



ELSEVIER

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.

© 2015 Published by Elsevier GmbH. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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). Its 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: Shuox@sjzue.edu.cn (L. Shuo).

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), GSMB (general stochastic block model).

Table 1 Comparison of typical models of community detection based on statistical inference.

Category	Model	Structure	Type	Scale	Overlap	Time complexity
Vertex community	PPM	Traditional	Non-direction	N	N	$O(N)$
	NMM	Generalized	All	N	A	$O(KL)$
	MMM	Traditional	Direction	N	Y	$O(KN^2)$
	MMSB	Traditional	Direction	N	Y	$O(KN^2)$
	DCSBM	Generalized	Non-direction	Y	N	$O(NK^2)$
Link community	SPAEM	Traditional	Non-direction	N	Y	$O(KL)$
	SBMLC	Generalized	Non-direction	Y	A	$O(KN)$
	GSB	Generalized	All	N	A	$O(LK^2)$

Simple probabilistic algorithm expectation maximization
The SPAEM model is a special kind of symmetric joint link model. It is considered that there are many types of links in the network (Ren et al., 2009). It estimates community membership and community distribution and probability of community point node with maximum likelihood parameter. EM algorithm will not stop to iterate until convergence. Compared to the modular community detection method, it can better identify the different size, different degrees of the sequence of asymmetric network in the community; Compared to the Newman hybrid model, it is found that the effect of the traditional community is more excellent. SPAEM also provides a solution to the optimal class number selection. Using the minimum description length scheme can get a compromise between the maximum likelihood and the number of classes.

Stochastic block model for link community

Authors (Karrer and Newman, 2011a,b) proposed a method based on the global overlapping community of the linked community. Different from the existing heuristic link community detection method, it is not only assignment community for the link, but also to overcome the shortcomings which the existing community could not find a link to different scales, different structure of community by fitting the observed data through the formation of the model. The model assumes that the network is a non-direction multiple network, and one is colored in a variety of colors, each color corresponds to a community (a role that represents a node in a social network). The model can calculate the probability of the nodes belonging to the community by the node to the color edge. It can also be used to find the non-overlapping community, that is, the community which has the largest node to the color edge is selected as the community.

General stochastic block model

The general stochastic block model (Duan et al., 2011) can find a variety of intrinsic structural rules to the network without any prior information. It is assumed that the network nodes are divided into several groups, and all nodes in any two groups have similar link patterns. The model sets the group assignment as an implicit variable, and the relationship between groups is modelled as a block matrix, which is used to fit the observed data.

Analysis based on statistical inference of community detection model

Qualitative comparison of models

In order to facilitate the user to choose a community detection algorithm based on statistical inference, the following selection of various models of several representative algorithms, from many aspects of the comparison, the results see Table 1. Some of the symbols in the time complexity: N represents the number of network nodes, L represents the number of edges of the network, K represents the number of communities.

Analysis

The advantages of community detection model based on statistical inference include: it can be used to realize overlapping and non-overlapping community detection; it can find the relationship matrix between the community and the generalized community. But the model is still in the research stage, and there are many problems need to be solved. The analysis should be as follows.

- (1) There is no standard network data set for the generalized community. The advantage of statistical inference model is that it can identify the actual network in the broad community. The existing benchmark data set is mainly aimed at the traditional community detection method, which cannot be used to test the statistical inference model and can be found in many kinds of structures.
- (2) The method of overlapping community evaluation needs to be improved: the statistical inference model is obtained by the edge or node's membership degree, although the existing NMI and the module function have been extended to the overlapping community metrics, but can not effectively measure the accuracy rate of the joint membership degree.
- (3) It is not effective to link community thought and network potential. It can effectively deal with the problem of overlapping in the community. The hierarchical structure is the actual phenomenon in the network. The existing statistical inference model cannot consider the network structure.

- (4) The complexity of the algorithm is high: the current generation model uses EM algorithm to study the parameters of the algorithm, the number of iterations, the amount of each iteration, and EM algorithm to overcome the problem of local optimum, the operation efficiency is very low.
- (5) The number selection problem of the model: the community detection also belongs to the graph clustering problem. The existing SPAEM and GSB all adopt the minimum description length method to select the appropriate number, but also need the user to specify the number of classes.

