Journal of Machine Learning Research 18 (2017) 1-45 Submitted 5/16; Revised 2/17; Published 6/17

Achieving Optimal Misclassification Proportion in Stochastic Block Models

Chao Gao CHAOGAO@GALTON.UCHICAGO.EDU
University of Chicago

Zongming Ma ZONGMING@WHARTON.UPENN.EDU
University of Pennsylvania

Anderson Y. Zhang YE.ZHANG@YALE.EDU
Harrison H. Zhou HUIBIN.ZHOU@YALE.EDU
Yale University

Editor: Sara van de Geer

Abstract

Community detection is a fundamental statistical problem in network data analysis. In this paper, we present a polynomial time two-stage method that provably achieves optimal statistical performance in misclassification proportion for stochastic block model under weak regularity conditions. Our two-stage procedure consists of a refinement stage motivated by penalized local maximum likelihood estimation. This stage can take a wide range of weakly consistent community detection procedures as its initializer, to which it applies and outputs a community assignment that achieves optimal misclassification proportion with high probability. The theoretical property is confirmed by simulated examples.

Keywords: Clustering, Community detection, Minimax rates, Network analysis, Spectral clustering.

1. Introduction

Network data analysis (Wasserman, 1994; Goldenberg et al., 2010) has become an important topic in statistics. In fields such as physics, computer science, social science and biology, one observes a network among a large number of subjects of interest such as particles, computers, people, etc. The observed network can be modeled as an instance of a random graph and the goal is to infer structures of the underlying generating process. A structure of particular interest is *community*: there is a partition of the graph nodes in some suitable sense so that each node belongs to a community. Starting with the proposal of a series of methodologies (Girvan and Newman, 2002; Newman and Leicht, 2007; Handcock et al., 2007; Karrer and Newman, 2011), we have seen a large literature devoted to algorithmic solutions to uncovering community structure. Great advances have also been made in recent years on the theoretical understanding of the problem in terms of statistical consistency and thresholds for detection and exact recoveries. See, for instance, Bickel and Chen (2009); Decelle et al. (2011); Zhao et al. (2012); Mossel et al. (2012, 2013b); Massoulié (2014); Abbe et al. (2014); Mossel et al. (2014); Hajek et al. (2014), among others. The major goal of the present paper is to propose a computationally feasible algorithm for community detection in stochastic block models with adaptive minimax optimal performance.

To describe network data with community structure, we focus on the stochastic block model (SBM) proposed by Holland et al. (1983). Let $A \in \{0,1\}^{n\times n}$ be the symmetric adjacency matrix of an undirected random graph generated according to an SBM with k communities. The diagonal entries of A are all zeros and each $A_{uv} = A_{vu}$ for $u > v$ is an independent Bernoulli random variable with mean $P_{uv} = B_{\sigma(u)\sigma(v)}$ for some symmetric connectivity matrix $B \in [0,1]^{k\times k}$ and some label function $\sigma : [n] \to [k]$. Here and after, for any positive integer m, $[m] = \{1, \ldots, m\}$. In other word, if the u^{th} node and the v^{th} node belong to the i^{th} and the j^{th} community respectively, then $\sigma(u) = i$, $\sigma(v) = j$ and there is an edge connecting u and v with probability B_{ij}. We define a and b through $\min_i B_{ii} = a/n$ and $\max_{i\neq j} B_{ij} = b/n$. Community detection then refers to the problem of estimating the label function σ subject to a permutation of the community labels $\{1, \ldots, k\}$. A natural loss function for such an estimation problem is the proportion of wrong labels (subject to a permutation of the label set $[k]$), which we refer to as misclassification proportion from here on.

Literature review The field of community detection in SBMs has been growing fast. Results on fundamental limits and various algorithms for achieving them have been obtained in the literature. We first review the most relevant results prior to our work.

1. *Detection.* In ground breaking works by Mossel et al. (2012, 2013b) and Massoulié (2014), the authors established sharp threshold for the regimes in which it is possible and impossible to achieve a misclassification proportion strictly less than $\frac{1}{2}$ (so that it is better than random guess) when $k = 2$ and both communities are of the same size. This solved the conjecture in Decelle et al. (2011) that was only justified in physics rigor. The necessary and sufficient condition for doing better than random guess is $(a-b)^2 > 2(a+b)$.

2. *Weak consistency.* In the current context, weak consistency means recovering all but a vanishing proportion of the community labels. As was shown in Mossel et al. (2014), the necessary and sufficient condition for achieving weak consistency is $(\sqrt{a}-\sqrt{b})^2 \to \infty$ in the equal-sized two community setting. Conditions of weak consistency in general SBMs were obtained and discussed in Zhang and Zhou (2015); Yun and Proutiere (2014a); Abbe and Sandon (2015a).

3. *Strong consistency.* Abbe et al. (2014); Mossel et al. (2014) established the necessary and sufficient condition for ensuring zero misclassification proportion (usually referred to as "strong consistency") with high probability when $k = 2$ and community sizes are equal. The result was later generalized to general SBMs with possibly unequal community sizes by Abbe and Sandon (2015a) using the notion of Chernoff-Hellinger (CH) divergence.

