Electronic Journal of Statistics
Vol. 12 (2018) 1842–1866
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1445

Community detection by L_0-penalized graph Laplacian

Chong Chen, Ruibin Xi

School of Mathematical Sciences and Center for Statistical Science, Peking University
5 Yihe Yuan Road, 100871, Beijing, China
e-mail: cheung1990@126.com*;* ruibinxi@math.pku.edu.cn

and

Nan Lin

Department of Mathematics, Washington University in St. Louis
One Brookings Drive, 63130, St. Louis, USA
e-mail: nlin@wustl.edu

Abstract: Community detection in network analysis aims at partitioning nodes into disjoint communities. Real networks often contain outlier nodes that do not belong to any communities and often do not have a known number of communities. However, most current algorithms assume that the number of communities is known and even fewer algorithm can handle networks with outliers. In this paper, we propose detecting communities by maximizing a novel model free tightness criterion. We show that this tightness criterion is closely related with the L_0-penalized graph Laplacian and develop an efficient algorithm to extract communities based on the criterion. Unlike many other community detection methods, this method does not assume the number of communities is known and can properly detect communities in networks with outliers. Under the degree corrected stochastic block model, we show that even for networks with outliers, maximizing the tightness criterion can extract communities with small misclassification rates when the number of communities grows to infinity as the network size grows. Simulation and real data analysis also show that the proposed method performs significantly better than existing methods.

MSC 2010 subject classifications: Primary 62-09; secondary 62P10.
Keywords and phrases: Consistency, degree corrected stochastic block model, spectral clustering, outlier, social network, gene regulatory network.

Received December 2017.

Contents

1. Introduction

Community detection has attracted tremendous research attention, initially in the physics and computer science community [22, 25, 24] and more recently in the statistics community [3, 4, 34, 13]. Considering an undirected network $G = (V, E)$, where V is the set of nodes and E is the set of edges. Community detection is to find an "optimal" partition of the nodes $V = G_1 \bigcup \cdots \bigcup G_K$ such that nodes within the communities G_k $(k = 1, \cdots, K)$ are more closely connected than nodes between the communities.

One class of community detection algorithms detects community by optimizing a heuristic global criterion over all possible partitions of the nodes. For example, modularity [25] has been very popular in community detection and fast algorithms for maximizing modularity [23] have been developed and widely used. The well-known spectral clustering algorithms [13, 2, 6, 27, 14] can be traced back as continuous approximation methods of global criterion such as ratio cut [10] or normalized cut [28]. Spectral clustering methods are fast in computation and easy to implement since they usually only require calculation of a few eigenvectors of the Laplacian matrix.

Probabilistic model-based methods are another class of community detection algorithms. They detect communities by fitting a probabilistic model [4, 26, 21, 8] or by optimizing a criterion derived from a probabilistic model [3, 15]. One of the most commonly used models is the stochastic block model (SBM) [12]. Given the adjacency matrix $A = (A_{ij})_{1 \leq i,j \leq n}$ of a network G with n nodes, the SBM assumes that true node labels c_i are independently sampled from a multinomial distribution with parameters $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)^T$, i.e. $\pi_k = P(c_i = k)$, $k = 1, \cdots, K$. Conditional on the community labels, the edges A_{ij} $(i < j)$ are independent Bernoulli random variables with $P(A_{ij} = 1 | c_i, c_j) = p_{c_i c_j}$. The SBM assumes that the expected degrees are the same for all nodes in the same community and thus cannot allow hubs in the network. To remove this constraint, the degree corrected stochastic block model (DCSBM) [15] introduces a degree correction variable θ_i to each node such that $P(A_{ij} = 1 | c_i, c_j, \theta_i, \theta_j) = \theta_i \theta_j p_{c_i c_j}$, where $\theta_i > 0$ and $E(\theta_i) = 1$.

Consistency results were developed for a number of community detection algorithms, mostly based on the SBM or DCSBM. Under the assumption that the community number is fixed, Bickel and Chen [3] laid out a general theory under the SBM for checking consistency of community detection criteria when the network size grows to infinity, and similar theories were also developed for DCSBM [34, 13]. With a fixed community number, the community size would linearly grow as the number of nodes grows. However, this is not a realistic assumption, because real networks often have tight communities at small scales, even when networks contain millions of nodes [20]. Recent researches [27, 7, 5] generalized these consistency results by allowing the number of communities grows to infinity. However, as far as we know, similar results for the DCSBM are not available yet.

Despite all these progresses, current algorithms implicitly assume that all nodes of the network belong to a community. However, many real networks contain outlier nodes. These outlier nodes do not belong to any community and they just loosely connect to other nodes in the network. Ignoring these outlier nodes can significantly influence the accuracy of community detection. In addition, real networks often do not have a known number of communities. Although several methods have recently been proposed to estimate the number of communities [18, 29], but these methods are also based on the assumption that all nodes belong to a community. A few available algorithms [17, 33] can detect communities for networks with ourliers and unknown community numbers. However, there is no theoretical result to guarantee the consistency of these methods when there are outliers in the network.

In this paper, we propose a novel model-free tightness criterion for community detection. Community detection based on this criterion iteratively extracts single communities and no prior knowledge about the community number is needed. Maximizing this criterion is closely related with the L_0-penalized graph Laplacian. An efficient algorithm is developed based on the alternating direction method of multiplier (ADMM) to maximize this penalized Laplacian. A permutation-based test is performed to filter the extracted communities that are likely to be outliers or false communities. Under the DCSBM and the DCSBM with outliers, we establish asymptotic consistency allowing the community number K increases as the number of nodes grows. Simulation and real data analysis show that the proposed method can computationally efficiently recover the community structure with high resolution and accuracy. This paper is organized as follows. The model-free criterion and the ADMM algorithm are described in Section 2. Theoretical results are given in Section 3. Section 4 presents simulation comparison with existing methods and Section 5 is the real data analysis. Proofs of the theorems are given in the Appendix.

2. Method and algorithm

Assume that nodes of a graph $G = (V, E)$ are indexed by $\{1, 2, ..., n\}$ and each node i belongs to exactly one of K non-overlapping communities denoted by a latent label $c_i \in \{1, ..., K\}$. Given a set $S \subset V$, the complementary set of S is denoted by $\bar{S}$ and the number of elements in S is denoted as $|S|$. Define $W(S) = \sum_{i,j \in S} A_{ij}$, $B(S) = \sum_{i \in S, j \in \bar{S}} A_{ij}$ and $V(S) = W(S) + B(S)$. Then, $W(S)$ is twice the number of edges between nodes in S, $B(S)$ is the total number of edges between S and $\bar{S}$ and $V(S)$ is the total degrees in S. Given a vector $\mathbf{u}$, we denote $\|\mathbf{u}\|_0$ as the number of nonzero elements in $\mathbf{u}$ and $\|\mathbf{u}\|_2$ as the L_2-norm of the vector $\mathbf{u}$.

2.1. A tightness criterion

Given a set $S \subset V$, if it is a true community, we expect that most of its connections are within S itself and thus $W(S)/V(S)$ should be large. However, directly maximizing $W(S)/V(S)$ has a trivial solution $S = V$. We instead introduce a penalty to the size of the community and consider the following tightness criterion,

$$\psi(S) = \frac{W(S)}{V(S)} - \eta\,|S|\,, \tag{2.1}$$

where η is a tuning parameter. In Section 3, we will show that with a proper choice of η, maximizing this tightness criterion can render consistency in community detection.

The quantity $B(S)$ is known as the cut between S and $\bar{S}$ [10]. True communities should have a small cut value. However, the entire network V or single nodes with no connections all have zero cut values. To avoid these trivial solutions, the ratio cut minimizes $B(S)/(|S||\bar{S}|)$ for community detection [10]. The denominator $|S||\bar{S}|$ can be viewed as a penalty to guard against too large or too small communities. Similarly, the normalized cut minimizes $B(S)/V(S)+B(S)/V(\bar{S})$ for community detection [28]. The denominators $V(S)$ and $V(\bar{S})$ are penalties for the community size. The criterion proposed in Zhao et al. 2011 [33] also has a penalty for both S and $\bar{S}$. Since these criteria penalize both $|S|$ and $|\bar{S}|$, they perform best when the community sizes are similar. In comparison, the tightness criterion (2.1) only penalizes $|S|$. This endows the tightness criterion with a high detection power for both large and small communities. However, only penalizing $|S|$ can also lead to small spurious communities and we use a resampling procedure to remove these potential false communities in section 2.3.

The tightness criterion 2.1 is closely related to a penalized graph Laplacian. More specifically, let $Q = D^{-1/2}AD^{-1/2}$ be the graph Laplacian, where A is the adjacency matrix and $D = \mathrm{diag}\{d_1, \cdots, d_n\}$ is the nodal degree matrix with d_i being the degree of the ith node. Then, we have the following proposition.

Proposition 2.1. *Given a set $S \subset V$, define its membership vector by*

$$\mathbf{u}_S(i) = \begin{cases} \frac{\sqrt{d_i}}{\sqrt{W(S)+B(S)}}, & \textit{if} \quad i \in S, \\ 0, & \textit{if} \quad i \in \bar{S}. \end{cases} \tag{2.2}$$

Then we have $\psi(S) = \mathbf{u}_S^t Q \mathbf{u}_S - \eta\|\mathbf{u}_S\|_0$ and $\|\mathbf{u}_S\|_2 = 1$.

Therefore, maximizing the tightness criterion (2.1) is equivalent to the following optimization problem

$$\max_{S \subset V, \mathbf{u}=\mathbf{u}_S} \mathbf{u}^t Q \mathbf{u} - \eta\|\mathbf{u}\|_0. \tag{2.3}$$

Finding the global solution to (2.3) is difficult in general, because we have to search over all possible subsets of V. In the next section, we will develop an efficient algorithm based on the ADMM to find a local optimal.