Conclusions

In reality, many complex networks have complex structures, which can help us to understand the complex systems better. The community detection model based on statistical inference is trying to use the network "latent" structure to generate observation network, and use Bayesian inference to solve the problem of community detection. In addition, the method based on the stochastic block model can not only identify the node community assignment, but also find the interaction between the communities. This feature allows us to grasp the structure of the entire network from the global interaction matrix. The research of this kind of model is also related to Gibbs's algorithm, belief propagation algorithm, variational inference and so on. Research Based on statistical inference of community detection model approximate reasoning not only has important significance to the community detection, the theory research and application of the approximate solution method has the positive significance.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgement

We thank the support of VŠB-TUO activities with China with financial support from the Moravian-Silesian Region and the support of beforehand Research Foundation on National Natural Science Foundation from Shijiazhuang University of Economics.

References

- Ahn, Y.Y., Bagrow, J.P., Lehmann, S., 2010. Link communities reveal multiscale complexity in networks. *Nature* 466 (7307), 761–764.
- Airoldi, E.M., Fienberg, S.E., et al., 2006. Mixed membership stochastic block models for relational data with application to protein–protein interactions. *Proc. Int. Biom. Soc.*
- Airoldi, E.M., Blei, D.M., Fienberg, S.E., et al., 2008. Mixed membership stochastic block models. *J. Mach. Learn. Res.* 9, 1981–2014.
- Blei, D.M., Ng, A.Y., Jordan, M.I., 2013. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.
- Cohn, D., Chang, H., 2000. Learning to Probabilistically Identify Authoritative Documents Citeseer., pp. 167–174.
- Condon, A., Karp, R.M., 2001. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms* 18 (2), 116–140.
- Duan, D., Li, Y., Li, R., et al., 2011. MEI: mutual enhanced infinite generative model for simultaneous community and topic detection. *Discov. Sci.*, 9–106.
- Erosheva, E., Fienberg, S., Lafferty, J., 2004. Mixed-membership models of scientific publications. *Proc. Natl. Acad. Sci. U. S. A.* 101, 5220–5223.
- Evans, T.S., Lambiotte, R., 2009. Line graphs, link partitions, and overlapping communities. *Phys. Rev. E* 80 (1), 06105.
- Karrer, B.B.B., Newman, M.E.J., 2011a. An efficient and principled method for detecting communities in networks. *Phys. Rev. E* 84 (3), 036103.
- Karrer, B., Newman, M.E.J., 2011b. Stochastic block models and community structure in networks. *Phys. Rev. E* 83 (1), 016107.
- Mungan, M., Ramasco, J.J., 2010. Stability of maximum likelihood based clustering methods: exploring the backbone of classifications. *J. Stat. Mech.* 4, 04028.
- Nallapati, R.M., Ahmed, A., Xing, E.P., et al., 2008. Joint latent topic models for text and citations. *ACM*, 542–550.
- Parkkinen, J., Sinkkonen, J., Gyenge, A., et al., 2009. A block model suitable for sparse graphs. In: Proceeding of the 7th International Workshop on Mining and Learning with Graphs.
- Psorakis, I., Roberts, S., Sheldon, B., 2010. Partitioning in networks via Bayesian Non-negative Matrix Factorization. In: The 24th Conference on NIPS.
- Ren, W., Yan, G., Liao, X., et al., 2009. Simple probabilistic algorithm for detecting community structure. *Phys. Rev. E* 79 (3), 036111.
- Vazquez, A., 2008. Population stratification using a statistical model on hypergraphs. *Phys. Rev. E* 77 (6), 066106.
- Vazquez, A., 2009. Finding hypergraph communities: a Bayesian approach and variational solution. *J. Stat. Mech.: Theory Exp.*, 07006.
- Yang, T., Jin, R., Chi, Y., et al., 2009a. Combining link and content for community detection: a discriminative approach. *KDD*, 927–936.
- Yang, T., Chi, Y., Zhu, S., et al., 2009b. A Bayesian framework for community detection integrating content and link. *BUAI*, 615–622.
- Yang, T., Chi, Y., Zhu, S., et al., 2010. Directed network community detection: a popularity and productivity link model. *SDM*, 742–753.