Stochastic Blockmodels Meets Overlapping Community Detection

Qiqi Zhao, Huifang Ma, Zhixin Li, Lijun Guo

Stochastic Blockmodels Meets Overlapping Community Detection

Qiqi Zhao[1], Huifang Ma[1,2], Zhixin Li[2], and Lijun Guo[3]

[1]Northwest Normal University, Lanzhou Gansu 730070, China
zhaoqiqi@nwnu.edu.cn
[2]Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin Guangxi 541004, China
mahuifang@yeah.net, lizx@gxnu.edu.cn
[3]College of Information science and engineering, Ningbo University, Ningbo Zhejiang 315000, China
guolijun@nbu.edu.cn

Abstract. It turns out that the Stochastic Blockmodel (SBM) and its variants can successfully accomplish a variety of tasks, such as discovering community structures. Note that the main limitations are inferencing high time complexity and poor scalability. Our effort is motivated by the goal of harnessing their complementary strengths to develop a scalability SBM for graphs, that also enjoys an efficient inference process and discovery interpretable communities. Unlike traditional SBM that each node is assumed to belong to just one block, we wish to use the node importance to also infer the community membership(s) of each node (as it is one of the goals of SBMs). To this end, we propose a multi-stage maximum likelihood strategy for inferring the latent parameters of adapting the Stochastic Blockmodels to Overlapping Community Detection (OCD-SBM). The intuitive properties to build the model, is more in line with the real-world network to reveal the hidden community structural characteristics. Particularly, this enables inference of not just the node's membership into communities, but the strength of the membership in each of the communities the node belongs to. Experiments conducted on various datasets verify the effectiveness of our model.

Keywords: Overlapping Community Detection, Stochastic Blockmodels, Maximum Likelihood.

1 Introduction

Studies show that classical network modeling statistical models have been explored for decades. Such real-world networks contain omnipresent features to reflect small world phenomena, overlapping clusters or community structures [1,2]. Additionally, a crowd of the recent works have focused on recomposing classical statistical models to boost the performance of model statistical inference [3,4,5].

In order to partition the graph such that nodes within each group are structurally equivalent and/or tightly connected, statistical and/or probabilistic methods are typically used to partition the graph structure. Among them, stochastic blockmodel (SBM) is one prominent model for such purposes. Suppose that each node belongs to only one of the K groups is the simplest form of SBMs. Then the goal of the statistical learning is to infer the probability of connection between these unobserved groups and groups based on the observable edges of the entire graph [6,7,8]. Due to its computational flexibility and structural interpretation, SBM and its extensions have been popularizing in a variety of network analysis tasks.

Non-overlapping. Recent years have seen work on SBM implementations for non-overlapping community detection algorithm [9,10]. These methods have the same assumption that nodes in the network can only be assigned to one cluster, and the possibility of existence of edges between pairs of nodes depends only on the cluster to which they belong. Snijders [5] first present method of revealing such a cluster structure using posteriori information. The approach named ML-SBM [11] is to use SBM to develop a scalable non-overlapping community detection method on large graphs, which simply based on multi-stage MLE approach to learn latent parameters.

Overlapping. In their seminal work, Airoldi [12] proposes the first mixture-based model with overlapping communities and successfully applied to the real networks. This model, called the Mixed-Member Stochastic Blockmodel (MMSBM), is an adaptation of earlier mixed membership models [13] to the context of networks. Latouche et al [14] propose another extension of the SBM to overlapping classes, called Overlapping Stochastic Blockmodel (OSBM). The main difference between OSBM and MMSB is that the latent classes are no longer drawn from the multinomial distributions but from a product of the Bernoulli distribution.

In general, comparing to many non-attribute community detection methods, ML_SBM [11] method is based on SBM to effectively infer and learn model parameters for community detection tasks, and this algorithm performs well on most networks compared to most existing methods. It is worth noting that our work is a significant extension of ML_SBM. Yet, we consider not only learning and inferring the model latent parameters, but also introducing the importance of intuitive attributes and instinct consistent with real-world network features in overlapping community detection tasks.

