Physica A 392 (2013) 5218–5230

Contents lists available at ScienceDirect

Physica A

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

Overlapping community detection using a generative model for networks

Zhenwen Wang*, Yanli Hu, Weidong Xiao, Bin Ge

College of Information System and Management, National University of Defense Technology, Changsha, China

HIGHLIGHTS

- We propose a generative model that describes the probability of generating a network.
- The communities of nodes are detected by fitting this model to the network.
- The degree of participation in each community is calculated.

ARTICLE INFO

Article history:
Received 6 January 2013
Received in revised form 19 May 2013
Available online 19 June 2013

Keywords:
Networks
Community detection
Probabilistic model and statistics
Generative model

ABSTRACT

Detecting overlapping communities is a challenging task in analyzing networks, where nodes may belong to more than one community. Many present methods optimize quality functions to extract the communities from a network. In this paper, we present a probabilistic method for detecting overlapping communities using a generative model. The model describes the probability of generating a network with the model parameters, which reflect the communities in the network. The community memberships of each node are determined based on a probabilistic approach using those model parameters, whose values can be obtained by fitting the model to the network. This method has the advantage that the node participation degrees in each community are also computed. The proposed method is compared with some other community detection methods on both synthetic networks and real-world networks. The experiments show that this method is efficient at detecting overlapping communities and can provide better performance on the networks where a majority of nodes belong to more than one community.

1. Introduction

Many complex systems consist of interconnected entities in the real world [1]. The network provides a formal method to represent such systems, where nodes represent the entities and edges represent the connections between the entities. In the past decade, networks have been extensively studied. Community structure is one of the interesting properties which are revealed in the study of networks. Communities are groups of nodes with relatively denser edges within groups than those between them [2]. Detecting communities from a network can provide valuable information about the network. For example, in collaboration networks of scientists, detecting communities can uncover research groups of similar fields [2].

Many methods have been developed to divide nodes into disjoint communities where a node belongs to only one community, including the edge betweenness method [3], spectral partition method [4], hierarchical clustering [5] and extremal optimization [6]. In the real world, it is common that a node belongs to more than one community. Researchers have recently focused on detecting overlapping communities where nodes may belong to more than one community. A significant

* Corresponding author. Tel.: +86 15973136545.
E-mail address: wang_zhen_wen@163.com (Z. Wang).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.06.038

approach is the clique percolation method [7], which believes that communities are the unions of adjacent complete subgraphs with k nodes. From then on, many methods have been proposed to detect overlapping communities. Most methods have heuristically optimized quality functions to discover communities. Such functions include the modularity function [8,9], the local fitness function [10], the function based on edge density [11], the function based on node similarity [12]. These optimization-based methods provide reasonable results in practice. However, they cannot return a unique optimum [13] and the modularity cannot be utilized to find very small communities [14]. The research in Ref. [15] points out that such deficiencies can be mended by fitting generative models to networks.

Generative models describe the probability of generating a network based on the model parameters, which describe how nodes are connected to each other [2]. The communities in a network are uncovered by fitting the model to the actual network.

Generative models have been considered in many methods for detecting communities. In Ref. [16], an edge is generated with the probability p_{in} if its two nodes belong to the same community or p_{out} if they belong to different communities. The method in Ref. [17] gives the probability of generating a network based on the community memberships of the nodes. These two methods both take the number of communities as an external input. To overcome this problem, Ref. [18] extends the method in Ref. [17] by defining the average entropy of the classification to infer the number of communities. Ref. [19] optimizes the posterior probability of the number of communities to infer the most probable number of communities.

The above methods based on generative models assume that a node belongs to only one community. To extract overlapping communities from networks, Brian Ball et al. in Ref. [15] develops a method using a generative model where nodes may belong to more than one community. However, the model in Ref. [15] assumes that networks may have more than one edge between a pair of nodes. Such an assumption makes the model somewhat unrealistic, since most real-world networks have only one edge between two nodes.

In this paper, a new generative model is built to detect overlapping communities based on the idea that communities are sets of edges. This idea corresponds to the nature of overlapping communities. When communities are sets of edges, each edge belong to some community. A node will naturally belong to more than one community if it is connected to the edges which belong to different communities. In order to detect overlapping community, the idea is utilized to build a generative model in this paper. When the generative model is built, overlapping communities can be discovered using the model parameters which best fit a network. Experiments on synthetic benchmarks and real-world networks show the effectiveness of the proposed method.

The outline of this paper is as follows: a generative model for networks is described in the next section. In Section 2, we describe a method which discovers overlapping communities by fitting the proposed generative model to a network. In Section 3, experiments are performed. Finally, the conclusions are given.

2. A method for detecting overlapping communities

A method for detecting overlapping communities is described in this section. The first step is to build a generative model for networks. Next, the model parameters are calculated by fitting this model to the network. Finally, it is demonstrated how overlapping communities are detected using these parameters.

2.1. A generative model for networks

Suppose there is a network G with N nodes and M directed edges, in which v_i is the i'th node. In order to extract communities from the network G, a generative model is presented to describe the probability of generating the network G.

The network G can be represented by its adjacency matrix $\mathbf{A}$, where the element $A_{ij} = 1$ denotes that there is an edge from v_i to v_j otherwise $A_{ij} = 0$. The probability of generating the network G is the probability of generating the adjacency matrix $\mathbf{A}$, which can be written as the product of the probabilities of the elements that are equivalent to 1 [17]. Let e_i denote the i'th edge in the network G. Consequently, the probability of generating the network G can be written as

$$P(G) = P(\mathbf{A}) = \prod_{i=1}^{N}\prod_{j=1}^{N} P(A_{ij})^{A_{ij}} = \prod_{i=1}^{M} P(e_i), \tag{1}$$

where $P(e_i)$ is the probability of generating the edge e_i.

Obviously, Eq. (1) cannot provide the exact probability of the network G, since it excludes the elements of the adjacency matrix that are 0. The exact formula of $P(G)$ should include all elements of the adjacency matrix. The purpose of this paper is to develop a method for detecting communities by computing the parameters in $P(G)$. However, it is hard to compute the parameters in $P(G)$ when using the exact formula. In order to make the computation easier, the formula of $P(G)$ is simplified in Eq. (1), where the elements of the adjacency matrix that are 0 are excluded.

The next step of building this model is to give the probability of generating the edge e_i. Assume that there are K communities in the network G. The value of K will be calculated later. According to the idea that communities are sets of edges, each edge belongs to some community, which is the community of the edge. The variants $\{z_i\}_{i=1}^{M}$ are utilized to indicate the communities of all M edges, where the value of z_i indicates which community is the community of the edge e_i.