2.2. Algorithm

Before introducing the algorithm, we first give some notations. For any $\mathbf{u}$ with $\|\mathbf{u}\|_2 = 1$, we denote its nonzero element index set $S(\mathbf{u}) = \{i : \ \mathbf{u}(i) \neq 0\} \subset V$. On the other hand, given $S(\mathbf{u})$, we can define a new membership vector $\mathbf{u}_d = \mathbf{u}_{S(\mathbf{u})}$ using (2.2). The vector $\mathbf{u}_d$ is obtained just by reassigning values of the nonzero elements of $\mathbf{u}$ according to the degrees of $S(\mathbf{u})$. Note that $\mathbf{u}_d$ satisfies $\|\mathbf{u}_d\|_2 = 1$. Given $\lambda_1 \geq 0$, we consider the following optimization problem

$$\max_{\|\mathbf{u}\|_2=1} \ \mathbf{u}^t Q\mathbf{u} - \eta\|\mathbf{u}\|_0 - 2\lambda_1\|\mathbf{u} - \mathbf{u}_d\|_2^2, \tag{2.4}$$

which can be viewed as the augmented Lagrangian of (2.3). When λ_1 is sufficient large, $\mathbf{u}$ will be forced to be $\mathbf{u}_d$. It is easy to see that $\mathbf{u}$ can only take discrete values in the optimization (2.3). So it is much easier to solve (2.4) than solving (2.3) since $\mathbf{u}$ can be any vector with norm 1 in (2.4). By introducing an intermediate variable $\mathbf{v}$ with $\mathbf{v} = \mathbf{u}$, the augmented Lagrangian of (2.4) is

$$\max_{\|\mathbf{u}\|_2=1, \|\mathbf{v}\|_2=1} \mathbf{u}^t Q\mathbf{v} - \lambda\|\mathbf{u}-\mathbf{v}\|_2^2 - \frac{\eta}{2}(\|\mathbf{u}\|_0 + \|\mathbf{v}\|_0) - \lambda_1\|\mathbf{u}-\mathbf{u}_d\|_2^2 - \lambda_1\|\mathbf{v}-\mathbf{v}_d\|_2^2, \tag{2.5}$$

which is equivalent to

$$\max_{\|\mathbf{u}\|_2=1, \|\mathbf{v}\|_2=1} \mathbf{u}^t(Q + 2\lambda I)\mathbf{v} - \frac{\eta}{2}(\|\mathbf{u}\|_0 + \|\mathbf{v}\|_0) + 2\lambda_1\mathbf{u}^t\mathbf{u}_d + 2\lambda_1\mathbf{v}^t\mathbf{v}_d. \tag{2.6}$$

We alternatively update $\mathbf{u}$, $\mathbf{u}_d$, $\mathbf{v}$ and $\mathbf{v}_d$ to solve (2.6). Given $\mathbf{u}$ or $\mathbf{v}$, we could easily get $\mathbf{u}_d$ and $\mathbf{v}_d$ using the d-operator defined above. Given other variables, updating $\mathbf{u}$ or $\mathbf{v}$ reduces to a simple linear programming problem which has an explicit solution given by the following proposition.

Proposition 2.2. *For a given vector $\mathbf{z} = (z_1, ..., z_n)^t \in \mathbb{R}^n$, we denote its rth largest absolute value as $|z|_r$, and let $\mathbf{z}_r^h$ be the vector with the ith element as $\mathbf{z}_r^h(i) = z_i\mathbf{I}(|z_i| > |z|_{r+1})$. Then for a constant $\rho > 0$, the solution to*

$$\max_{\|\mathbf{u}\|_2=1} \mathbf{u}^t\mathbf{z} - \rho\|\mathbf{u}\|_0 \tag{2.7}$$

is $\mathbf{u} = L(\mathbf{z}, \rho) = \mathbf{z}_r^h/\|\mathbf{z}_r^h\|_2$, where r is the smallest integer that satisfies

$$|z|_{r+1} \leq \sqrt{\rho^2 + 2\rho\|\mathbf{z}_r^h\|_2}. \tag{2.8}$$

The proof of Proposition (2.7) is given in [16] and we omit it here. We summarize the algorithm for the optimization problem (2.6) in the following L0Lap algorithm.

Algorithm 1 L_0-Penalized Laplacian Algorithm(L0Lap)

```
Require: $Q$, $\lambda$, $\lambda_1$, $\eta$ and $\epsilon$
  Initialize $\mathbf{v}^0$, $\mathbf{u}^0$. For each $k = 1, 2, \cdots$,
  repeat
    $\mathbf{z}_1^k = (Q + 2\lambda I)\mathbf{v}^{k-1} + 2\lambda_1 \mathbf{u}_d^{k-1}$, $\mathbf{u}^k = L(\mathbf{z}_1^k, \eta/2)$;
    $\mathbf{z}_2^k = (Q + 2\lambda I)^t \mathbf{u}^k + 2\lambda_1 \mathbf{v}_d^{k-1}$, $\mathbf{v}^k = L(\mathbf{z}_2^k, \eta/2)$;
  until $\|\mathbf{u}^k - \mathbf{v}^k\| < \epsilon$.
  return $S_\eta = \{i, \mathbf{u}^k(i) \neq 0$ and $\mathbf{v}^k(i) \neq 0\}$
```

In all simulation and real data analysis, we set the convergence tolerance parameter ϵ as 10^{-4}. The parameters λ and λ_1 are the penalty parameters in the augmented Lagrangian and they can be chosen as fixed [35]. Throughput the paper, we set $\lambda = 1/\sqrt{n}$. For λ_1, we first set $\lambda_1 = 0$ and run Algorithm 1 with the initial value $\mathbf{v}^0 = (1/\sqrt{n}, ..., 1/\sqrt{n})$ to get $\hat{\mathbf{v}}^0$. Then, we set $\lambda_1 = 1$ and run Algorithm 1 with the initial value $\hat{\mathbf{v}}^0$ to get the final solution. Although Algorithm 1 cannot guarantee a global maximum for (2.4), we find that this process achieves robust results and good performance in the numerical analyses.

The parameter η is the most important tuning parameter and we introduce a criterion to tune the parameter η. Given a subset $S \subset V$, define $\bar{p}_W(S) = W(S)/(|S|(|S|-1))$ and $\bar{p}_B(S) = B(S)/(|S||\bar{S}|)$. Thus, $\bar{p}_W(S)$ is the average connection within S and $\bar{p}_B(S)$ is the average connection between S and $\bar{S}$. A true community S should has relative large $\bar{p}_W(S)$ and small $\bar{p}_B(S)$. Define

$$\phi(S) = \frac{\bar{p}_W(S)}{\bar{p}_W(S) + \bar{p}_B(S)}. \tag{2.9}$$

A large $\phi(S)$ implies that S has more connections within itself and thus would be more likely to be a community. From the theoretical results in section 3, we know that the best η is at the order of $O(1/n)$. Therefore, we run Algorithm 1 for $\eta = 0, 1/10n, 2/10n, \cdots, 10/10n$ and choose the η such that the resulted S_η achieves the largest $\phi(S_\eta)$.

2.3. The permutation test

After a community is extracted by Algorithm 1, we remove it from the network and iteratively apply Algorithm 1 to the remaining network until there is no edge left. However, it may lead to some small spurious communities during this process, because even Erdös-Rényi (ER) networks can have small community-like structures. To filter these spurious communities, we introduce the following permutation test.

Suppose that $S_1, \cdots, S_c$ are all the identified communities with less than M nodes and G_0 is the sub-network of G composed of nodes in $\bigcup_{i=1}^c S_i$. Let $\bar{p} = \sum_{i,j} A_{ij}/(n^2 - n)$. Given a subset S of G_0, if S is an ER-graph with a connecting probability $\bar{p}$, given any m nodes, the probability of observing no more than E edges between these m nodes is

$$p(m,E)=\sum_{i=0}^{E}\binom{m(m-1)/2}{i}\bar{p}^{i}(1-\bar{p})^{m(m-1)/2-i}. \tag{2.10}$$

Let n_i and E_i be the number of nodes and number of edges in S_i $(i=1,\cdots,c)$, respectively. Each detected community S_i has an associated probability $p(n_i,E_i)$ using (2.10). We permute N times the edges in G_0 to generate N ER-graphs and run Algorithm 1 to each of the N ER-graphs. The first extracted community of the jth ER-graph also has a probability p_j^{ER} using (2.10). Note that $p(n_i,E_i)$ should be less than most p_j^{ER} if S_i is a true tight community. We assign the permutation p-value for S_i as $p_i=|\{j:\ p_j^{ER}\geq p(n_i,E_i),\ j=1,\cdots,N\}|/N$. The detected community S_i is filtered out if $p_i\geq\alpha$. In our simulation and real data, we set $M=20,N=100$ and $\alpha=0.05$.

3. Theoretical properties

In this section, we discuss theoretical results about the estimator S that maximizes the tightness criterion (2.1) under the DCSBM and the DCSBM with outliers. We first give the exact definition of the DCSBM.

Definition 3.1. *A network $G=(V,E)$ is said to follow a DCSBM, if it satisfies the following assumptions.*

(A1) Each node is independently assigned a pair of latent variables (c_i,θ_i), where c_i is the community label taking values in $\{1,2,...,K\}$, and θ_i is a "degree variable" taking discrete values in $\{h_1,\cdots,h_M\}$ $(0<h_1<...<h_M)$.

(A2) The marginal distribution of c_i is a multinomial distribution with parameters $\boldsymbol{\pi}=(\pi_1,...,\pi_K)^T$, and the random variable θ_i satisfies $E[\theta_i]=1$ for identifiability.

(A3) Given $\mathbf{c}=(c_1,...,c_n)$ and $\boldsymbol{\theta}=(\theta_1,...,\theta_n)$, the edges A_{ij} $(i<j)$ are independent Bernoulli random variables with $P(A_{ij}=1|\mathbf{c},\boldsymbol{\theta})=\theta_i\theta_j p_{c_ic_j}$.

(A4) Denote $\pi^-=\min_{1\leq k\leq K}\pi_k$, $p^-=\min_{1\leq k\leq K}\{p_{kk}\}$ and $q^+=\max_{k\neq m}\{p_{km}\}$. Then, $p^->q^+$.