In this paper, an overlapping community detection approach based on SBM, i.e. adapting the Stochastic Blockmodels to Overlapping Community Detection (OCD-SBM), is proposed to conquer the limitation of high time complexity and poor scalability of SBM. Our model explicitly encodes the importance of overlapping nodes characteristic, and thus is capable to correct the bias caused by statistical inference in the traditional SBM. In summary, the contributions of this paper are as follows:

1) We develop a fast algorithm that uses an SBM to adjust overlapping community detection in the undirected graph to address the limitations of existing algorithms of high-time complexity and scalability of large-scale networks. Contrary to other community detection methods, we use the SBM generation model to mine better clustering results in the network to preserve the characteristics of the real network.

2) Different from the rules of establishing edge between two nodes using simple SBM, we not only consider the strength of the connection between the communities,

but also the importance of nodes-to-communities. To this end, we model a method of detecting large-scale overlapping community structures in the real world via introducing the importance of intuitive attributes and instincts consistent with real-world network features in overlapping community detection tasks.

3) Various verification experiments performed on synthetic datasets and real-world datasets with ground-truth show that this is a new possibility to combine the advances in overlapping community detection and SBMs to broaden the understanding of organizing principles of complex networks.

The rest of this paper is organized as follows. Section 2 introduces the motivation and framework of our proposed model. Section 3 describes the inference algorithms in OCD-SBM. We describe the experimental results of simulations in Section 4. Section 5 concludes this paper.

2 Preliminaries

2.1 Motivation

Notation. Consider an undirected graph $G = (V, E)$, where V is the node set of size $N=|V|$, and E is the edge list of size $M=|E|$. The corresponding $N\times N$ adjacency matrix is denoted by $\mathbf{A}$, where $A_{ij} = 1$ when there is an undirected and unweighted edge for the dyad (i, j), $A_{ij} = 0$ otherwise. Let matrix $\mathbf{Z}\in[0,1]^{N\times K}$. the importance of a node is different to K blocks, where Z_{ij} represents the importance of node i for j block. And each node must subject to $\sum_{j=1}^{K} Z_{ij} = 1$. Let matrix $\mathbf{B}\in[0,1]^{K\times K}$, suggesting the probability of connection between the parameterized blocks, i.e., a node from cluster r is connected to a node from cluster s. If $r = s$, B_{rs} represents the probability of a connection within the block. The stochastic blockmodel is a special type of probability distribution over the space of adjacency arrays.

We then define the probability matrix $\boldsymbol{\theta} = \mathbf{ZBZ}^{\mathrm{T}}$ using matrices $\mathbf{B}$ and $\mathbf{Z}$. From the following model, the adjacency matrix $\mathbf{A}$ of a sample network can then be generated:

$$P(A_{ij}) = \begin{cases} \theta_{ij}, & if\ A_{ij} = 1 \\ 1-\theta_{ij}, & if\ A_{ij} = 0 \end{cases} \tag{1}$$

for $i,j\in\{1,2,\ldots,N\}$ and $i \neq j$, indicating that A_{ij} is a sample from the Bernoulli distribution with success rate θ_{ij}.

Usually in practice, the adjacency matrix $\mathbf{A}$ can be observed from the network data set. The main purpose is to ultimately estimate $\mathbf{Z}$, i.e. the block labels.

Motivation. Our motivation for proposing the overlapping version of SBM, i.e. OCD-SBM, comes from following intuitive properties:

(1) If a node is important to a community, there are edges with most nodes in the community.

(2) The connection between node i and j is affected by the connection between the community that i and j belongs to respectively, in addition to their own importance of the community they exist.

(3) Communities can overlap, as individual nodes may belong to multiple communities.

