Perspectives in Science (2016) 7, 122—125

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc


# Discussion of the community detection algorithm based on statistical inference☆

**Liangxun Shuo***, **Bianfang Chai**

*Shijiazhuang University of Economics, Shijiazhuang 050031, China*

Received 27 October 2015; accepted 11 November 2015
Available online 10 December 2015

**KEYWORDS**
Statistical inference;
Community detection;
Probabilistic model

**Summary** This paper aims to solve the model and parameters with the discussion from the algorithm characteristics of model, the analysis of each algorithm, solving the difficulties, problems and development direction. The paper tries to analyze and summarize the evolution law of algorithm and solution thinking. Some improvements are given on the existing community detection algorithm based on statistical inference.


## Introduction

The goal of the community detection is to resolve the modular community structure in complex network with the information contained in the graph topology structure. This is the key of network analysis. At present, it has been widely used in the fields of sociology, biology, physics and computer science. It is very important for people to understand the characteristics of complex system. A good community detection algorithm can find a variety of network structure and deal with all kinds of network (including a directed, undirected or weighted network). It's time complexity and space complexity can be controlled in large network. It must has a reliable theoretical basis and not be only empirically based heuristic method.

A community detection method based on statistical inference can identify the structure of the network with structural equivalence and regular equivalence, and fit the observed network with the generated model to obtain the node's division and the structure of the network. The community detection method based on statistical inference has a complete probability theory and interpretation, and can better meet the standard of community detection algorithm. This paper reviews the research status of community detection model based on statistical inference and the main problems. The principle and application of each model are analyzed in detail. Finally, this paper discusses the future development prospects and problems of community detection method based on statistical inference.

## Community detection model based on statistical inference

According to the different community elements in the generative model, it can be divided into vertex community and linked community (Ahn et al., 2010; Evans and Lambiotte,

☆ This article is part of a special issue entitled ''Proceedings of the 1st Czech-China Scientific Conference 2015''.

* Corresponding author.
*E-mail address:* Shuolx@sjzue.edu.cn (L. Shuo).

http://dx.doi.org/10.1016/j.pisc.2015.11.020

2009). Nodes are assigned to each community in vertex community, and the link is assigned to each community in linked community. Because the edges of the vertices can be assigned to different communities, the idea of linked community is easy to explain the phenomenon of overlapping communities. The following are discussed in detail from the following aspects, such as the modelling idea, the network generation process, the characteristics of the network (direction, overlap), the complexity of the problem solving and so on.

## Statistical inference model based on vertex community

Statistical inference model based on vertex community includes: PPM (Planted partition model), NMM (Newman's mixture model), MMM (mixed membership model) MMSBM (mixed membership stochastic block model) and DCSBM (degree-corrected stochastic block model).

### Partition model

The model of planted area model is used to generate the model of benchmark test network, which belongs to the special random block model. The diagonal elements and the non diagonal elements of the random block matrix are respectively representing the link probability of the nodes in the community and the link probability of the community (Condon and Karp, 2001). The model turns community detection problem into a statistical inference problem for the ''Function'', which can be used to find the non-overlapping of the traditional community, the complexity is higher, but the speed is relatively fast on the sparse graph.

### Mixed model

The mixed model of Newman is used to detect the community that has a similar link pattern (Mungan and Ramasco, 2010; Vazquez, 2008, 2009). It can identify the traditional community, also can identify the ''non-coordinated mixed'' structure with the similar link pattern. The model assumes that nodes in the community have a similar link, not caring in the same community. The idea is similar to the random block model, but it has no clear description of the link probability relationship between the communities, and describes the relationship between the community and the node. This method can find the structure of the traditional community and the ''non-coordinated mixed'' structure, but it can't explain which kind of structure, the structure can not clearly describe the structure of the network. Its time complexity and space complexity are O (KN), which can be used to find the medium size of the network community.

### Mixed membership model

In 2003, Blei et al. proposed the LDA mixed membership model (Psorakis et al., 2010; Parkkinen et al., 2009; Blei et al., 2013; Cohn and Chang, 2000; Erosheva et al., 2004; Nallapati et al., 2008; Yang et al., 2010, 2009a,b). The model assumes that each node belongs to each class, with a probability that the membership degree vector describes the probability of each class, each vector is independent of the distribution, and the value of the vector is represented by the probability of the data. Observation object belongs to a number of classes, compared to the mixed model and simple random block model, which is more close to reality. This model mainly deals with the problem of link analysis to the network, and the existing research shows that it can improve the clustering results with the content of nodes, and the time complexity is O ($N^2K$). The current model can only deal with the problem of community detection in medium scale sparse networks.

### Mixed Membership Degree Random Block Model

The mixed membership degree model and the random block model are combined with (Airoldi et al., 2008; Airoldi and Fienberg, 2006), and the model is established. The model combines the global parameters (the block link matrix) and the local parameters (the mixed membership of the link), so as to solve the problem of the function of the pair. MMSB on the assumption that the nodes are more community and the community membership degree vector is more close to the reality, and the matrix of the membership degree of the nodes can be obtained quickly by using the variable Bayesian algorithm. The disadvantage is that for the node assignment community of ideas is not easy to extend into the hierarchical model, also network of two nodes belonging to the similarity between bigger and more easy to create a link, it implies the overlap region of the edge density higher than non-overlap region edges, which in many cases can not reflect the characteristics of real networks. The time complexity of this kind of model is O ($KN^2$), which is suitable for modelling of small scale.

### Fusion Node Degree of Random Block Model

The paper proves that the random block model without considering the degree of the node degree is easy to be combined with the large sum of nodes and the smaller community (Karrer and Newman, 2011a,b). The proposed model considers the effect of node degree on the network, and can identify the real structure of the network. In order to simplify the model, it is assumed that the network contains multiple edges and self loops, which is almost not affected by the large sparse graphs, but it is convenient to compute. Karrer et al. also designed a fast Monte Carlo iterative algorithm, which time complexity is O ($K^2$). The model can deal with large non-multiple networks, which is a non-overlapping community detection algorithm, and can be used to improve the performance of the original model in the model of overlapping community detection model and mixed membership model. The model can generate the non realistic degree sequence, and can not represent the multi-scale community structure.

## Statistical Inference Model Based On Link Community

The main statistical Inference Model Based On Link Community are: SPAEM(Simple Probabilistic Algorithm for Community Detection Employing Expectation Maximization), SBMLC (stochastic block model for link community), GSBM (general stochastic block model).