Physica A 419 (2015) 48–54

Contents lists available at ScienceDirect

Physica A

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

Inferring overlapping community structure with degree-corrected block model

Yingfei Qu *, Weiren Shi, Xin Shi

Institution of Intelligence Science & Advanced Integration Technologies, Chongqing University, Chongqing 400030, China

HIGHLIGHTS

- Inferring community structure based on the degree-corrected block model.
- Our algorithm can detect overlapping communities.
- Our algorithm has low time complexity.
- Experiments on synthetic and real-world networks certify the validity of our algorithm.

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form 2 September 2014
Available online 18 October 2014

Keywords:
Community detection
Stochastic block model
Spectral partitioning
Maximum likelihood method

ABSTRACT

Recent research has shown great interest in statistical inference methods for community detection, not only in models and algorithms but also in the detectability. In this paper we propose a fast community detection algorithm based on the degree-corrected block model. By introducing a parameter to select the candidate solutions, our algorithm is able to detect overlapping communities. Experiments on a range of networks have achieved state-of-the-art results. Moreover, we show that the algorithm based on the degree-corrected block model also suffers the detectability limitation, which is in accord with the most recent research on the detectability threshold.

1. Introduction

In real-world networks, vertices act neither in isolation nor simply in whole. They are usually organized in some local structure called communities in network research. Communities are believed to play the role of functional units within a networked system [1]. For instance, in the biological domain community detection algorithms have been used to find the interaction of metabolic networks [2]. As a consequence, community detection has become a fundamental research orientation of network science. For a better understanding of structural and dynamical properties of networks, recent years have witnessed an explosion of community detection methods, such as graph cut methods based on the principle of maximum flow and minimum cut [3], hierarchical clustering methods [4], spectral clustering methods [5,6], modularity maximization methods and its transformations [7–10], random walk methods [11], label propagation methods [12], and so on. Although community detection methods can be diverse, we have discovered several disadvantages that influence many algorithms. On the one hand, quite a lot of community detection algorithms have inferior theoretical basis. Therefore, interest has been attracted to the statistical inference methods for their solid mathematical foundations [13–15], but most inference algorithms cannot detect overlapping communities. On the other hand, the detectability threshold, below which community detection will be unfeasible, has drawn wide attention and is believed to affect all community detection algorithms [15–20].

* Corresponding author.
E-mail address: quyf@cqu.edu.cn (Y. Qu).

http://dx.doi.org/10.1016/j.physa.2014.10.025
0378-4371/

However, current research is not general enough to support this belief as most investigations are based on the special case of the standard stochastic block model [16,18]. The effect of the detectability threshold on other complex models such as the degree-corrected block model is not clear and needs experiments for research.

In this paper we focus our attention on community inference methods based on the degree-corrected block model. By analyzing the spectral properties and computing the likelihood, our algorithm can find many possible partitions of the network. Then we introduce a parameter to select the candidate solutions from the partitions. Finally, by comparing the candidate solutions our algorithm is able to infer overlapping community structure. With experiments on the Lancichinetti–Fortunato–Radicchi (LFR) benchmark networks [21] and some real-world networks, we show that our algorithm can detect overlapping communities efficiently considering both speed and quality. At the same time, we find that the detection will be unfeasible if the average mixing parameter of the LFR benchmark exceeds a threshold, even though there exists community structure in theory. The results are in line with what has gotten so far for the detectability threshold research.

2. Degree-corrected block model

At first, we must define the particular network or networks for study. The standard stochastic block model is most widely used and has been deep researched for community detection [18,22]. But in this paper we prefer to consider the degree-corrected block model, which is developed from the standard one. Compared with the standard one, the degree-corrected block model takes the broad degree distribution of the networks into consideration and gives better performances in most real-world networks [22]. The model builds a graph of n vertices with different probabilities for edges within and between communities. Assuming that a network is generated according to this model, we can partition the network into some number of communities by computing the edge probabilities or the model parameters based on the observed features of the targeted network. Although most real-world networks may not be generated by the model, this method provides an amazingly good estimate of the true community structure.

We will here concentrate on the case of two communities in a network based on the degree-corrected block model. Usually, the edges in the model are created independently at random with probability p_U for vertices in the same community and p_V for vertices between different communities. Here we employ another mode to create the edges, which is based on the Poisson distribution with mean λ_U for edges in the same community and mean λ_V for edges between different communities. Actually, the random mode and the Poisson mode are extremely similar because the edges in real-world networks are very sparse, which has been well studied in the research of small-world [23] and scale-free [24] networks. We prefer the Poisson mode because its analysis is more concise.

Considering a network G, we denote by A the adjacency matrix of the network, and its elements $A_{ij} = 1$ if vertices i and j in G are connected and $A_{ij} = 0$ otherwise. Self-edges are not discussed in this paper, so $A_{ii} = 0$. c_i denotes the community to which vertex i belongs. Given the community memberships c and the Poisson parameters λ, we can denote the likelihood of generating a particular network G by

$$P(G|c, \lambda) = \prod_{i<j} \frac{\left(d_i d_j \lambda_{ij}\right)^{A_{ij}}}{A_{ij}!} \mathrm{e}^{-d_i d_j \lambda_{ij}}, \tag{1}$$

where $d_i d_j \lambda_{ij}$ is the expected number of edges between vertices i and j, and d_i is the degree of vertex i. $d_i d_j \lambda_{ij}$ can be either λ_U or λ_V, depending on whether the edges are in the same community or in the different communities. Then we can rewrite Eq. (1) as

$$P(G|c, \lambda) = \frac{{\lambda_U}^{b_U}}{b_U!} \mathrm{e}^{-\lambda_U} \cdot \frac{{\lambda_V}^{b_V}}{b_V!} \mathrm{e}^{-\lambda_V}, \tag{2}$$

where b_U and b_V are the observed numbers of edges within and between communities respectively for a given partition of the network. In fact, maximizing the likelihood and maximizing its logarithm are equivalent, but the latter is easier to work with. We neglect an unimportant additive constant and obtain the log of likelihood of the model.

$$Q = b_U \ln \lambda_U + b_V \ln \lambda_V. \tag{3}$$

We can estimate the most likely values of λ_U and λ_V with b_U and b_V by

$$\lambda_U = \frac{2b_U}{d_1^2 + d_2^2}, \qquad \lambda_V = \frac{b_V}{d_1 d_2}, \tag{4}$$

where d_1 and d_2 are the sums of the degrees of the vertices in the two communities. Substituting Eq. (4) back into Eq. (3) gives the object function which should be maximized to find the community structure.

$$Q = b_U \ln \frac{2b_U}{d_1^2 + d_2^2} + \ln \frac{b_V}{d_1 d_2}. \tag{5}$$

As the possible partitions are too many in a network, the maximization of Eq. (5) faces high computational cost and low speed. So we just come back to Eq. (1) and get its logarithm directly, assuming that λ_U and λ_V have been known.

$$R = \ln P(G|c, \lambda) = \sum_{i<j} \left[A_{ij} \ln d_i d_j \lambda_{ij} - d_i d_j \lambda_{ij} - \ln A_{ij}!\right]. \tag{6}$$