(4) If two nodes are important to multiple public communities, they are more likely to belong to the same community. (i.e., overlapping communities are more intensive).

Our ultimate goal is to capture the following three instincts that conform to the assumptions of real-world network characteristics:

(1) the possibility that a node community membership affects whether a pair of nodes are linked,

(2) the extent of the impact (probability of node connections belonging to the same community) depends on community that node belongs to, and

(3) the connection probability is independently influencing each community.

For special probability statistical models, the maximum likelihood estimation (MLE) is a setting that maximizes the parameters of likelihood function.

As defining in Eq. (1), if only **A** is given, the log-likelihood function is

$$H_1(B, Z \mid A) = \sum_{i \neq j} \log P(A_{ij}) = \sum_{i \neq j} \log[(1 - A_{ij}) + (2A_{ij} - 1)\theta_{ij}] \tag{2}$$

For large graphs, directly maximizing this likelihood function with traditional optimization methods takes too much time since there are at least N^2K unknown variables to estimate.

2.2 Framework

Figure 1 clarifies the proposed generative model. Rectangle (A_{ij}) is an entry of the observed network adjacency matrix **A**. Circles denote two latent variables: node importance strength **Z** and probability of connection **B**. In the following section, we will reveal how to estimate community memberships from node connections of the network structure (i.e., how to infer **W** from **Z** and **B**).

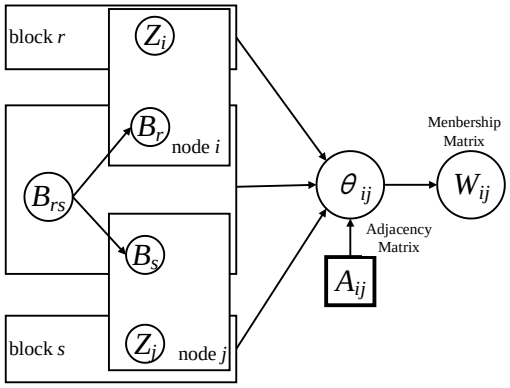


Figure 1. Plate representation of OCD-SBM. θ_{ij}: Probability that A_{ij} = 1; Z_i: Importance strength of node *i* to block *r*; Z_j: Importance strength of node *j* to block *s*; B_{rs}: the probability of connections to block *r* and *s*; W_{ij}: the project of overlapping membership node matrix.

Note that the above probability model generative process satisfies our three aforementioned requests. The network edges are created due to the importance of node-to-

block (Request (1)). Furthermore, each membership W_{ij} of a node i is regarded as an independent variable to allow a node to belong to multiple blocks simultaneously (Request (2)). This is in stark contrast to 'soft-membership' models, setting constraints $\sum_{j=1}^{K} W_{ij} = 1$ so that W_{ij} is a probability that a node i belongs to a particular block. Finally, because each block r generates connections between its members independently, nodes belonging to multiple common blocks have a higher probability of connection than that they share just a single community (Request (3)).

3 Method

3.1 Algorithm and Complexity Analysis

Our method is summarized in Algorithm 1. Our algorithm has the following advantages:

(1) **Interpretable Method.** Adapting the SBMs to overlapping community detection, we conquer the limitation of non-interpretable community detection.

(2) **Membership Strength.** In particular, this enables inference of not just the node's membership into communities, but also the strength of the membership in each of the communities the node belongs to.

(3) **More realistic.** Intuitive observations consistent with real network characteristics are proposed to quantify the importance of nodes to the community, which make sense actually.

Algorithm 1 Inference for OCD-SBM

Input: Initialization for model parameters matrixes **B**, $\mathbf{Z}^{(t)}$ into many tiny communities, $Z^{(t-1)} = Z^{(t)}$, membership matrix **W**, membership threshold ε, the number of communities K, stop criterion δ.

Output: Learned model parameters **Z**, **B**, cluster structure **W**.