Throughout this paper, we assume that $\alpha,\tau,\gamma,\delta$ are fixed constants such that $0\leq 2\delta<\alpha<1/2$ and $0<\tau<\gamma<\alpha-2\delta$. The constant α is to control the lower bound for the within-community connection probabilities and the constant δ is to control the smallest community size. The constant τ is to separate different communities by a feature depending on the community size (π_k) and the within-community connection probabilities (p_{kk}). Given any two sets S_1 and S_2, we denote $S_1\Delta S_2=S_1\bigcup S_2-S_1\bigcap S_2$ as their symmetric difference. For two nonnegative sequences a_n and b_n, we write $a_n\gtrsim b_n$ if there exists a constant $C_0>0$ such that $a_n\geq C_0b_n$. Define $\Gamma_\delta=\{S\subset V,|S|^2/K\gtrsim n^{2-2\delta}\}$. Similar to [34], we assume Π is the $K\times M$ matrix representing the joint distribution of (c_i, θ_i) with $\mathbb{P}(c_i = k, \theta_i = h_l) = \Pi_{kl}$. Denote $\pi_k^d = \sum_l h_l \Pi_{kl}$. Note that since $E[\theta_i] = 1$, we have $\sum_k \pi_k^d = 1$. Let $\rho_k^d = p_{kk}/\sum_{l=1}^K \pi_l^d p_{kl}$.

Given the community label $\mathbf{c}$ and a set of nodes $S \in \Gamma_\delta$, denote $G_k = \{i | c_i = k,\ i = 1, \cdots, n\}$, $S_k = \{i | i \in S, c_i = k\}$, $\pi_k = |G_k|/n$ and $r_k(S) = |S_k|/n$ for $1 \le k \le K$. We define $\hat{\pi}_k^d = \pi_k^d \hat{\pi}_k/\pi_k$, $r_k^d(S) = \pi_k^d r_k(S)/\pi_k$, $r^d(S) = \sum_{k=1}^K r_k^d(S)$ and $\hat{\rho}_k^d = p_{kk}/\sum_{l=1}^K \hat{\pi}_l^d p_{kl}$. For $S = G_1$, we have $r_1(G_1) = \hat{\pi}_1$, $r_1^d(G_1) = \hat{\pi}_1^d$, $r_k(G_1) = 0$, $r_k^d(G_1) = 0$ $(k = 2, \cdots, K)$ and $r^d(G_1) = \hat{\pi}_1^d$. Let $x_k = p_{kk}$, $y_k = \sum_{l=1}^K \hat{\pi}_l^d p_{kl}$ for $1 \le k \le K$. For any $t_k \ge 0$ $(k = 1, \cdots, K)$ and $\sum_{k=1}^K t_k = 1$, define

$$f(t_1, ..., t_K) = \frac{\sum_{k=1}^K t_k (t_k x_k + \sum_{l \neq k}^K t_l p_{kl})}{\sum_{k=1}^K t_k y_k}.$$

Theorem 3.1. *Under the assumptions of DCSMB, assume* $\rho_1^d - \max_{2 \le k \le K} \rho_k^d \gtrsim n^{-\tau}$ *and* $\pi_1^d/\pi_1 \ge \max_{2 \le k \le K} \pi_k^d/\pi_k$. *If* $p^- \gtrsim \log n/n^{1-2\alpha}$ *and* $\pi^- \gtrsim n^{-\delta/2}$, *there is a constant* C *such that, with probability at least* $1 - 2Kn^{-2}$, *we can choose* $\eta > 0$ *such that*

$$f(1, 0, ..., 0)\frac{\pi_1^d}{\pi_1} - \frac{C}{n^\gamma} > n\eta > \max_{t_1 \le 1 - 1/n^{\gamma - \tau}} f(t_1, t_2, ..., t_K)\frac{\pi_1^d}{\pi_1} + \frac{C}{n^\gamma}. \tag{3.1}$$

With such a choice of η, *suppose that* $S \subset V$ *is such that the tightness criterion (2.1) is maximized in* Γ_δ, *then with probability at least* $1 - (2K)n^{-2} - 2^{n+2}/n^n$,

$$\frac{|S \Delta G_1|}{|S \bigcup G_1|} \le 2h_M h_1^{-1}/n^{\gamma - \tau} + \log n/n^{\alpha - 2\delta - \gamma}. \tag{3.2}$$

This theorem says that under a number of regularity conditions, if the tuning parameter is chosen properly, the detected community S is very close to the underlying true community G_1 which has the largest ρ_k^d.

Remark 3.1. *In Theorem 3.1, the maximum is taken over* $S \in \Gamma_\delta$. *This constraint is needed because small spurious communities could generate a smaller tightness criterion (2.1) than the true communities. However, we find it difficult to develop an efficient algorithm with this constraint and hence this constraint is not added in Algorithm 1. Instead, we filter the potential small spurious communities by a resampling procedure.*

Remark 3.2. *The condition* $\pi_1^d/\pi_1 \ge \max_{2 \le k \le K} \pi_k^d/\pi_k$ *is not as restrictive as it looks. For example, if* $\mathbf{c}$ *and* $\boldsymbol{\theta}$ *are independent, then* $\pi_k^d = \pi_k$ *for all* $1 \le k \le K$ *and this condition is naturally satisfied. The SBM clearly also satisfies this condition, since in this case* $M = 1$ *and* $h_1 = 1$.

Remark 3.3. *The condition* $\rho_1^d - \max_{2 \le k \le K} \rho_k^d \gtrsim n^{-\tau}$ *is to make sure that the first community is separable from the other communities. Consider a simple case of SBM when* $p_{kl} = p_0$ *for all* $k \neq l$, *we have* $\rho_k^d = 1/\left((1 - p_0/p_{kk})\pi_k + p_0/p_{kk}\right)$. *The ratio* $\beta_k = p_0/p_{kk}$ *can be viewed as the "out-in-ratio" defined in [8]. If all* π_i*'s are the same, the first extracted community* G_1 *is the community with the smallest out-in-ratio. If all out-in-ratios* β_k *are the same, the first extracted community* G_1 *is the community with the smallest size.*

Since $p^- \gtrsim \log n/n^{1-2\alpha}$ and $\pi^- \gtrsim n^{-\delta/2}$, we have $p^- n^{1-2\alpha+\delta/2} \gtrsim K \log n$. Consider a special case when K is finite and the community sizes are all $O(n)$. In this case, the lower bound of the connecting probability within communities should satisfies $p^- \gtrsim \log n/n^{1-2\alpha+\delta/2}$ and thus $np^-/\log n \gtrsim n^{2\alpha-\delta/2}$. This condition is similar to the condition $np^-/\log n \to \infty$ in [34], especially when α is close to 0. If $p^- = O(1)$ and $\delta = 1/4 - 2\epsilon$ for some $\epsilon > 0$ very close to 0, then $n_{min} = O(n^{7/8+\epsilon})$ and $K = O(n^{1/8-\epsilon})$. Thus, the upper bound of K is $O(n^{1/8})$. Consider the simplest case when $K = 2$. Let $\tau = 0$ and $\gamma = \alpha/2 - \delta$, the misclassification rate is about $O_p(\log n/n^{\alpha/2-\delta})$ by the inequality (3.2). This improves the results in [27] and [19] where the misclassification rate was $O_p(1/\log n)$.

When there are outliers in networks, we also have a consistency result similar to Theorem 3.1. We first give the definition of the DCSBM with outliers.

Definition 3.2. *A network $G = (V, E)$ is said to follow a DCSBM with outliers, if it satisfies all assumptions (A1)-(A3) and the following assumption.*

(A4′) Denote $\pi^- = \min_{1\leq k\leq K-1} \pi_k$, $p^- = \min_{1\leq k\leq K-1}\{p_{kk}\}$ and $q^+ = \max_{k\neq m}\{p_{km}\}$. Then, $p^- > q^+ \geq p_{KK}$. The Kth community is called the outlier community.

For a DCSBM with outliers, all communities are well-defined communities except the Kth outlier community. We also assume $p^- \gtrsim \log n/n^{1-2\alpha}$ and $\pi^- \gtrsim n^{-\delta/2}$ for the DCSBM with outliers. We have the following theorem.

Theorem 3.2. *Suppose that G is a DCSBM with outliers. Assumes that all conditions in Theorem 3.1 hold. In addition, assume that the outlier community G_K satisfies $|G_K|^2/K = o(n^{2-2\delta})$. Then, the conclusions in Theorem 3.1 hold.*

This theorem says that as long as the outlier community is not too large, the first extracted community will be very close to the community with the largest ρ_k^d.

4. Simulation study

In this section, we perform simulation to compare the proposed method with state-of-the-art algorithms including SCORE [13], nPCA [28], OSLOM [17], Zhao [33], and PLH [29]. Since SCORE and nPCA require a known community number, we provide the true community number to these algorithms in the simulation. For the algorithm developed in this paper, we consider two versions of the algorithm, with or without the permutation test. This helps us to see the effect of the permutation test on removing false communities. We call these algorithms L0Lap (without the permutation test), L0LapT (with the permutation test). We evaluate the performance of the algorithms by the normalized mutual information (NMI) [31] between the detected community and true community. For methods that can automatically determine the community number, we also compare their estimated community numbers. In addition, we also consider another two algorithms, NB and BH [18], when comparing the accuracy for the community number estimation. Since NB and BH can only estimate the community number, we do not evaluate their performance in terms of NMI. For OSLOM, we use the C++ implementation available at http://www.oslom.org/software.htm. The computer codes of the algorithms Zhao, NB, BH and PLH were provided by the original authors. For the other methods, we implement the algorithms using Matlab according to their respective descriptions. L0Lap and L0LapT were implemented by Matlab which are available at https://github.com/ChongC1990/L0Lap.

Computationally, we find that SCORE and nPCA are computationally most efficient methods. For a network with 1000 nodes, they can finish computation in less than 0.1 second. The proposed method can finish computation in 3 seconds for a 1000-node network. The Zhao method and OSLOM need 50 seconds and 282 seconds to process a 1000-node network, respectively. For a larger network with 10,000 nodes, SCORE and nPCA can finish computation in a few seconds. The proposed method can finish in 70 seconds. In comparison, the Zhao method requires more than 3500 seconds and OSLOM is unable to give any result.