4. *Minimax optimal rates.* The foregoing three categories of results mainly focused on sharp conditions of achieving detection, weak consistency and strong consistency, respectively. Arguably, what is of more interest to statisticians is the optimal rates of misclassification proportion. When $(\sqrt{a}-\sqrt{b})^2/(k\log k) \to \infty$, it was derived by Zhang and Zhou (2015) that the optimal rate of misclassification proportion takes the

form of

$$\exp\left(-(1+o(1))\frac{nI^*}{k}\right), \tag{1}$$

for a k-community SBM with equal community sizes, where I^* is the Rényi divergence of order $\frac{1}{2}$ between Bern $\left(\frac{a}{n}\right)$ and Bern $\left(\frac{b}{n}\right)$ (Rényi, 1961). See Theorem 1 below for a more general and precise statement of the result. A special case of this result for symmetric SBMs with $k = 2$ was obtained in Yun and Proutiere (2014a).

5. *Algorithms.* Various algorithms have been proposed in the literature to achieve detection, weak consistency and strong consistency. A popular approach is spectral clustering. Its application on network data dates back to Hagen and Kahng (1992); McSherry (2001). Its performance on SBMs has been investigated by Coja-Oghlan (2010); Rohe et al. (2011); Sussman et al. (2012); Fishkind et al. (2013); Qin and Rohe (2013); Joseph and Yu (2013); Lei and Rinaldo (2014); Vu (2014); Chin et al. (2015); Jin (2015); Le et al. (2015), among others. Various ways for refining spectral clustering have been proposed, such as those in Amini et al. (2013); Abbe et al. (2014); Mossel et al. (2014); Lei and Zhu (2014); Yun and Proutiere (2014a); Chin et al. (2015), which lead to strong consistency or convergence rates that are exponential in signal-to-noise ratio, while Mossel et al. (2013a) studied the problem of minimizing a non-vanishing misclassification proportion. However, in the regime of weak consistency, with the exception of Yun and Proutiere (2014a) for equal-sized two community case, these refinement methods cannot attain the optimal misclassification proportion.

   Another important line of research is devoted to the investigation of likelihood-based methods, which was initiated by Bickel and Chen (2009) and later extended to more general settings by Zhao et al. (2012); Choi et al. (2012). To tackle the intractability of optimizing the likelihood function, an EM algorithm using pseudo-likelihood was proposed by Amini et al. (2013). Another way to overcome the intractability of the maximum likelihood estimator (MLE) is by convex relaxation. Various semi-definite relaxations were studied by Cai and Li (2014); Chen and Xu (2014); Amini and Levina (2014), and the sharp threshold for strong consistency can indeed be achieved by semi-definite programming (Hajek et al., 2014, 2015).

Main contribution The current paper proposes a computationally feasible algorithm that provably achieves the optimal misclassification proportion established in Zhang and Zhou (2015) adaptively under weak regularity conditions. It covers the cases of both finite and diverging number of communities and both equal and unequal community sizes, and it achieves both weak and strong consistency in the respective regimes. In addition, the algorithm is guaranteed to compute in polynomial time even when the number of communities diverges with the number of nodes. Since the error bound of the algorithm matches the optimal misclassification proportion (1) in Zhang and Zhou (2015) under weak conditions, it achieves various existing detection boundaries in the literature. For instance, for any fixed number of communities, the procedure is weakly consistent under the necessary and sufficient condition of Mossel et al. (2012, 2013b), and strongly consistent under the necessary and sufficient condition of Abbe et al. (2014); Mossel et al. (2014). Moreover, it could match the optimal misclassification proportion in Zhang and Zhou (2015) even

when k diverges. It is remarkable that the same algorithm can lead to optimal performance in both strong and weak consistency regiems, while most papers in the literature require different algorithms in the two regimes.

The core of the algorithm is a refinement scheme for community detection motivated by penalized MLE, an idea that was previously explored in Amini et al. (2013); Abbe et al. (2014); Mossel et al. (2014); Lei and Zhu (2014); Yun and Proutiere (2014a); Chin et al. (2015). As long as there exists an initial estimator that satisfies a certain weak consistency condition, the refinement scheme is able to obtain an improved estimator that achieves the optimal misclassification proportion with high probability. The key to achieve this goal is to optimize the *local* penalized likelihood function for each node separately. This local optimization step is completely data-driven and has a closed form solution, and hence can be computed very efficiently. The additional penalty term is indispensable as it plays a key role in ensuring the optimal performance when the community sizes are unequal and when the within community and/or between community edge probabilities are unequal.

To obtain a qualified initial estimator, we show that both spectral clustering and its normalized variant could satisfy the desired condition needed for subsequent refinement, though the refinement scheme works for any other method satisfying a certain weak consistency condition. Note that spectral clustering can be considered as a *global* method, and hence our two-stage algorithm runs in a "*from global to local*" fashion. In essence, with high probability, the global stage pinpoints a local neighborhood in which we shall search for solution to each local penalized maximum likelihood problem, and the subsequent local stage finds the desired solution.

Notable results after initial posting of this manuscript After the initial posting of this manuscript on arXiv (arXiv:1505.03772), there have appeared a number of papers with notable and related results. Below, we highlight them with brief discussions:

1. *Detection.* The algorithm proposed in this paper achieves optimal misclassification proportion in the weak and the strong consistency regimes. In the detection regime, optimal misclassification proportion and its statistical-computational gap have been studied in Deshpande et al. (2015); Mossel and Xu (2015); Abbe and Sandon (2015b).

2. *Strong and weak consistency.* Two algorithms were proposed in Abbe and Sandon (2015c) for community detection in general SBMs. One for the strong consistency regime, and the other for weak consistency and detection. The one for strong consistency was shown to achieve the goal all the way to the CH divergence threshold, which is tighter than the strong consistency result in the present paper for asymmetric SBMs. However, the convergence rate for the other algorithm is sub-optimal in the weak consistency regime even for symmetric SBMs. See also the discussion below for the further improvement in Yun and Proutiere (2015).