1. Compute variational likelihood H_{new} by (5)
2. **repeat**
3. $H_{old} = H_{new}$
4. **inference Z, B**
5. randomly set community number to K
6. **repeat**
7. **for** $i = 1: N$ do
8. compute $N^{(d)}$ for node i
9. update Z_{ij} by block coordinate descent method via (7)
10. $Z_{ij}^{(t-1)} = Z_{ij}^{(t)}$
11. **update** $N^{(c)}$
12. **end for**
13. **update B** via (6)
14. **determine overlapping community membership**
15. update **W** via (8)
16. Compute variational likelihood H_{new} with updated parameters by (5)
17. **until** $| H_{new} - H_{old} | < \delta$

The time complexity for each updating is $O(NK^2)$. However, if we only consider the pairs of communities that have at least one edge between them, time complexity becomes $O(M)$. Community updating runs at the end of each stage after updating **Z** and **B**. The overall time cost of the algorithm is $O(tNK^2 + M)$.

3.2 Parameter Inference

The ultimate aim is to maximize the model posterior given the observations. To speed up the inferring process, a fast algorithm is proposed, which updates **B** and **Z** in turn in order to maximize the objective function $H_1(B,Z|W)$, and uses a two-stage updating framework to deal with the global optimum solution approach. Given **A**, MLE for (**B**; **Z**) can be defined as

$$\begin{aligned} &\arg\max_{0\le B,Z\le 1}\{H_1(B,Z\mid A)=\sum_{i\neq j}\log[(1-A_{ij})+(2A_{ij}-1)\theta_{ij}] \\ &=\sum_{i\neq j}\log[(1-A_{ij})+(2A_{ij}-1)\frac{\sum_{m=1}^{K}\sum_{n=1}^{K}Z_{in}B_{nm}Z_{jm}{}^{T}}{\sum_{j=1}^{N}(\sum_{m=1}^{K}\sum_{n=1}^{K}Z_{in}B_{nm}Z_{jm}{}^{T})}\} \end{aligned} \tag{3}$$

subject to $\sum_{j=1}^{K} Z_{ij} = 1$. We solve the above optimization problem by alternatively updating **B** and **Z**.

As we can see that the form of θ_{ij} is too complicated for the process of maximization. In addition, maximizing the objective function is to solve a relatively apposite value not an exact value. Therefore, we rewrite Eq. (3) as the truly function of maximum likelihood:

$$\begin{aligned} &\arg\max_{0\le B,Z\le 1}\{H_1(B,Z\mid A)=\sum_{i\neq j}\log[(1-A_{ij})+(2A_{ij}-1)\theta_{ij}^{'}] \\ &=\sum_{i\neq j}\log[(1-A_{ij})+(2A_{ij}-1)\sum_{m=1}^{K}\sum_{n=1}^{K}Z_{in}B_{nm}Z_{jm}{}^{T}]\} \end{aligned} \tag{4}$$

Such,

$$H(B,Z\mid A)=\sum_{i\neq j}\log[(1-A_{ij})+(2A_{ij}-1)\theta_{ij}^{'}] \tag{5}$$

where $\theta_{ij}^{'}=\sum_{m=1}^{K}\sum_{n=1}^{K}Z_{in}B_{nm}Z_{jm}{}^{T}$. Next we use the Optimization Strategy to update matrices **Z** and **B** in turn. When **Z** is fixed and **B** is considered as unknown, **B** is updated Gradient descent method, we have the updating strategy for elements in matrix **B**

$$B_{rs}=\frac{E}{(E+\hat{E})\sum_{m=1}^{K}\sum_{n=1}^{K}Z_{nr}Z_{ms}^{T}} \tag{6}$$

When **B** is fixed, **Z** is updated row by row utilizing the block coordinate descent method. The updating strategy for elements in matrix **Z** can be written as Eq. (7) to reduce time complexity, Z_{ij} is defined as:

$$Z_{ij}=\sum_{j=1}^{K}[B_{ij}^{N_k^{(d)}}(1-B_{ij})^{N_k^{(c)}-N_k^{(d)}}] \tag{7}$$

where $N^{(c)}\in R^K$, the entries are the number of nodes in each community, and $N^{(d)}$ is defined as the vector with the number of nodes connected to node *i* in each community.