4.1. Simulation without Outliers

We perform the simulation under both SBM and DCSBM. All simulated networks have $n = 1,000$ nodes and $K = 21$ communities of different sizes. Among the 21 communities, 5 of them have 100 nodes, 6 have 50 nodes and 10 have 20 nodes. For the DCSBM, $\mathbf{\Theta} = (\theta_{ij})$ are drawn independently from $U\,[0.5, 1]$. For SBM, all elements of $\mathbf{\Theta}$ are set as 1. Similar to [1], the connecting matrix $\mathbf{P}$ is constructed depending on an "out-in-ratio" parameter β [8]. Given a β, we set the diagonal elements of matrix $\mathbf{P}^{(0)}$ as β^{-1} and set all off-diagonal elements as 1. Then, given an overall expected network degree Λ, we rescale $\mathbf{P}^{(0)}$ to give the final $\mathbf{P}$:

$$\mathbf{P} = \frac{\Lambda}{(n-1)(\boldsymbol{\pi}^T\mathbf{P}^{(0)}\boldsymbol{\pi})(\mathbb{E}\mathbf{\Theta})^2}\mathbf{P}^{(0)},$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_{21})$ is the proportion of nodes in each community. Conditional on the labels and $\mathbf{P}$, the edges between nodes are generated as independent Bernoulli variables with parameter $\theta_i\theta_j P_{ij}$. The methods NB, BH and PLH require a candidate set of K, we provide the candidate set by all possible values from 1 to 25. For L0LapT, OSLOM and Zhao, since there will be unclassified nodes, we only consider nodes that are assigned with a community label when calculating the NMI.

We first fix $\Lambda = 50$ and vary the out-in-ratio parameter β from 0.02 to 0.2. For each β, the mean NMI of each algorithm is summarized by 100 repetitions (Figure 1). For all algorithms, the NMIs tend to decrease as β increases. We clearly see that our algorithm achieves the highest NMI compared with other methods. As expected, after the permutation test, the NMI can be significantly improved. This is because the permutation test successfully removes small false communities. Especially, when β is large, the test is more effective in terms of improving the NMI. For example, under the SBM, when $\beta = 0.2$, the NMI of

L0Lap is similar to that of nPCA, but after applying the permutation test, the NMI of L0Lap becomes close to 1. Furthermore, since nodes in the DCSBM are heterogeneous, as expected, all methods perform better in the SBM than in the DCSBM. In terms of the community number, L0LapT and PLH give comparable estimates and are usually better than other algorithms (Figure 1, the bottom panel). When the out-in-ratio β is large, other than the Zhao method, all other methods tend to underestimate the community number. The Zhao method prefers to find a big community and many small communities, so it often overestimates the number of communities.

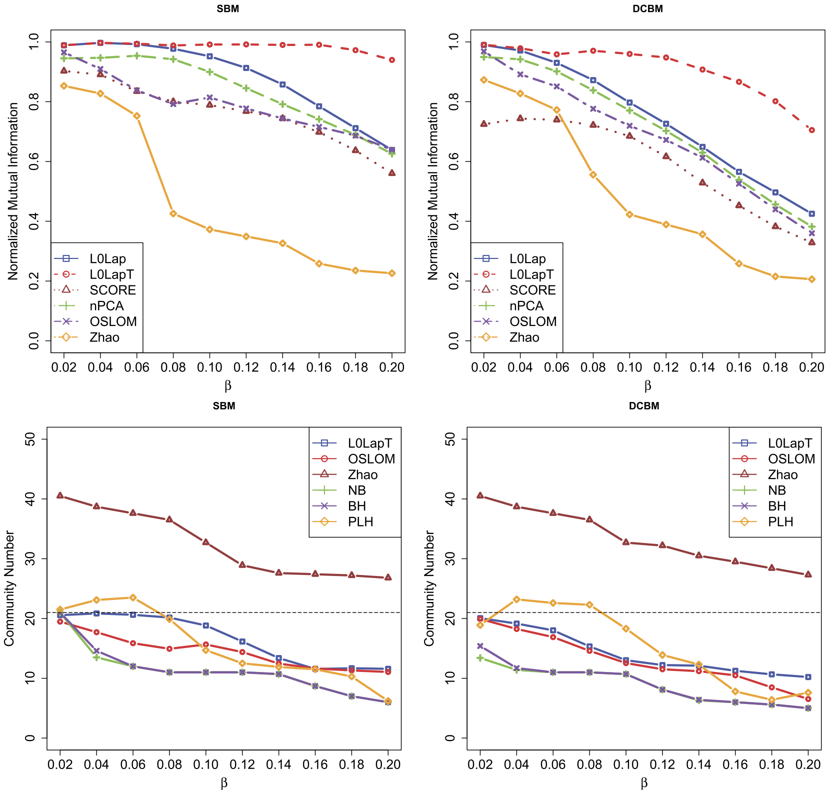

FIG 1. *Under the SBM and DCSBM, the mean NMI (top panel) and the mean detected community number (bottom panel) over 100 simulated networks with varying out-in-ratio parameter β. The degree parameter Λ is fixed as 50.*

We then fix $\beta = 0.1$ and vary Λ from 2 to 100 to compare different algorithms. The mean NMI of each algorithm is shown in Figure 2. Again, we see that our algorithm generally performs better than other algorithms. When Λ is very small, since OSLOM and Zhao often divide networks to many small connected subsets, they tend to have much larger NMIs than other methods and also tend to significantly overestimate the community number.

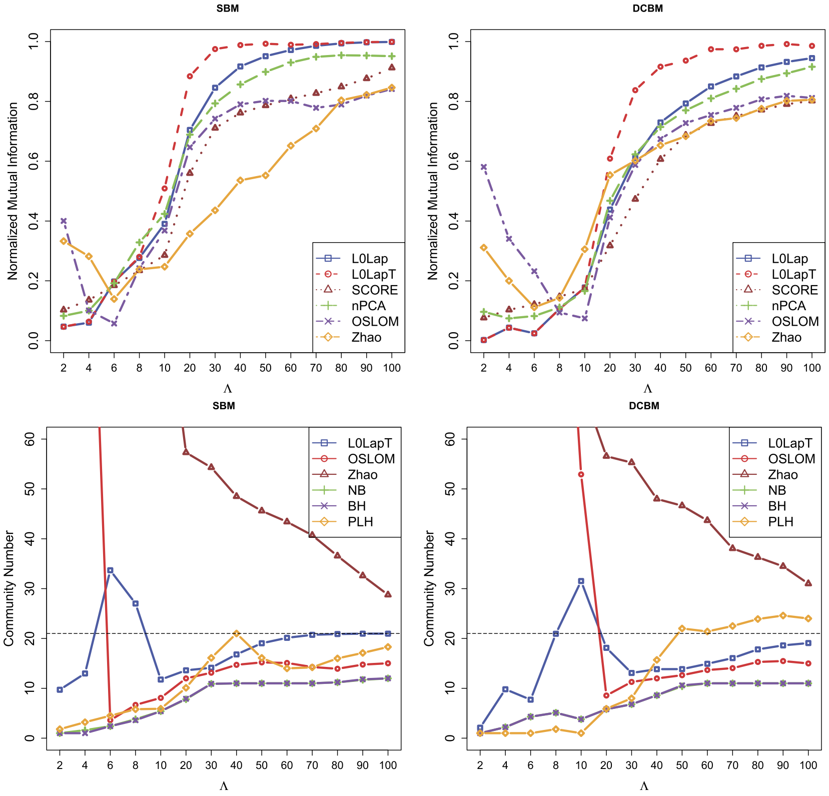

FIG 2. *Under the SBM and DCSBM, the mean NMI (top panel) and the mean detected community number (bottom panel) over 100 simulated networks with varying network degree* Λ. *The out-in-ratio* β *is fixed as 0.1.*

4.2. Simulation with outliers

In this section, we compare the performance of each method under the DCSBM with outliers. The simulated networks are similar to the simulated networks in Section 4.1 except that the 5 communities with 20 nodes are treated as outliers. The connecting probability between outliers is the same as the between-community connection probability. For SCORE and nPCA, we set the community number as 17 in this simulation (16 communities and 1 outlier community). To compute the reasonable NMIs, the outlier nodes are viewed as in the 17th true community. Figure 3 shows the NMIs and the number of detected communities of these algorithms. Because there are outliers, even when β is very small, there is still an nonignorable gap between the NMIs and its upper bound 1. However, after applying the permutation test, the NMI of L0Lap is significantly improved. Furthermore, the community number found by L0LapT is significantly better than all other methods.

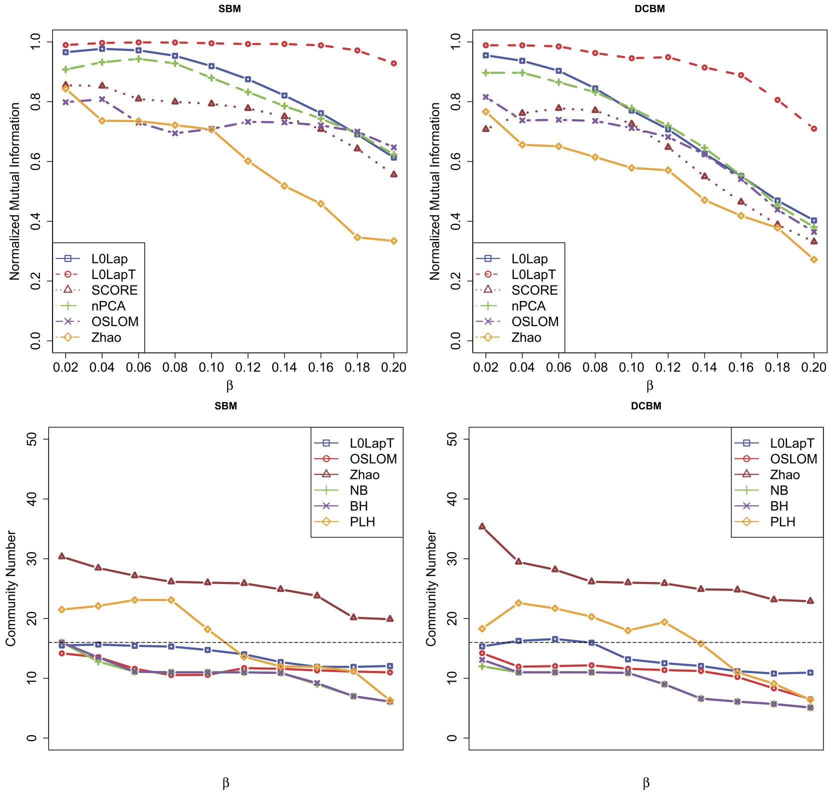

Fig 3. *Under the SBM and DCSBM with outliers, the mean NMI (top panel) and the mean detected community number (bottom panel) over 100 simulated networks with varying out-in-ratio parameter β. The degree parameter Λ is fixed as 50.*

5. Real data analysis

We consider two real data sets in this section, the college football network data [30] and the protein-protein network data in yeast [32].