3. *Weighted/labeled SBMs.* Jog and Loh (2015) extends the results in Zhang and Zhou (2015) to networks with weighted/labeled edges using Rényi divergence of order $\frac{1}{2}$. A more recent paper Yun and Proutiere (2015) studied optimal misclassification proportion for weighted/labeled SBMs. The proposed algorithm in Yun and Proutiere (2015) is able to achieve the CH limit (Abbe and Sandon, 2015a) in the strong consistency regime. It also attains optimal misclassification proportion in the weak consistency

regime with respect to a potentially smaller class than the ones used in our paper. However, their stronger results also require a relatively stronger set of conditions. For example, they require $k = O(1)$. Moreover, for equal-sized two community binary SBMs, with a and b defined at the beginning of this section, Yun and Proutiere (2015) requires $a \asymp b$ and $a - b \asymp b$. In comparison, we do not require the latter for any result in this manuscript. We can even drop the former if we are willing to accept any $1 - \epsilon$ relaxation of the tight constant of the exponent in our error rates. See, for instance, Theorem 12.

4. *Degree-corrected block models.* The paper Gao et al. (2016) studied degree-corrected block models by deriving the minimax rates for misclassification proportion and proposing an adaptive algorithm.

In summary, progress has been made along several different directions after initial posting of the present manuscript. However, none of the aforementioned results dominates those we are to present in the rest of the paper.

Organization and notation The rest of the paper is organized as follows. Section 2 formally sets up the community detection problem and presents the two-stage algorithm. The theoretical guarantees for the proposed method are given in Section 3, followed by numerical results demonstrating its competitive performance on simulated datasets in Section 4. A discussion on the results in the current paper and possible directions for future investigation is included in Section 5. Section 6 presents the proofs of main results with some technical details deferred to the appendix.

We close this section by introducing some notation. For a matrix $M = (M_{ij})$, we denote its Frobenius norm by $\|M\|_{\mathrm{F}} = \sqrt{\sum_{ij} M_{ij}^2}$ and its operator norm by $\|M\|_{\mathrm{op}} = \max_l \lambda_l(M)$, where $\lambda_l(M)$ is its l^{th} singular value. We use M_{i*} to denote its i^{th} row. The norm $\|\cdot\|$ is the usual Euclidean norm for vectors. For a set S, $|S|$ denotes its cardinality. The notation $\mathbb{P}$ and $\mathbb{E}$ are generic probability and expectation operators whose distribution is determined from the context. For two positive sequences $\{x_n\}$ and $\{y_n\}$, $x_n \asymp y_n$ means $x_n/C \le y_n \le Cx_n$ for some constant $C \ge 1$ independent of n, while $x_n = o(y_n)$ means $x_n/y_n \to 0$ as $n \to \infty$. Throughout the paper, unless otherwise noticed, we use C, c and their variants to denote absolute constants, whose values may change from line to line.

2. Problem formulation and methodology

In this section, we give a precise formulation of the community detection problem and present a new method for it. The method consists of two stages: initialization and refinement. We shall first introduce the second stage, which is the main algorithm of the paper. It clusters the network data by performing a node-wise penalized neighbor voting based on some initial community assignment. Then, we will discuss several candidates for the initialization step including a new greedy algorithm for clustering the leading eigenvectors of the adjacency matrix or of the graph Laplacian that is tailored specifically for stochastic block models. Theoretical guarantees for the algorithms introduced in the current section will be presented in Section 3.

2.1 Community detection in stochastic block model

Recall that a stochastic block model is completely characterized by a symmetric connectivity matrix $B \in [0,1]^{k\times k}$ and a label vector $\sigma \in [k]^n$. One widely studied parameter space of SBM is

$$\Theta_0(n,k,a,b,\beta) = \Bigg\{(B,\sigma):\ \sigma:[n]\to[k], |\{u\in[n]:\sigma(u)=i\}| \in \left[\frac{n}{\beta k}-1, \frac{\beta n}{k}+1\right],\ \forall i\in[k],$$
$$B=(B_{ij})\in[0,1]^{k\times k}, B_{ii}=\frac{a}{n} \text{ for all } i \text{ and } B_{ij}=\frac{b}{n} \text{ for all } i\neq j\Bigg\} \quad (2)$$

where $\beta \geq 1$ is an absolute constant. This parameter space $\Theta_0(n,k,a,b,\beta)$ contains all SBMs in which the within community connection probabilities are all equal to $\frac{a}{n}$ and the between community connection probabilities are all equal to $\frac{b}{n}$. In the special case of $\beta = 1$, all communities are of nearly equal sizes.