Due to space limitations, we have omitted relevant proof.

3.3 Determine Community Membership

After learning **Z**, the ultimate goal is to determine whether node *i* belongs to block *j*. To achieve this, if Z_{ij} is below a threshold β, it can be considered that node *i* does not belong to block *j*. Otherwise ($Z_{ij} > \beta$), it can be regard *i* as belonging *j*. Specifically, let community membership matrix $W \in \{0,1\}^{N \times K}$, where W_{ij} indicates that node *i* belongs to *j* block.

$$P(Z_{ir}) = \begin{cases} W_{ij} = 1, & if\ Z_{ij} \geq \beta \\ W_{ij} = 0, & if\ Z_{ij} < \beta \end{cases} \tag{8}$$

Solving this inequality, let $\beta = \sqrt{-\log(1-\varepsilon)}$. For all our experiments we set $\varepsilon \approx 10^{-8}$ .It is worth noting that other values of β are also tested in practice, but the above-mentioned β setting provides overall good performance.

4 Empirical Study

In this section, we empirically evaluate our method with the aim of answering the following research questions:

• RQ1: How does OCD-SBM perform as compared with state-of-the-art community detection methods?

• RQ2: How does the overlapping community detection benefit from the importance of node-to-block assignment?

4.1 Experiments Settings

Datasets. Experiments are conducted on synthetic networks and several well-studied real-world datasets[1] (Table 1) with ground-truth community information to verify the effectiveness and efficiency. To be more objective and fairer, the results on the synthetic networks are omitted in experimental part.

Evaluation Metrics. For evaluation purposes, we use the metrics, Avg F1[7] and Avg NMI [14], to quantify the degree of correspondence between the detected community and the ground truth community. In view of an agreement between the ground-truth community *C** and the detected community *C*, we adopt two evaluation procedures previously used in [7] [14] to quantify performance.

Baselines for comparison. Experiments are conducted on various networks to demonstrate the effectiveness, and we compare OCD-SBM with following community detection algorithms:

MMSBM [13]: This is a dynamic model-based approach and it is a state-of-the-art overlapping community detection method using SBM.

BIGCLAM [7]: The method is an optimization-based method for overlapping community detection approach that scales to large networks of millions of nodes and edges.

ML_SBM [11]: This is a multi-stage maximum likelihood approach to recover the

[1] Networks are available at http://snap.stanford.edu and http://www.voidcn.com/article/p-plritjbe-rv.html.

latent parameters based SBM for non-overlapping community detection.

CD-SBM: In order to further explore the benefit of node-to-community assignment to overlapping community detection methods, we denote CD-SBM as the variant method of CD-SBM as we do not perform step 15 in algorithm 1 thus each row of **Z** contains only one nonzero entry.

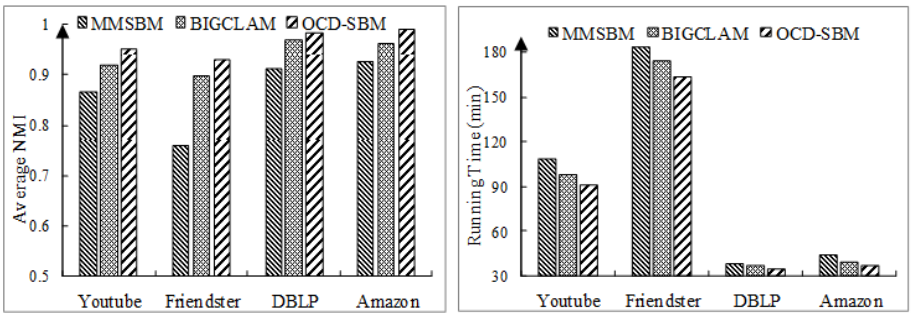
Table 1. Real-world Network Datasets statistics. *N*: number of nodes, *E*: number of edges, *C*: number of communities, *S*: average community size, *A*: community memberships per node. On average 95% of all communities overlap with at least one other community.