5.1. College football data

The college football network data is the 2006 National Collegiate Athletic Association (NCAA) Football Bowl Subdivision (FBS) schedule [30]. The data set consists of 115 schools belonging to 11 conferences in FBS, 4 independent schools and 61 lower division schools. Schools within conferences play more often against each other, so the 11 conferences are 11 communities. The four independent schools are hubs. They play against many schools in different conferences but do not belong to any conferences. The 61 lower division schools connect loosely with other nodes and are outliers of the network. We apply all methods considered in the simulation study to this data set. The algorithms L0Lap, L0LapT, OSLOM, Zhao and PLH can automatically estimate the community number. For SCORE and nPCA, we provide them with the true community number 12, including 11 communities and one outlier community. The outlier community includes both the hub nodes and the outlier nodes. Table 1 shows the NMI and the detected community number (CN) of each algorithm. This clearly show that L0LapT have the largest NMI compared with other methods. PLH also works well. Its NMI is 0.929 and ranks the second best among all algorithms. In terms of outlier identification, although OSLOM and Zhao are designed to be able to identify outliers, OSLOM fails to report any outlier and the Zhao method assigns most outlier nodes to its largest detected community. In comparison, L0LapT identifies 80 nodes as outliers and 62 of them are true outliers.

Table 1

Performance comparison on the college football network data. CN is the detected or the provided community number. We set the community number as 12 for SCORE and nPCA.

	L0Lap	L0LapT	SCORE	nPCA	OSLOM	Zhao	PLH
NMI	0.856	0.985	0.674	0.640	0.681	0.651	0.929
CN	22	10	12	12	11	25	9

To look into more details of the detected communities of each algorithm, we examine the pairwise overlaps between detected communities with true communities. Specifically, given a detected community C_i^D and a true community C_j^T, we calculate an overlapping score between these two communities by $o_{ij} = |C_i^D \bigcap C_j^T|/|C_i^D \bigcup C_j^T|$. Thus, we get a matrix $O = (o_{ij})_{CN\times 12}$ for each algorithm, where CN is the detected community number. Figure 4 shows heat maps of these matrices for L0LapT, PLH, OSLOM and Zhao. Since Zhao extracted too many communities, we only consider the top 12 biggest communities. All communities identified by L0LapT are highly similar to or exactly the same as the true communities, which is shown in Figure 4. This demonstrates that L0LapT can give high quality communities. However, L0LapT fails to detect the community 11 and nodes in this community are filtered as outliers. For OSLOM, most of the diagonal overlapping scores are less than 0.71 and the largest overlapping score is only 0.86, showing that many detected communities by OSLOM contain substantial amount of nodes not belonging to these communities. PLH performs well for most communities, but members from true communities 4, 7 and 11 are mixed up. Zhao performs poorly in this data. Most of its detected communities are far away from true communities.

5.2. Protein-protein interaction data in yeast

In this section, we consider a protein-protein interaction (PPI) network data in yeast [32]. After removing isolated nodes, we get a network with 1,540 nodes and 7,123 edges. Different proteins often interact with each other to achieve one biological function. The communities of the PPI network should then represent

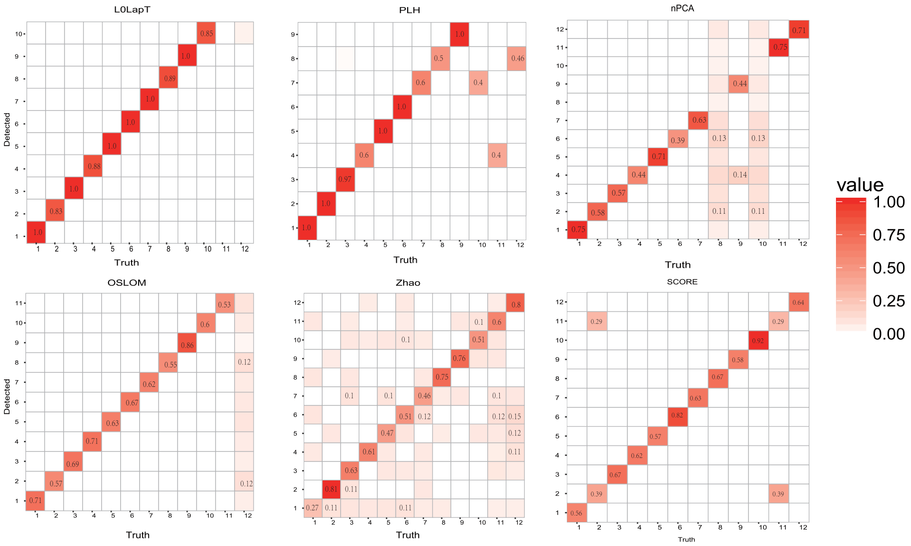

FIG 4. *Heatmap of the overlapping scores o_{ij} between the detected communities with the true communities for the college football network data. The true community 12 consists of outliers. The numbers in the figure are the overlapping scores o_{ij} with $o_{ij} > 0.1$.*

different cellular functions. We apply all methods in the simulation study to this PPI network. L0LapT finds 22 communities with their sizes ranging from 8 to 138. The Zhao method finds 15 communities ranging from 2 to 632 nodes. OSLOM finds 114 communities ranging from 3 to 103. For SCORE and nPCA, since the number of communities is unknown, we set the community number as 50, which roughly is the average number of communities detected by L0LapT, Zhao and OSLOM. The candidate community number set for PLH is set as all integers between 20 and 50. Finally, PLH find 29 communities ranging from 14 to 181. We further filter out communities with less than 5 nodes, since these are unlikely to be true communities.

There is no true community structure to evaluate the quality of detected communities. We instead use gene oncology (GO) enrichment analysis to compare different methods. We download yeast gene GO annotation database from http://www.yeastgenome.org/ and only focus only on GO terms with at least 10 annotated genes. For each community, we calculate a list of p-values with every GO term by Fisher's exact test. If the detected communities are biological meaningful, the communities should be highly significant with a number of GO terms. After $\log_{10}$ transformation of these p-values, define $\text{ratio}_t = |-\log_{10}\text{p-value} > t|/|-\log_{10}\text{p-value} > 0|$ for a threshold t. This ratio could be viewed as an indicator of biological relatedness of the detected communities. At the same cutoff t, larger ratio value should correspond to more biologically meaningful communities. The ratio curves of these methods are shown in Figure 5, left panel. We see that the curve of L0LapT is largely above other curves.

However, when t is large, it is hard to see the difference. Therefore, we further consider only p-values less than 0.1 and define $\text{ratio}_t^r = |\{-\log_{10}\text{p-value} > t\}|/|\{-\log_{10}\text{p-value} > 1\}|$ for any threshold $t \geq 1$. The new ratio curves are shown in Figure 5, right panel. We can now clearly see that L0LapT is always above other methods.

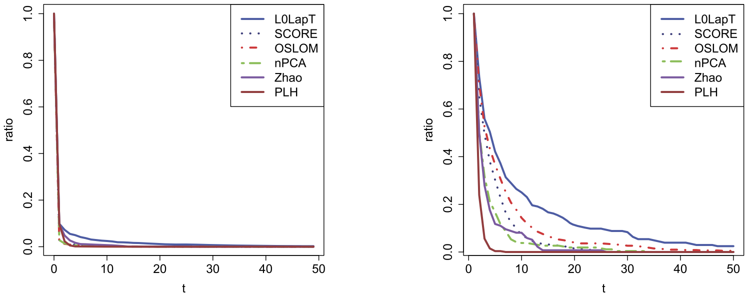

Fig 5. *GO enrichment analysis.*

6. Conclusion and discussion

In this paper, we propose a community detection method by maximizing a tightness criterion. This method does not require a known community number and it can detect communities in networks with outliers. We prove a consistency result for DCSBM with or without outliers. Simulation studies and real data applications show that the proposed method generally performs better than other available algorithms. One problem we found is that although the proposed method generally gives more accurate estimation of the community number, when networks contain more noise or when the network is too sparse, it still cannot give a very accurate estimate of community number. In addition, the statistical test used in this paper is based on permutation. Although simulation shows that this permutation works well in general in terms filtering false communities, we were not able to develop theoretical guarantees for this test.

The ADMM Algorithm 1 cannot guarantee a global maximum. Recently, a few paper showed that global optimizer could be identified by local adjustments [36]. These methods could be generalized to our optimization problem (2.4) and deserve future research. If the community number K is known, the tightness criterion (2.1) can be generalized to a partition of V. Assume $V = G_1 \bigcup ... \bigcup G_K$ is a partition of V, define

$$\psi(G_1, ..., G_K) = \sum_{i=1}^{K} \frac{W(G_i)}{V(G_i)}.$$

True communities should have a large $\psi(G_1, ..., G_K)$ and we may detect communities by maximizing $\psi(G_1, ..., G_K)$ over all partitions of V. Similarly, this optimization problem can be approximated by the Graph Laplician problem

$\max_{\mathbf{u}_1,\ldots,\mathbf{u}_K} \sum_{i=1}^{k} \mathbf{u}_i^T Q \mathbf{u}_i$ subject to $\|\mathbf{u}_i\|_2 = 1, \mathbf{u}_i \succeq 0,\ 1 \le i \le K$ and $\mathbf{u}_i^T \mathbf{u}_j = 0,\quad 1 \le i \ne j \le K$. Simulation analyses show that this formulation results in accurate communities. However, we have not found an efficient algorithm and future research is needed in this direction.