Assuming equal within and equal between connection probabilities can be restrictive. Thus, we also introduce the following larger parameter space

$$\Theta(n,k,a,b,\lambda,\beta;\alpha) = \Bigg\{(B,\sigma):\ \sigma:[n]\to[k], |\{u\in[n]:\sigma(u)=i\}| \in \left[\frac{n}{\beta k}-1, \frac{\beta n}{k}+1\right],\ \forall i\in[k],$$
$$B=B^T=(B_{ij})\in[0,1]^{k\times k}, \frac{b}{\alpha n}\leq \frac{1}{k(k-1)}\sum_{i\neq j}B_{ij}\leq \max_{i\neq j}B_{ij}=\frac{b}{n},$$
$$\frac{a}{n}=\min_i B_{ii}\leq \max_i B_{ii}\leq \frac{\alpha a}{n},$$
$$\lambda_k(P)\geq\lambda \text{ with } P=(P_{uv})=(B_{\sigma(u),\sigma(v)})\Bigg\}. \quad (3)$$

Throughout the paper, we treat $\beta \geq 1$ and $\alpha \geq 1$ as absolute constants, while k, a, b and λ should be viewed as functions of the number of nodes n which can vary as n grows. Moreover, we assume $0 < \frac{b}{n} < \frac{a}{n} \leq 1-\epsilon$ throughout the paper for some numeric constant $\epsilon \in (0,1)$. Thus, the parameter space $\Theta(n,k,a,b,\lambda,\beta;\alpha)$ requires that the within community connection probabilities are bounded from below by $\frac{a}{n}$ and the connection probabilities between any two communities are bounded from above by $\frac{b}{n}$. In addition, it requires that the sizes of different communities are comparable. In order to guarantee that $\Theta(n,k,a,b,\lambda,\beta;\alpha)$ is a larger parameter space than $\Theta_0(n,k,a,b,\beta)$, we always require λ to be positive and sufficiently small such that

$$\Theta_0(n,k,a,b,\beta)\subset\Theta(n,k,a,b,\lambda,\beta;\alpha). \quad (4)$$

According to Proposition 24 in the appendix, a sufficient condition for (4) is $\lambda \leq \frac{a-b}{2\beta k}$. We assume (4) throughout the rest of the paper.

The labels on the n nodes induce a community structure $[n] = \cup_{i=1}^k \mathcal{C}_i$, where $\mathcal{C}_i = \{u\in[n]:\sigma(u)=i\}$ is the i^{th} community with size $n_i = |\mathcal{C}_i|$. Our goal is to reconstruct this partition, or equivalently, to estimate the label of each node modulo any permutation of label symbols. Therefore, a natural error measure is the misclassification proportion defined as

$$\ell(\widehat{\sigma},\sigma) = \min_{\pi\in S_k}\frac{1}{n}\sum_{u\in[n]}\mathbf{1}_{\{\widehat{\sigma}(u)\neq\pi(\sigma(u))\}}, \quad (5)$$

where S_k stands for the symmetric group on $[k]$ consisting of all permutations of $[k]$.

2.2 Main algorithm

We now present the main method of the paper – a refinement algorithm for community detection in stochastic block model motivated by penalized local maximum likelihood estimation.

To motivate our proposal, for any SBM in the parameter space $\Theta_0(n, k, a, b, 1)$ with equal community size, the MLE for σ (Cai and Li, 2014; Chen and Xu, 2014; Zhang and Zhou, 2015) is

$$\widehat{\sigma} = \underset{\sigma:[n]\to[k]}{\operatorname{argmax}} \sum_{u<v} A_{uv} \mathbf{1}_{\{\sigma(u)=\sigma(v)\}}, \tag{6}$$

which is a combinatorial optimization problem and hence is computationally intractable. However, node-wise optimization of (6) has a simple closed form solution. Suppose the values of $\{\sigma(u)\}_{u=2}^n$ are known and we want to estimate $\sigma(1)$. Then, (6) reduces to

$$\widehat{\sigma}(1) = \underset{i\in[k]}{\operatorname{argmax}} \sum_{\{v\neq 1:\sigma(v)=i\}} A_{1v}. \tag{7}$$

For each $i \in [k]$, the quantity $\sum_{\{v\neq 1:\sigma(v)=i\}} A_{1v}$ is the number of neighbors that the first node has in the i^{th} community. Therefore, the most likely label for the first node is the one it has the most connections with when all communities are of equal sizes. In practice, we do not know any label in advance. However, we may estimate the labels of all but the first node by first applying a community detection algorithm σ^0 on the subnetwork excluding the first node and its associated edges, the adjacency matrix of which is denoted by A_{-1} since it is the $(n-1)\times(n-1)$ submatrix of A with its first row and first column removed. Once we estimate the remaining labels, we can apply (7) to estimate $\sigma(1)$ but with $\{\sigma(v)\}_{v=2}^n$ replaced with the estimated labels.

For any $u \in [n]$, let A_{-u} denote the $(n-1)\times(n-1)$ submatrix of A with its u^{th} row and u^{th} column removed. Given any community detection algorithm σ^0 which is able to cluster any graph on $n-1$ nodes into k categories, we present the precise description of our refinement scheme in Algorithm 1.

The algorithm works in two consecutive steps. The first step carries out the foregoing heuristics on a node by node basis. For each fixed node u, we first leave the node out and apply the available community detection algorithm σ^0 on the remaining $n-1$ nodes and the edges among them (as summarized in the matrix $A_{-u} \in \{0,1\}^{(n-1)\times(n-1)}$) to obtain an initial community assignment vector σ_u^0. For convenience, we make σ_u^0 an n-vector by fixing $\sigma_u^0(u) = 0$, though applying σ^0 on A_{-u} does not give any community assignment for u. We then assign the label of the u^{th} node according to (10), which is essentially (7) with σ replaced with σ_u^0 except for the additional penalty term. The additional penalty term is added to ensure the optimal performance even when both the diagonal and the off-diagonal entries of the connectivity matrix B are allowed to take different values and the community sizes are not necessarily equal. To determine the penalty parameter ρ_u in an adaptive way as spelled out in (11) – (12), we first estimate the connectivity matrix B based on A_{-u} in (8) – (9). After we obtain the community assignment for u, we organize the assignment for all n vertices into an n-vector $\widehat{\sigma}_u$. We call this step "penalized neighbor

Algorithm 1: A refinement scheme for community detection

Input: Adjacency matrix $A \in \{0,1\}^{n\times n}$,
number of communities k,
initial community detection method σ^0.