Datasets	Description	*N*	*E*	*C*	*S*	*A*
Youtube	Youtube online social network	1,134,890	2,987,624	8,385	26.72	0.33
Friendster	Friendster online socialnetwork	65,608,366	1,806,067,135	957,154	9.75	0.26
DBLP	DBLP collabora-tion network	317,080	1,049,866	13,477	429.79	2.57
Amazon	Amazon product network	334,863	925,872	75,149	99.86	14.83
Polblogs	political blog network	1,490	19,090	2	745	1

4.2 Performance Comparison (RQ1)

To answer (RQ1), we start by comparing the performance of all the methods, and then explore how the modeling of community membership improves on synthetic datasets and real-world networks.

Results of Real-word Networks. We conduct experiment on each dataset 500 times, comparing the average NMI with three different community detection methods. Jointly analyzing Figure 2, we have the following observations:



(a). Average NMI (b). Running Time

Figure 2. Performance Evaluation for networks on Avg NMI and Running time.

MMSBM: Although MMSBM can detect overlapping communities, the performance of this method is the worst among comparison methods in large networks. It may be that MMSBM is not suitable for large-scale community structures when dynamically updating community assignments.

BIGCLAM: BIGCLAM and OCD-SBM maintain high average NMI value on four datasets. However, BIGCLAM cannot correctly extract overlapping communities in the network because the BIGCLAM method implicitly assumes overlapping sparse connections between communities.

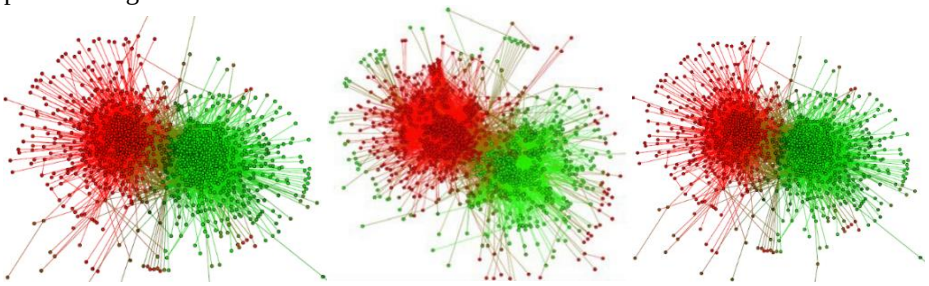
OCD-SBM: OCD-SBM performs best on four real-world networks. The reason for this may be to adopt a multi-stage maximum likelihood estimation method, which uses an important definition of nodes-to-community to accurately detect overlapping communities that are highly similar to the benchmark. OCD-SBM nearly perfectly reveals the hidden structure of the overlapping network.

4.3 Study of OCD-SBM (RQ2)

In this section, we attempt to understand how the overlapping community detection benefit from the importance of node-to-block assignment (RQ3). We observe how their representations are influenced w.r.t. the depth of OCD-SBM on political blog network. The performance evaluation of ML_SBM and CD-SBM is shown in Table 2. We have the following findings:

(a) Although the ML_SBM method can reveal the community structure, our method shows outstanding performance results in terms of NMI value and running time.

(b) Comparing the performance and clustering structure of ML_SBM and CD-SBM methods. This is consistent with the intuition that two parties have a few significant overlapping blogs of over a hundred links and the rest of the blogs with clearly connections membership. Obviously, the CD-SBM method outperforms the ML_SBM method for nonoverlapping community detection. This is verified via their clustering accuracies reported in Figure 3.