7. Appendix

In this section, we give proofs of our theoretical results. Before proving the main theorem, we first give some lemmas.

Lemma 7.1. *Under the assumptions of DCSBM, we have*

$$\mathbb{E}\left(W(S)|\mathbf{c}\right) = \sum_{k=1}^{K} nr_k^d(S) \left(\sum_{l=1}^{K} nr_l^d(S) p_{kl} \right) \quad \textit{and}$$
$$\mathbb{E}(V(S)|\mathbf{c}) = \sum_{k=1}^{K} nr_k^d(S) \left(\sum_{l=1}^{K} n\hat{\pi}_l^d p_{kl} \right)$$

Proof. Under the assumptions of DCSBM, we have

$$\mathbb{E}(A_{ij}|c_i = k, c_j = l) = \mathbb{E}(\theta_i|c_i = k)\mathbb{E}(\theta_j|c_j = l)p_{kl} = \frac{\pi_k^d}{\pi_k}\frac{\pi_l^d}{\pi_l}p_{kl}.$$

So we have

$$\begin{aligned} \mathbb{E}(W(S)|\mathbf{c}) &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i\in S_k, j\in S_l} \mathbb{E}(A_{ij}|\mathbf{c}) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i\in S_k, j\in S_l} \frac{\pi_k^d}{\pi_k}\frac{\pi_l^d}{\pi_l}p_{kl} = \sum_{k=1}^{K} nr_k^d(S)(\sum_{l=1}^{K} nr_l^d(S)p_{kl}), \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(V(S)|\mathbf{c}) &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i\in S_k, j\in G_l} \mathbb{E}(A_{ij}|\mathbf{c}) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i\in S_k, j\in G_l} \frac{\pi_k^d}{\pi_k}\frac{\pi_l^d}{\pi_l}p_{kl} = \sum_{k=1}^{K} nr_k^d(S)(\sum_{l=1}^{K} n\hat{\pi}_l^d p_{kl}). \end{aligned}$$ □

We need Chernoff's inequality [9] and Hoeffding's inequality [11] to prove Theorem 3.1.

Lemma 7.2. *(Chernoff's inequality) Let $X_1, ..., X_n$ be independent random variables with*

$$\mathbb{P}(X_i = 1) = p_i,\ \mathbb{P}(X_i = 0) = 1 - p_i.$$

Then the sum $X = \sum_{i=1}^n X_i$ has expectation $\mathbb{E}(X) = \sum_{i=1}^n p_i$, and we have

$$\mathbb{P}\left(X < \mathbb{E}\left(X\right) - \lambda\right) < \exp\big\{-2^{-1}\lambda^2/\mathbb{E}(X)\big\},$$

$$\mathbb{P}\left(X > \mathbb{E}\left(X\right) + \lambda\right) < \exp\big\{-2^{-1}\lambda^2/(\mathbb{E}(X) + \lambda/3)\big\}.$$

Lemma 7.3. *(Hoeffding's inequality) Let $X_1, ..., X_n$ be independent random variables and X_i's are strictly bounded by the intervals $[a_i, b_i]$. We define the empirical mean of these variables by $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, then we have*

$$\mathbb{P}\left(\left|\bar{X} - \mathbb{E}(\bar{X})\right| > t\right) \leq 2\exp\left\{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

Lemma 7.4. *Define $\hat{\psi}(S) = \mathbb{E}(W(S)|\mathbf{c})/\mathbb{E}(V(S)|\mathbf{c}) - \eta|S|$. Under the assumptions of DCSBM, we have*

$$\max_{S\in\Gamma_\delta}\left|\psi(S) - \hat{\psi}(S)\right| \lesssim n^{\delta-\alpha}$$

with probability at least $1 - 2^{n+2}/n^n$ when n is sufficiently large.

Proof. By Lemma 7.1 and the condition $p^- \gtrsim \log n/n^{1-2\alpha}$, we have

$$\begin{aligned}\mathbb{E}(W(S)|\mathbf{c}) &= \sum_{k=1}^K nr_k^d(S)(\sum_{l=1}^K nr_l^d(S)p_{kl}) \\ &\geq p^- \sum_{k=1}^K n^2(r_k^d(S))^2 \gtrsim n^{1+2\alpha}\log n\sum_{k=1}^K (r_k^d(S))^2. \qquad (7.1)\end{aligned}$$

Since $\sum_{k=1}^K r_k^d(S) = \sum_{k=1}^K r_k(S)\pi_k^d/\pi_k$ and $0 < h_1 \leq \pi_k^d/\pi_k \leq h_M$, we have $\sum_{k=1}^K (r_k^d(S))^2 \geq (\sum_{k=1}^K r_k^d(S))^2/K \geq h_1^2|S|^2/(Kn^2)$ by the Cauchy-Schwarz inequality. Then we have $\mathbb{E}(W(S)|\mathbf{c}) \gtrsim n^{1+2(\alpha-\delta)}\log n$ if $S \in \Gamma_\delta$. Let $\lambda = 2\sqrt{n\log n\mathbb{E}(W(S)|\mathbf{c})}$ and by Chernoff's inequality, we have

$$\mathbb{P}\left(W(S) - \mathbb{E}(W(S)|\mathbf{c}) < -\lambda\right) < n^{-n}.$$

Since $\mathbb{E}(W(S)|\mathbf{c}) \gtrsim n^{1+2(\alpha-\delta)}\log n$, we have $\lambda/3 < \mathbb{E}(W(S)|\mathbf{c})$ with sufficiently large n and thus

$$\mathbb{P}\left(W(S) - \mathbb{E}(W(S)|\mathbf{c}) > \lambda\right) < n^{-n}.$$

So we have

$$\mathbb{P}\left(|W(S) - \mathbb{E}(W(S)|\mathbf{c})| > \lambda\right) < 2n^{-n}.$$

For $V(S)$, we have

$$\mathbb{E}(V(S)|\mathbf{c}) \geq \mathbb{E}(W(S)|\mathbf{c}) \gtrsim n^{1+2(\alpha-\delta)}\log n$$

Similarly, let $\tilde{\lambda} = 2\sqrt{n\log n\mathbb{E}(V(S)|\mathbf{c})}$, and we have

$$\mathbb{P}\left(|V(S) - \mathbb{E}(V(S)|\mathbf{c})| > \tilde{\lambda}\right) < 2n^{-n}.$$

In addition, we have

$$\frac{\lambda}{\mathbb{E}(W(S)|\mathbf{c})} \lesssim \frac{1}{n^{\alpha-\delta}} \text{ and } \frac{\tilde{\lambda}}{\mathbb{E}(V(S)|\mathbf{c})} \lesssim \frac{1}{n^{\alpha-\delta}}.$$

Thus, with probability at least $1 - 4/n^n$, we have

$$\begin{aligned}
&\left|\frac{W(S)}{V(S)} - \frac{\mathbb{E}(W(S)|\mathbf{c})}{\mathbb{E}(V(S)|\mathbf{c})}\right| \\
&\leq \max\left\{\left|\frac{\mathbb{E}(W(S)|\mathbf{c}) - \lambda}{\mathbb{E}(V(S)|\mathbf{c}) + \tilde{\lambda}} - \frac{\mathbb{E}(W(S)|\mathbf{c})}{\mathbb{E}(V(S)|\mathbf{c})}\right|, \left|\frac{\mathbb{E}(W(S)|\mathbf{c}) + \lambda}{\mathbb{E}(V(S)|\mathbf{c}) - \tilde{\lambda}} - \frac{\mathbb{E}(W(S)|\mathbf{c})}{\mathbb{E}(V(S)|\mathbf{c})}\right|\right\} \\
&= \max\left\{\left|\frac{\mathbb{E}(W(S)|\mathbf{c})\tilde{\lambda} + \mathbb{E}(V(S)|\mathbf{c})\lambda}{\mathbb{E}(V(S)|\mathbf{c})(\mathbb{E}(V(S)|\mathbf{c}) + \tilde{\lambda})}\right|, \left|\frac{\mathbb{E}(W(S)|\mathbf{c})\tilde{\lambda} + \mathbb{E}(V(S)|\mathbf{c})\lambda}{\mathbb{E}(V(S)|\mathbf{c})(\mathbb{E}(V(S)|\mathbf{c}) - \tilde{\lambda})}\right|\right\} \\
&\leq \left|\frac{\lambda}{\mathbb{E}(V(S)|\mathbf{c}) - \tilde{\lambda}}\right| + \left|\frac{\tilde{\lambda}}{\mathbb{E}(V(S)|\mathbf{c}) - \tilde{\lambda}}\right| \lesssim \frac{1}{n^{\alpha-\delta}}.
\end{aligned}$$

Therefore, with probability at least $1 - 2^{n+2}/n^n$ we have

$$\max_{S\in\Gamma_\delta}\left|\psi(S) - \hat{\psi}(S)\right| \lesssim \frac{1}{n^{\alpha-\delta}},$$

when n is sufficiently large. □

Lemma 7.5. *For DCSBM, with probability at least* $1 - 2Kn^{-2}$, *we have*

$$|\hat{\rho}_k^d - \rho_k^d| \lesssim \frac{1}{n^{\alpha-\delta}},$$

for all $1 \leq k \leq K$.

Proof. By definition, we have

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}\{c_i = k\}.$$

Since $\mathbb{E}(\hat{\pi}_k) = \pi_k$ and $\mathbf{I}\{c_i = k\}$ are strictly bounded by the intervals $[0, 1]$, we have

$$\mathbb{P}\left(|\hat{\pi}_k - \pi_k| > \sqrt{\frac{\log n}{n}}\right) \leq \frac{2}{n^2},$$

by Hoeffding's inequality.