Output: Community assignment $\widehat{\sigma}$.

Penalized neighbor voting:

for $u = 1$ **to** n **do**

Apply σ^0 on A_{-u} to obtain $\sigma_u^0(v)$ for all $v \neq u$ and let $\sigma_u^0(u) = 0$;

Define $\widetilde{\mathcal{C}}_i^u = \{v : \sigma_u^0(v) = i\}$ for all $i \in [k]$; let $\widetilde{\mathcal{E}}_i^u$ be the set of edges within $\widetilde{\mathcal{C}}_i^u$, and $\widetilde{\mathcal{E}}_{ij}^u$ the set of edges between $\widetilde{\mathcal{C}}_i^u$ and $\widetilde{\mathcal{C}}_j^u$ when $i \neq j$;

Define

$$\widehat{B}_{ii}^u = \frac{|\widetilde{\mathcal{E}}_i^u|}{\frac{1}{2}|\widetilde{\mathcal{C}}_i^u|(|\widetilde{\mathcal{C}}_i^u|-1)}, \quad \widehat{B}_{ij}^u = \frac{|\widetilde{\mathcal{E}}_{ij}^u|}{|\widetilde{\mathcal{C}}_i^u||\widetilde{\mathcal{C}}_j^u|}, \quad \forall i \neq j \in [k], \tag{8}$$

and let

$$\widehat{a}_u = n \min_{i\in[k]} \widehat{B}_{ii}^u \quad \text{and} \quad \widehat{b}_u = n \max_{i\neq j\in[k]} \widehat{B}_{ij}^u. \tag{9}$$

Define $\widehat{\sigma}_u : [n] \to [k]$ by setting $\widehat{\sigma}_u(v) = \sigma_u^0(v)$ for all $v \neq u$ and

$$\widehat{\sigma}_u(u) = \operatorname*{argmax}_{l\in[k]} \sum_{\sigma_u^0(v)=l} A_{uv} - \rho_u \sum_{v\in[n]} \mathbf{1}_{\{\sigma_u^0(v)=l\}} \tag{10}$$

where for

$$t_u = \frac{1}{2}\log\frac{\widehat{a}_u(1-\widehat{b}_u/n)}{\widehat{b}_u(1-\widehat{a}_u/n)}, \tag{11}$$

we define

$$\rho_u = -\frac{1}{2t_u}\log\left(\frac{\frac{\widehat{a}_u}{n}e^{-t_u}+1-\frac{\widehat{a}_u}{n}}{\frac{\widehat{b}_u}{n}e^{t_u}+1-\frac{\widehat{b}_u}{n}}\right), \tag{12}$$

end

Consensus:

Define $\widehat{\sigma}(1) = \widehat{\sigma}_1(1)$. For $u = 2, \ldots, n$, define

$$\widehat{\sigma}(u) = \operatorname*{argmax}_{l\in[k]} |\{v : \widehat{\sigma}_1(v) = l\} \cap \{v : \widehat{\sigma}_u(v) = \widehat{\sigma}_u(u)\}|. \tag{13}$$

voting" since the first term on the RHS of (10) counts the number of neighbors of u in each (estimated) community while the second term is a penalty term proportional to the size of each (estimated) community.

Once we complete the above procedure for each of the n nodes, we obtain n vectors $\widehat{\sigma}_u \in [k]^n$, $u = 1, \ldots, n$, and turn to the second step of the algorithm. The basic idea behind the second step is to obtain a unified community assignment by assembling $\{\widehat{\sigma}_u(u) : u \in [n]\}$ and the immediate hurdle is that each $\widehat{\sigma}_u$ is only determined up to a permutation of the community labels. Thus, the second step aims to align these different permutations by (13) before we assemble the $\widehat{\sigma}_u(u)$'s. We call this step "consensus" since we are essentially looking for a consensus on the community labels for n possibly different community assignments, under the assumption that all of them are close to the ground truth up to some permutation.

2.3 Initialization via spectral methods

In this section, we present algorithms that can be used as initializers in Algorithm 1. Note that for any model in (3), the matrix P has rank at most k and $\mathbb{E}A_{uv} = P_{uv}$ for all $u \neq v$. We may first reduce the dimension of the data and then apply some clustering algorithm. Such an approach is usually referred to as spectral clustering (von Luxburg, 2007). The application of spectral clustering on network data goes back to Hagen and Kahng (1992); McSherry (2001), and its performance under the stochastic block model has been investigated by Coja-Oghlan (2010); Rohe et al. (2011); Sussman et al. (2012); Fishkind et al. (2013); Qin and Rohe (2013); Joseph and Yu (2013); Lei and Rinaldo (2014); Vu (2014); Chin et al. (2015); Jin (2015); Le et al. (2015), among others. Technically speaking, spectral clustering refers to the general method of clustering eigenvectors of some data matrix. For random graphs, two commonly used methods are unnormalized spectral clustering (USC) and normalized spectral clustering (NSC). The former refers to clustering the eigenvectors of the adjacency matrix A itself and the latter refers to clustering the eigenvectors of the associated graph Laplacian $L(A)$. To formally define the graph Laplacian, we introduce the notation $d_u = \sum_{v\in[n]} A_{uv}$ for the degree of the u^{th} node. The graph Laplacian operator $L : A \mapsto L(A)$ is defined by $L(A) = ([L(A)]_{uv})$ where $[L(A)]_{uv} = d_u^{-1/2} d_v^{-1/2} A_{uv}$. Although there have been debates and studies on which one works better (see, for example, von Luxburg et al. (2008); Sarkar and Bickel (2013)), for our purpose, both of them can lead to sufficiently decent initial estimators.