(a). Truth. (b). ML_SBM (c). CD-SBM.

Figure 3. Prediction on the political blog network: (a) Truth, manually labeled two groups by [4]. (b)ML_SBM. (c) CD-SBM. Red represents the Liberal Party cluster; green represents the Democratic Party and yellow represents the overlap.

Table 2. Performance evaluation of political blog networks: ML_SBM and CD-SBM.

Grade	ML_SBM		CD-SBM	
	0	1	0	1
Cluster result	727	743	754	736
Running time(min)	43.5		36.6	
Avg NMI	0. 9432		0.9855	

In short, even though both methods use the MLE to update community assignment parameters, the OCD-SBM approach integrates node-to-community importance into the optimization process of maximal community assignment parameters, which makes the detection community more accurate and meets real-world networks characteristics.

5 Conclusions

In this paper, we propose a fast overlapping community detection algorithm, OCD-SBM, to uses an SBM to perform on undirected graph. Intuitive observations consistent with real network characteristics are proposed to quantify the importance of nodes to the community. Combining the overlapping intuitions, we adapt SBM to overlapping community detection tasks. Our model explicitly encodes the importance of overlapping node features and is therefore able to correct for deviations caused by statistical inference in traditional SBM. OCD-SBM broadens our understanding of the organization of complex social networks and opens up new possibilities to combine community detection with advances in SBMs.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (61762078, 61363058, 61966004), Ningbo Municipal Natural Science Foundation of China (No.2018A610057), Major project of young teachers' scientific research ability promotion plan (NWNU-LKQN2019-2) and Research Fund of Guangxi Key Lab of Multi-source Information Mining and Security (MIMS18-08).

References

1. Chang H, Feng Z, Ren Z. Community detection using dual representation chemical reaction optimization[J], IEEE Transactions on Cybernetics, 47(12): 4328-4341, (2017).
2. Yang L, Cao X. A unified semi supervised community detection framework using latent space graph regularization[J], IEEE Transactions on Cybernetics,45(11):2585-2598, (2015).
3. Qiao M, Yu J, Bian W, et al. Adapting Stochastic Block Models to Power-Law Degree Distributions[J]. IEEE Transactions on Cybernetics, 1-12. (2018).
4. Goldenberg A, Zheng A X, Fienberg S E, et al. A survey of statistical network models[J]. Foundations and Trends® in Machine Learning, 2(2): 129-233, (2010).
5. Chen J, Xu G, Wang Y, et al. Community Detection in Networks Based on Modified PageRank and Stochastic Block Model[J]. IEEE Access, 6, 77133-77144, (2018).
6. Lee C, Wilkinson DJ. A Review of Stochastic Block Models and Extensions for Graph Clustering[J]. (2019).
7. Yang J, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach[C]// Proceedings of the sixth ACM international conference on Web search and data mining. ACM, (2013).
8. Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multi-scale complexity in networks[J]. Nature, 466,761-764, (2010).
9. Cherifi H. Non-overlapping community detection[J]. arXiv:1805.11584 (2018).
10. Boccaletti S, Latora V, Moreno Y, Chavez M. Complex networks: Structure and dynamics. Physical Review 424, 175-308, (2006).

11.Peng C, Zhang Z, et al. A scalable community detection algorithm for large graphs using stochastic block models[J]. Intelligent Data Analysis, 21(6):1463-1485, (2017).
12.Airoldi EM, Blei DM, SE Fienberg, EP Xing. Mixed membership stochastic blockmodels[J]. Journal of Machine Learning Research, 9:1981-2014, (2008).
13.Griffiths T, Ghahramani Z. Infinite latent feature models and the Indian buffet process[J]. In Neural Information Processing Systems, volume 18, 475-482, (2005).
14.Latouche P, Birmele E, Ambroise C. Overlapping stochastic block models with application to the french political blogosphere[J]. Annals of Applied Statistics, 5(1):309-336, (2011).