Since $\hat{\pi}_k^d = \pi_k^d\hat{\pi}_k/\pi_k$ and $\pi_k^d/\pi_k \leq h_M$, we have

$$\mathbb{P}\left(\left|\hat{\pi}_k^d - \pi_k^d\right| > h_M\sqrt{\frac{\log n}{n}}\right) \leq \frac{2}{n^2},$$

by the inequality above. Therefore, with probability at least $1-2Kn^{-2}$ we have

$$|\hat{\rho}_k^d - \rho_k^d| = \left|\frac{p_{kk}}{\sum_{l=1}^K \pi_l^d p_{kl}} - \frac{p_{kk}}{\sum_{l=1}^K \hat{\pi}_l^d p_{kl}}\right| = p_{kk}\left|\frac{\sum_{l=1}^K (\hat{\pi}_l^d - \pi_l^d) p_{kl}}{(\sum_{l=1}^K \pi_l^d p_{kl})(\sum_{l=1}^K \hat{\pi}_l^d p_{kl})}\right|$$
$$\lesssim \sqrt{\frac{\log n}{n}}\frac{1}{(\pi^-)^2} \lesssim \frac{\sqrt{\log n}}{n^{1/2-\delta}} \lesssim \frac{1}{n^{\alpha-\delta}}. \qquad \square$$

Lemma 7.6. *Assume real numbers* $0 < x_k, y_k, z_{kl} \le 1$ *satisfy* $0 < C_1 \le x_k/y_k \le C_2$ *for all* $1 \le k \ne l \le K$. *Define*

$$f(t_1, ..., t_K) = \frac{\sum_{k=1}^K t_k(t_k x_k + \sum_{l\ne k} t_l z_{kl})}{\sum_{k=1}^K t_k y_k}, \tag{7.2}$$

where $t_k \ge 0$ *and* $\sum_{k=1}^K t_k = 1$. *If* $x_1/y_1 > \max_{2\le k\le K} x_k/y_k$ *and* $\min_{1\le k\le K} x_k > max_{k\ne l} z_{kl}$, *we have*

(1) $f(t_1, ..., t_K) \le f(1, 0, ..., 0) = \frac{x_1}{y_1}$,
(2) For any $0 < t < 1$,

$$f(1,0,...,0) - \max_{t_1 \le 1-t} f(t_1, ..., t_K) \ge \frac{1}{2}\left(\frac{x_1}{y_1} - \max_{2\le k\le K}\frac{x_k}{y_k}\right) t.$$

Proof. (1) Since $x_1/y_1 > \max_{2\le k\le K} x_k/y_k$ and $\min_{1\le k\le K} x_k > max_{k\ne l} z_{kl}$,

$$f(t_1, ..., t_K) \le \frac{\sum_{k=1}^K t_k x_k}{\sum_{k=1}^K t_k y_k} \le \frac{x_1}{y_1} = f(1, 0, ..., 0).$$

(2) Since $f(t_1, ..., t_K)$ is continuous and $\{(t_1, ..., t_K)|t_1 \le 1-t\}$ is a close set, $f(t_1, ..., t_K)$ can achieve its upper bound on $\{(t_1, ..., t_K)|t_1 \le 1-t\}$. Suppose that $f(t_1, ..., t_K)$ achieves its upper bound at $(t_1^*, ..., t_K^*)$ and define $\bar{x}, \bar{y}$ such that $(1-t_1^*)\bar{x} = \sum_{k=2}^K t_k^*(t_k^* x_k + \sum_{l\ne k} t_l^* z_{kl})$, $(1-t_1^*)\bar{y} = \sum_{k=2}^K t_k^* y_k$. We have $\bar{x}/\bar{y} \le \max_{2\le k\le K} x_k/y_k$ since $\min_{1\le k\le K} x_k > max_{k\ne l} z_{kl}$. Let $z^+ = max_{k\ne l} z_{kl}$, then we have

$$\begin{aligned} & f(1,0,...,0) - \max_{t_1\le 1-t} f(t_1, ..., t_K) \\ & \quad \ge \frac{x_1}{y_1} - \frac{t_1^*(t_1^* x_1 + (1-t_1^*)z^+) + (1-t_1^*)\bar{x}}{t_1^* y_1 + (1-t_1^*)\bar{y}} \\ & \quad = \frac{(1-t_1^*)t_1^* y_1(x_1 - z^+)}{y_1(t_1^* y_1 + (1-t_1^*)\bar{y})} + \frac{(1-t_1^*)(x_1\bar{y} - \bar{x}y_1)}{y_1(t_1^* y_1 + (1-t_1^*)\bar{y})}. \end{aligned}$$

Case I: $y_1 = \min_{1\le k\le K} y_k$.
We have $\bar{y}/\left[t_1^* y_1 + (1-t_1^*)\bar{y}\right] \ge 1$, and thus

$$\frac{(1-t_1^*)(x_1\bar{y} - \bar{x}y_1)}{y_1(t_1^* y_1 + (1-t_1^*)\bar{y})} = (1-t_1^*)(\frac{x_1}{y_1} - \frac{\bar{x}}{\bar{y}})\frac{\bar{y}}{t_1^* y_1 + (1-t_1^*)\bar{y}}$$

$$\geq t(\frac{x_1}{y_1} - \max_{2\leq k\leq K} \frac{x_k}{y_k}).$$

Case II: $y_1 > \min_{1\leq k\leq K} y_k$.
There exists $i \neq 1$ such that $y_i < y_1$, then we have $z^+/y_1 < x_i/y_i$.
If $\bar{y}(1-t_1^*) \geq y_1 t_1^*$, we have

$$\frac{\bar{y}}{t_1^* y_1 + (1-t_1^*)\bar{y}} \geq \frac{1}{2(1-t_1^*)},$$

and

$$\begin{aligned}\frac{(1-t_1^*)(x_1\bar{y} - \bar{x}y_1)}{y_1(t_1^*y_1 + (1-t_1^*)\bar{y})} &= (1-t_1^*)(\frac{x_1}{y_1} - \frac{\bar{x}}{\bar{y}})\frac{\bar{y}}{t_1^*y_1 + (1-t_1^*)\bar{y}} \\ &\geq \frac{1}{2}(\frac{x_1}{y_1} - \max_{2\leq k\leq K} \frac{x_k}{y_k}).\end{aligned}$$

If $\bar{y}(1-t_1^*) \leq y_1 t_1^*$, we have

$$\frac{t_1^* y_1}{t_1^* y_1 + (1-t_1^*)\bar{y}} \geq \frac{1}{2}$$

and $z^+/y_1 \leq \max_{2\leq k\leq K} x_k/y_k$, then we have

$$\begin{aligned}\frac{(1-t_1^*)t_1^*y_1(x_1 - z^+)}{y_1(t_1^*y_1 + (1-t_1^*)\bar{y})} &= (1-t_1^*)(\frac{x_1}{y_1} - \frac{z^+}{y_1})\frac{t_1^*y_1}{t_1^*y_1 + (1-t_1^*)\bar{y}} \\ &\geq \frac{1}{2}t(\frac{x_1}{y_1} - \max_{2\leq k\leq K} \frac{x_k}{y_k}).\end{aligned}$$

So we have

$$f(1,0,...,0) - \max_{t_1\leq 1-t} f(t_1,...,t_K) \geq \frac{1}{2}t\left(\frac{x_1}{y_1} - \max_{2\leq k\leq K}\frac{x_k}{y_k}\right). \qquad \square$$

Based on the lemmas given previously, we give the proof of Theorem 3.1.

Proof of Theorem 3.1. Based on Lemma 7.4, with probability at least $1-2^{n+2}/n^n$, we have

$$\max_{S\in\Gamma_\delta}\left|\psi(S) - \hat{\psi}(S)\right| \lesssim \frac{1}{n^{\alpha-\delta}}.$$

Let $t_k(S) = r_k^d(S)/r^d(S)$ for $1 \leq k \leq K$,

$$\begin{aligned}\hat{\psi}(S) &= \frac{\sum_{k=1}^K nr_k^d(S)(\sum_{l=1}^K nr_l^d(S)p_{kl})}{\sum_{k=1}^K nr_k^d(S)(\sum_{l=1}^K n\hat{\pi}_l^d p_{kl})} - n\eta\sum_{k=1}^K \frac{\pi_k}{\pi_k^d} r_k^d(S) \\ &= r^d(S)\left(\frac{\sum_{k=1}^K t_k(S)(t_k(S)p_{kk} + \sum_{l\neq k} t_l(S)p_{kl})}{\sum_{k=1}^K t_k(S)(\sum_{l=1}^K \hat{\pi}_l^d p_{kl})} - n\eta\sum_{k=1}^K \frac{\pi_k}{\pi_k^d} t_k(S)\right) \\ &= r^d(S)\left(f(t_1(S),...,t_K(S)) - n\eta\sum_{k=1}^K \frac{\pi_k}{\pi_k^d}t_k(S)\right).\end{aligned}$$

Note that if $S = G_1$, $t_1(G_1) = 1$, $t_k(G_1) = 0$ $(k = 2, \cdots, K)$ and

$$\hat{\psi}(G_1) = \hat{\pi}_1^d \left(f(1, 0, \cdots, 0) - n\eta \frac{\pi_1}{\pi_1^d} \right).$$

Based on Lemma 7.5, with probability at least $1 - 2Kn^{-2}$ we have

$$\hat{\rho}_1^d - \max_{2 \leq k \leq K} \hat{\rho}_k^d \gtrsim \frac{1}{n^\tau}.$$

Then, with probability at least $1 - 2Kn^{-2}$ we have

$$\begin{aligned} & f(1, 0, ..., 0) - \max_{t_1 \leq 1 - 1/n^{\gamma - \tau}} f(t_1, t_2, ..., t_K) \\ & \gtrsim \frac{1}{n^{\gamma - \tau}} (\hat{\rho}_1^d - \max_{2 \leq k \leq K} \hat{\rho}_k^d) \gtrsim \frac{1}{n^{\gamma - \tau}} \frac{1}{n^\tau} \gtrsim \frac{1}{n^\gamma}, \end{aligned} \tag{7.3}$$

and $f(t_1, ..., t_K) \leq f(1, 0, ..., 0) = \hat{\rho}_1^d$ by Lemma 7.6. From (7.3), it is easy to see that for a constant C, we can choose η satisfying the inequality (3.1) with probability at least $1 - 2Kn^{-2}$.