The performances of USC and NSC depend critically on the bounds $\|A - P\|_{\text{op}}$ and $\|L(A) - L(P)\|_{\text{op}}$, respectively. However, as pointed out by Chin et al. (2015); Le et al. (2015), the matrices A and $L(A)$ are not good estimators of P and $L(P)$ under the operator norm when the graph is sparse in the sense that $\max_{u,v\in[n]} P_{uv} = o(\log n/n)$. Thus, it is necessary to regularize A and $L(A)$ in order to achieve better performances for USC and NSC. The adjacency matrix A can be regularized by trimming those nodes with high degrees. Define the trimming operator $T_\tau : A \mapsto T_\tau(A)$ by replacing the u^{th} row and the u^{th} column of A with 0 whenever $d_u \geq \tau$, and so $T_\tau(A)$ and A are of the same dimensions. It is argued in Chin et al. (2015) that by removing those high-degree nodes, $T_\tau(A)$ has better convergence properties. Regularization method for graph Laplacian goes back to Amini et al. (2013) and its theoretical properties have been studied by Joseph and Yu (2013); Le et al. (2015). In particular, Amini et al. (2013) proposed to use $L(A_\tau)$ for NSC where $A_\tau = A + \frac{\tau}{n}\mathbf{1}\mathbf{1}^T$ and $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^n$. From now on, we use USC(τ) and NSC(τ) to denote unnormalized spectral clustering and normalized spectral clustering with

regularization parameter τ, respectively. Note that the unregularized USC is USC(∞) and the unregularized NSC is NSC(0).

Another important issue in spectral clustering lies in the subsequent clustering method used to cluster the eigenvectors. A popular choice is k-means clustering. However, finding the global solution to the k-means problem is NP-hard (Aloise et al., 2009; Mahajan et al., 2009). Kumar et al. (2004) proposed a polynomial time algorithm for achieving $(1+\epsilon)$ approximation to the k-means problem for any fixed k, which was utilized in Lei and Rinaldo (2014) to establish consistency for spectral clustering under stochastic block models with a fixed number of communities. However, a closer look at the complexity bound suggests that the smallest possible ϵ is proportional to k. Thus, applying the algorithm and the associated bound in Kumar et al. (2004) directly in our settings can lead to inferior error bounds when $k \to \infty$ as $n \to \infty$. To address this issue for stochastic block models, we propose a greedy clustering method in Algorithm 2. The method is inspired by the fact that the clustering centers in stochastic block models are well separated from each other on the population level. It is straightforward to check that the complexity of Algorithm 2 is polynomial in n.

Algorithm 2: A greedy method for clustering

Input: Data matrix $\widehat{U} \in \mathbb{R}^{n\times k}$, either the leading eigenvectors of $T_\tau(A)$ or that of $L(A_\tau)$,
number of communities k,
critical radius $r = \mu\sqrt{\frac{k}{n}}$ with some constant $\mu > 0$.

Output: Community assignment $\widehat{\sigma}$.

Set $S = [n]$;
for $i = 1$ **to** k **do**
  Let $t_i = \arg\max_{u\in S} \left|\left\{v \in S : \left\|\widehat{U}_{v*} - \widehat{U}_{u*}\right\| < r\right\}\right|$;
  Set $\widehat{\mathcal{C}}_i = \left\{v \in S : \left\|\widehat{U}_{v*} - \widehat{U}_{t_i*}\right\| < r\right\}$;
  Label $\widehat{\sigma}(u) = i$ for all $u \in \widehat{\mathcal{C}}_i$;
  Update $S \leftarrow S\backslash\widehat{\mathcal{C}}_i$.
end
If $S \neq \varnothing$, then for any $u \in S$, set $\widehat{\sigma}(u) = \operatorname{argmin}_{i\in[k]} \frac{1}{|\widehat{\mathcal{C}}_i|}\sum_{v\in\widehat{\mathcal{C}}_i}\left\|\widehat{U}_{u*} - \widehat{U}_{v*}\right\|$.

Last but not least, we would like to emphasize that one needs not limit the initialization algorithm to the spectral methods introduced in this section. As Theorem 4 below shows, Algorithm 1 works for any initialization method that satisfies a weak consistency condition.

3. Theoretical properties

Before stating the theoretical properties of the proposed method, we first review the minimax rate in Zhang and Zhou (2015), which serves as the optimality benchmark. The

minimax risk is governed by the following critical quantity,

$$I^* = -2\log\left(\sqrt{\frac{a}{n}}\sqrt{\frac{b}{n}} + \sqrt{1-\frac{a}{n}}\sqrt{1-\frac{b}{n}}\right), \tag{14}$$

which is the Rényi divergence of order $\frac{1}{2}$ between Bern $\left(\frac{a}{n}\right)$ and Bern $\left(\frac{b}{n}\right)$, i.e., Bernoulli distributions with success probabilities $\frac{a}{n}$ and $\frac{b}{n}$ respectively. Recall that $0 < \frac{b}{n} < \frac{a}{n} \le 1-\epsilon$ is assumed throughout the paper. It can be shown that $I^* \asymp \frac{(a-b)^2}{na}$. Moreover, when $\frac{a}{n} = o(1)$,

$$\begin{aligned} I^* &= (1+o(1))\frac{(\sqrt{a}-\sqrt{b})^2}{n} = (1+o(1))\left[\left(\sqrt{\frac{a}{n}}-\sqrt{\frac{b}{n}}\right)^2 + \left(\sqrt{1-\frac{a}{n}}-\sqrt{1-\frac{b}{n}}\right)^2\right] \\ &= (2+o(1))H^2\left(\text{Bern}\left(\tfrac{a}{n}\right), \text{Bern}\left(\tfrac{b}{n}\right)\right), \end{aligned}$$

where $H^2(P,Q) = \frac{1}{2}\int(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$ is the squared Hellinger distance between two distributions P and Q. The minimax rate for the parameter spaces (2) and (3) under the loss function (5) is given in the following theorem.

