihua@nbu.edu.cn (W. Zhan).

http://dx.doi.org/10.1016/j.physa.2015.09.031

Community structure is an important topological feature of real networks, which refers to the natural clusters of nodes such that the connections within clusters are significantly more dense than those between clusters [4]. Due to the intimate affiliation with function decomposition and various dynamics of systems, community structure detection has been extensively studied. A variety of methods for detecting communities have been proposed based on different principle and heuristics, such as divisive method based on betweenness [4], methods based on modularity optimization by simulated annealing [5], spectral method [6,7] or extremal optimization [8], methods based on dynamical process including random walks [9] or synchronization [10,11], methods based on different formal definitions of community [12,13], and methods based on minimum spanning tree [14].

For all of these methods a common assumption is that community structure is a disjoint division of the network, that is, any node should only belong to a community. However, it may be not the case for many real-world networks. In scientific collaboration network, for instance, an energetic scientist would have participated several research groups with different concerns. Hence an ideal algorithm for community detection should be able to automatically find an accurate overlapping division of the network if the community structure is indeed overlapping.

Recently there is growing interest in overlapping community detection [15–19]. A well-known method is the clique percolation method (CPM) [15] where a community is a union of some adjacent k-cliques (complete subgraph with k nodes) in the network. This method has been extensively applied to the analysis of social networks and biology network. How to select a best k is a practical problem to CPM since the divisions found with different k values generally differ from each other. It is notable that several methods based on extend a disjoint division of network to an overlapping division have been proposed [20,21].

Similar to detecting disjoint community, the detection of overlapping community also can be formulated as an optimization problem given an appropriate measure for the quality of an overlapping division. Nicosia et al. [16] extended the modularity to overlapping case, and then proposed a genetic algorithm to optimize their quality function. Shen et al. [22] also presented an overlapping measure and then employed the Blondel's algorithm to optimize it. Zhang et al. successively presented a fuzzy c-means method [23] and negative matrix factorization method [24] for finding a good overlapping division.

As opposed to other heuristics, evolutionary methods have a stronger ability of global search and stronger stability originated from the search mechanism based on population. Zhan et al. [25] proposed a modified adaptive genetic algorithm (MAGA), which is superior to standard genetic algorithms when applied to community detection. To deal with overlapping case, here we propose an evolutionary method for overlapping community detection, MAGA*, as an extension of the MAGA. In Section 2, we give a variation of overlapping modularity for evaluating the quality of an overlapping partition. In Section 3, we describe the evolutionary method for detecting overlapping communities in detail. In Section 4, we test the method on several real-world networks: karate network, high school network, and dolphin networks. At last, the conclusion is given.

2. Quality function for an overlapping partition

To measure the goodness of a partition of a network the modularity [26] was proposed by Newman and Girvan, which has been widely used as an objective function for community detection approaches based on optimization. There also exist other measures for the quality of a disjoint partition, such as the hamiltonian of potts model [27] and absolute potts model [28], modular density [29], surprise value [30,31].

The definition of modularity is based on the idea that the true community structure of the network should correspond to a statistically surprising arrangement of edges in a network, that is, the number of actual links within communities should be significantly beyond that of expected links of a null model. Configuration model, an extensively used null model, is employed in the definition of modularity. Let L be the total number of edges in the network, k_i be the degree of the node i, then in the null model the expectation of edges present between nodes i and j is $\frac{k_i k_j}{2L}$. The modularity thus can be written as follows:

$$Q = \frac{1}{2L} \sum_{c \in \mathcal{C}} \left(e_c^{in} - e_c^{exp} \right) \tag{1}$$

where $\mathcal{C}$ refers to a partition on the network, e_c^{in} and e_c^{exp} are the number of inner links in the cluster c and that of the expectation of inner links, which are counted as $\sum_{i,j \in G_c} A_{ij}$ and $\sum_{i,j \in G_c} \frac{k_i k_j}{2L}$, respectively. Q is the sum of the difference over $|\mathcal{C}|$ groups of the specific partition. The maximum value of Q is 1, and a value approaching 1 indicates strong community structure. Conversely, a value approaching 0 implies weaker community structure or indivisibility. For a network with strong community structure, it normally falls in the range from around 0.3 to 0.7.

Since the above definition of modularity is actually designed for simple networks, some variations have been presented for various types of network [32–35]. To evaluate the quality of an overlapping partition of the network, it requires redefining the number of inner links and the expectation of links in a cluster. The number of inner links in the cluster of c can be counted as

$$e_c^{in} = \sum_{i,j \in G_c} S_{ic} S_{jc} A_{ij}, \tag{2}$$