Since $\max_{2 \leq k \leq K} \pi_k^d / \pi_k \leq \pi_1^d / \pi_1 \leq h_M$, Using the inequality in the condition, we have with sufficiently large n,

$$f(t_1(S), ..., t_K(S)) - n\eta \sum_{k=1}^K \frac{\pi_k}{\pi_k^d} t_k(S) \begin{cases} > C/(h_M n^\gamma), & \text{if } \ t_1(S) = 1; \\ < -C/(h_M n^\gamma), & \text{if } t_1(S) \leq 1 - 1/n^{\gamma - \tau}. \end{cases}$$

So we have

$$\hat{\psi}(S) \begin{cases} > r^d(S) C/(h_M n^\gamma), & \text{if } \ t_1 = 1; \\ < -r^d(S) C/(h_M n^\gamma), & \text{if } \ t_1 \leq 1 - 1/n^{\gamma - \tau}. \end{cases}$$

with sufficiently large n. To maximize $\hat{\psi}(S)$, t_1 must be bigger than $1 - 1/n^{\gamma - \tau}$. If $r^d(S) > \hat{\pi}_1^d + \hat{\pi}_1^d / (n^{\gamma - \tau} - 1)$, then $t_1 \leq \hat{\pi}_1^d / r^d(S) < 1 - 1/n^{\gamma - \tau}$. So we must have $r^d(S) \leq \hat{\pi}_1^d + \hat{\pi}_1^d / (n^{\gamma - \tau} - 1)$. If $r^d(S) \leq \hat{\pi}_1^d - \hat{\pi}_1^d \log n / n^{\alpha - 2\delta - \gamma}$ and $\hat{\pi}_1^d \gtrsim n^{-\delta}$, by Lemma 7.6 we have

$$\begin{aligned} \hat{\psi}(G_1) - \hat{\psi}(S) & \geq (\hat{\pi}_1^d - r^d(S)) \left(f(1, 0, ..., 0) - n\eta \frac{\pi_1}{\pi_1^d} \right) \\ & \geq \hat{\pi}_1^d \frac{\log n}{n^{\alpha - 2\delta - \gamma}} \left(f(1, 0, ..., 0) - n\eta \frac{\pi_1}{\pi_1^d} \right) \gtrsim \frac{\log n}{n^{\alpha - \delta}}. \end{aligned}$$

Since $r^d(S) = \sum_{k=1}^K r_k(S) \pi_k^d / \pi_k \gtrsim 1/n^\delta$ and $\pi_1^d \gtrsim 1/n^\delta$, then with probability at least $1 - 2Kn^{-2}$ we have

$$\hat{\psi}(G_1) - \hat{\psi}(S) \begin{cases} \gtrsim \frac{\log n}{n^{\alpha - \delta}}, & \text{if } \ r^d(S) \leq \hat{\pi}_1^d - \frac{\hat{\pi}_1^d \log n}{n^{\alpha - 2\delta - \gamma}} \\ \gtrsim \frac{C}{n^{\gamma + \delta}}, & \text{if } \ t_1(S) \leq 1 - 1/n^{\gamma - \tau} \end{cases}$$

So with probability at least $1 - 2Kn^{-2} - 2^{n+2}/n^n$, $\psi(S) < \psi(G_1)$ for all S satisfying $r^d(S) \le \hat{\pi}_1^d - \hat{\pi}_1^d \log n/n^{\alpha-2\delta-\gamma}$ or $t_1(S) \le 1 - 1/n^{\gamma-\tau}$, which implies that $\psi(S)$ maximizes when $t_1(S) \ge 1 - 1/n^{\gamma-\tau}$ *and* $\hat{\pi}_1^d - \hat{\pi}_1^d \log n/n^{\alpha-2\delta-\gamma} \le r^d(S) \le \hat{\pi}_1^d + \hat{\pi}_1^d/(n^{\gamma-\tau} - 1)$ with probability at least $1-2Kn^{-2}-2^{n+2}/n^n$. Since $t_1(S) = r_1^d(S)/r^d(S)$, we therefore have $\hat{\pi}_1^d - \hat{\pi}_1^d \log n/n^{\alpha-2\delta-\gamma} \le r_1^d(S)/t_1(S) \le \hat{\pi}_1^d + \hat{\pi}_1^d/(n^{\gamma-\tau} - 1)$, and hence $(1 - 1/n^{\gamma-\tau})\hat{\pi}_1 \left(1 - \log n/n^{\alpha-2\delta-\gamma}\right) \le r_1(S) \le \hat{\pi}_1 t_1(S) n^{\gamma-\tau}/(n^{\gamma-\tau} - 1)$. From $t_1(S) \le 1-1/n^{\gamma-\tau}$, we get $h_1 h_M^{-1}(r(S)-r_1(S))/r(S) \le (r^d(S) - r_1^d(S))/r^d(S) \le 1/n^{\gamma-\tau}$ and $r_1(S)/r(S) \ge 1 - h_M h_1^{-1}/n^{\gamma-\tau}$. Note that $r_1(S)/\hat{\pi}_1 = |S \bigcap G_1|/|G_1|$ and $r_1(S)/r(S) = |S \bigcap G_1|/|S|$. Therefore, with probability at least $1 - 2Kn^{-2} - 2^{n+2}/n^n$, we have

$$\frac{|S\Delta G_1|}{|S \bigcup G_1|} \le 2h_M h_1^{-1}/n^{\gamma-\tau} + \log n/n^{\alpha-2\delta-\gamma}. \tag{7.4}$$

□

The proof of Theorem 3.2 is very similar to the proof of Theorem 3.1 and we omit it. The only difference is that we have to pay attention to the ourlier community. For example, in the proof of Lemma 7.4, the inequality (7.1) becomes $\mathbb{E}(W(S)|\mathbf{c}) \gtrsim n^{1+2\alpha} \log n \sum_{k=1}^{K-1} (r_k^d(S))^2$. Then, using the condition $|G_K|^2/K = o(n^{2-2\delta})$, we can also get $\mathbb{E}(W(S)|\mathbf{c}) \gtrsim n^{1+2(\alpha-\delta)} \log n$ and hence the conclusion of Lemma 7.4 also holds for DCSBM with outliers.

Acknowledgments

The authors gratefully acknowledge Bin Yu for her insightful suggestions. We thank Yunpeng Zhao, Can Le for sharing their codes. The authors also gratefully acknowledge *the National Key Basic Research Project of China (2015CB856000 and 2016YFC0207705), the National Natural Science Foundation of China (11471022, 71532001) and the Recruitment Program of Global Youth Experts of China.*

References

[1] Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, **41**, 2097–2122. MR3127859

[2] Balakrishnan, S., Xu, M., Krishnamurthy, A., and Singh, A. (2011). Noise thresholds for spectral clustering. *In Advances in Neural Information Processing Systems*, 954–962.

[3] Bickel, P. J., and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, **106**, 21068–21073.

[4] Bickel, P. J., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics,* **41**, 1922–1943. MR3127853

[5] Cai, T., and Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics,* **43**, 1027–1059. MR3346696

[6] Chaudhuri, K., Graham, F. C., and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research,* **35**, 1–23.

[7] Choi, D., Wolfe, P., and Airoldi, E. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika,* **99**, 273–284. MR2931253

[8] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E,* **84**, 66–106.

[9] Füredi, Z., and Komlós, J. (1981). The eigenvalues of random symmetric matrices. *Combinatorica,* **1**, 233–241. MR0637828

[10] Hagen, L., and Kahng, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems,* **11**, 1074–1085.

[11] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association,* **58**, 13–30. MR0144363

[12] Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks,* **5**, 109–137. MR0718088

[13] Jin, J. (2015). Fast community detection by SCORE. *The Annals of Statistics,* **43**, 57–89. MR3285600

[14] Joseph, A., and Yu, B. (2016). Impact of regularization on spectral clustering. *The Annals of Statistics,* **44**, 1765–1791. MR3519940

[15] Karrer, B., and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E,* **83**, 16–107. MR2788206

[16] Kim, S., and Shi, T. (2012). Scalable spectral algorithms for community detection in directed networks. *arXiv preprint arXiv:1211.6807*.

[17] Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLOS ONE,* **6**, e18961.

[18] Le, C. M., and Levina, E. (2015). Estimating the number of communities in networks by spectral methods. *arXiv preprint arXiv:1507.00827*.

[19] Lei, J., and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics,* **43**, 215–237. MR3285605

[20] Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. *In Proceedings of the 17th international conference on World Wide Web*, 695–704. MR2736090

[21] Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics,* **4**, 715–742. MR2758646

[22] Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences,* **101**, 5200–5205.

[23] Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E,* **69**, 66–133. MR1975193

[24] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences,* **103**, 8577–8582.

[25] Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E,* **69**, 26–113. MR2282139

[26] Nowicki, K., and Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association,* **96**, 1077–1087. MR1947255

[27] Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics,* **39**, 1878–1915. MR2893856

[28] Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **22**, 888–905.

[29] Wang, Y. R., Bickel, P. J., et al. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics,* **45**, 500–528. MR3650391

[30] Xu, X., Yuruk, N., Feng, Z., and Schweiger, T. A. (2007). Scan: a structural clustering algorithm for networks. *In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 824–833.

[31] Yao, Y. Y. (2003). Information-theoretic measures for knowledge discovery and data mining. *In Entropy Measures, Maximum Entropy Principle and Emerging Applications*, 115–136.

[32] Yu, H., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science,* **322**, 104–110.

[33] Zhao, Y., Levina, E., and Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences,* **108**, 7321–7326.

[34] Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics,* **40**, 2266–2292. MR3059083

[35] Noceda, J., and Wright, S. (2006). Numerical Optimization. Berlin, New York: Springer-Verlag. MR2244940

[36] Byokov, Y., Veksler, O. and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **23**, 1222–1239.