Theorem 1 (Zhang and Zhou (2015)) *When $\frac{(a-b)^2}{ak\log k} \to \infty$, we have*

$$\inf_{\widehat{\sigma}} \sup_{(B,\sigma)\in\Theta} \mathbb{E}_{B,\sigma}\ell(\widehat{\sigma},\sigma) = \begin{cases} \exp\left(-(1+\eta)\frac{nI^*}{2}\right), & k=2; \\ \exp\left(-(1+\eta)\frac{nI^*}{\beta k}\right), & k \ge 3, \end{cases}$$

for both $\Theta = \Theta_0(n,k,a,b,\beta)$ and $\Theta = \Theta(n,k,a,b,\lambda,\beta;\alpha)$ with any $\lambda \le \frac{a-b}{2\beta k}$ and any $\beta \in [1,\sqrt{5/3})$, where $\eta = \eta_n \to 0$ is some sequence tending to 0 as $n \to \infty$.

Remark 2 *The assumption $\beta \in [1,\sqrt{5/3})$ is needed in Zhang and Zhou (2015) for some technical reason. Here, the parameter β enters the minimax rates when $k \ge 3$ since the worst case is essentially when one has two communities of size $\frac{n}{\beta k}$, while for $k = 2$, the worst case is essentially two communities of size $\frac{n}{2}$. For all other results in this paper, we allow β to be an arbitrary constant no less than 1.*

Remark 3 *The rate in Theorem 1 is optimal in a minimax sense. It is optimal for the worse-case instances in $\Theta_0(n,k,a,b,1)$. More general instance-optimal fundamental limits are referred to Abbe and Sandon (2015a). Details discussion will be given in Section 5.*

To this end, let us show that the two-stage algorithm proposed in Section 2 achieves the optimal misclassification proportion. The essence of the two-stage algorithm lies in the refinement scheme described in Algorithm 1. As long as any initialization step satisfies a certain weak consistency criterion, the refinement step directly leads to a solution with optimal misclassification proportion. To be specific, the initialization step needs to satisfy the following condition.

Condition 1 *There exist constants $C_0, \delta > 0$ and a positive sequence $\gamma = \gamma_n$ such that*

$$\inf_{(B,\sigma)\in\Theta} \min_{u\in[n]} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma, \sigma_u^0) \le \gamma\right\} \ge 1 - C_0 n^{-(1+\delta)}, \tag{15}$$

for some parameter space Θ.

Under Condition 1, we have the following upper bounds regarding the performance of the proposed refinement scheme.

Theorem 4 *Suppose as $n \to \infty$, $\frac{(a-b)^2}{ak\log k} \to \infty$, $a \asymp b$ and Condition 1 is satisfied for*

$$\gamma = o\left(\frac{1}{k\log k}\right) \tag{16}$$

and $\Theta = \Theta_0(n, k, a, b, \beta)$. Then there is a sequence $\eta \to 0$ such that

$$\begin{aligned} &\sup_{(B,\sigma)\in\Theta} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma,\widehat{\sigma}) \ge \exp\left(-(1-\eta)\frac{nI^*}{2}\right)\right\} \to 0, \qquad \text{if } k = 2, \\ &\sup_{(B,\sigma)\in\Theta} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma,\widehat{\sigma}) \ge \exp\left(-(1-\eta)\frac{nI^*}{\beta k}\right)\right\} \to 0, \qquad \text{if } k \ge 3, \end{aligned} \tag{17}$$

where I^* is defined as in (14)*.*

If in addition Condition 1 is satisfied for γ satisfying both (16) *and*

$$\gamma = o\left(\frac{a-b}{ak}\right) \tag{18}$$

and $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$, then the conclusion in (17) *continues to hold for $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$.*

Theorem 4 assumes $a \asymp b$. The case when $a \asymp b$ may not hold is considered in Section 5. Compared with Theorem 1, the upper bounds (17) achieved by Algorithm 1 is minimax optimal. The condition (16) for the parameter space $\Theta_0(n, k, a, b, \beta)$ is very mild. When $k = O(1)$, it reduces to $\gamma = o(1)$ and simply means that the initialization should be weakly consistent. For $k \to \infty$, it implies that the misclassification proportion within each community converges to zero. Note that if the initialization step gives wrong labels to all nodes in one particular community, then the misclassification proportion is at least $1/k$. The condition (16) rules out this situation. For the parameter space $\Theta(n, k, a, b, \lambda, \beta; \alpha)$, an extra condition (18) is required. This is because estimating the connectivity matrix B in $\Theta(n, k, a, b, \lambda, \beta; \alpha)$ is harder than in $\Theta_0(n, k, a, b, \beta)$. If we do not pursue adaptive estimation, (18) is not needed.

Remark 5 *Theorem 4 is an adaptive result without assuming the knowledge of a and b. When these two parameters are known, we can directly use a and b in (11) of Algorithm 1. By scrutinizing the proof of Theorem 4, the conditions (16) and (18) can be weakened to $\gamma = o(k^{-1})$ in this case.*

Given the results of Theorem 4, it remains to check the initialization step via spectral clustering satisfies Condition 1. For matrix $P = (P_{uv}) = (B_{\sigma(u)\sigma(v)})$ with (B, σ) belonging to either $\Theta_0(n, k, a, b, \beta)$ or $\Theta(n, k, a, b, \lambda, \beta; \alpha)$, we use λ_k to denote $\lambda_k(P)$. Define the average degree by

$$\bar{d} = \frac{1}{n} \sum_{u \in [n]} d_u. \tag{19}$$

Theorem 6 *Assume $e \le a \le C_1 b$ for some constant $C_1 > 0$ and*

$$\frac{ka}{\lambda_k^2} \le c, \tag{20}$$

for some sufficiently small $c \in (0, 1)$. Consider USC(τ) with a sufficiently small constant $\mu > 0$ in Algorithm 2 and $\tau = C_2 \bar{d}$ for some sufficiently large constant $C_2 > 0$. For any constant $C' > 0$, there exists some $C > 0$ only depending on C', C_1, C_2 and μ such that

$$\ell(\widehat{\sigma}, \sigma) \le C \frac{a}{\lambda_k^2},$$

with probability at least $1 - n^{-C'}$. If k is fixed, the same conclusion holds without assuming $a \le C_1 b$.

Remark 7 *Theorem 6 improves the error bound for spectral clustering in Lei and Rinaldo (2014). While Lei and Rinaldo (2014) requires the assumption $a > C \log n$, our result also holds for $a = o(\log n)$. A result close to ours is that by Chin et al. (2015), but their clustering step is different from Algorithm 2. Moreover, the conclusion of Theorem 6 holds with probability $1 - n^{-C'}$ for an arbitrary large C', which is critical because the initialization step needs to satisfy Condition 1 for the subsequent refinement step to work. On the other hand, the bound in Chin et al. (2015) is stated with probability $1 - o(1)$.*

Remark 8 *For the parameter space $\Theta_0(n, k, a, b, \beta)$, we have $\lambda_k \ge \frac{a-b}{\beta k}$. Then, Theorem 6 implies that consistency is achieved when $\frac{(a-b)^2}{a} \to \infty$ in the case $k = O(1)$, and when $\frac{(a-b)^2}{ak^3} > C$ for some sufficiently large $C > 0$ in the case $k \to \infty$.*

When $k = O(1)$, Theorem 4 and Theorem 6 jointly imply the following result.

Corollary 9 *Consider Algorithm 1 initialized by σ^0 with USC(τ) for $\tau = C\bar{d}$, where C is a sufficiently large constant. Suppose as $n \to \infty$, $k = O(1)$, $\frac{(a-b)^2}{a} \to \infty$ and $a \asymp b$. Then, there exists a sequence $\eta \to 0$ such that*

$$\sup_{(B,\sigma) \in \Theta} \mathbb{P}_{B,\sigma} \left\{ \ell(\sigma, \widehat{\sigma}) \ge \exp\left(-(1 - \eta) \frac{nI^*}{2} \right) \right\} \to 0, \qquad \text{if } k = 2,$$

$$\sup_{(B,\sigma) \in \Theta} \mathbb{P}_{B,\sigma} \left\{ \ell(\sigma, \widehat{\sigma}) \ge \exp\left(-(1 - \eta) \frac{nI^*}{\beta k} \right) \right\} \to 0, \qquad \text{if } k \ge 3,$$

where the parameter space is $\Theta = \Theta_0(n, k, a, b, \beta)$.

Compared with Theorem 1, the proposed procedure achieves the minimax rate under the condition $\frac{(a-b)^2}{a} \to \infty$ and $a \asymp b$. When $k = O(1)$, the condition $\frac{(a-b)^2}{a} \to \infty$ is necessary and sufficient for weak consistency in view of Theorem 1. More general results including the case of $k \to \infty$ are stated and discussed in Section 5.

The following theorem characterizes the misclassification rate of normalized spectral clustering.

Theorem 10 *Assume $e \le a \le C_1 b$ for some constant $C_1 > 0$ and*

$$\frac{ka \log a}{\lambda_k^2} \le c, \tag{21}$$

for some sufficiently small $c \in (0,1)$. Consider NSC(τ) with a sufficiently small constant $\mu > 0$ in Algorithm 2 and $\tau = C_2 \bar{d}$ for some sufficiently large constant $C_2 > 0$. Then, for any constant $C' > 0$, there exists some $C > 0$ only depending on C', C_1, C_2 and μ such that

$$\ell(\widehat{\sigma}, \sigma) \le C \frac{a \log a}{\lambda_k^2},$$

with probability at least $1 - n^{-C'}$. If k is fixed, the same conclusion holds without assuming $a \le C_1 b$.

Remark 11 *A slightly different regularization of normalized spectral clustering is studied by Qin and Rohe (2013) only for the dense regime, while Theorem 10 holds under both dense and sparse regimes. Moreover, our result also improves that of Le et al. (2015) due to our tighter bound on $\|L(A_\tau) - L(P_\tau)\|_{\mathrm{op}}$ in Lemma 22 below. We conjecture that the $\log a$ factor in both the assumption and the bound of Theorem 10 can be removed.*

Note that Theorem 6 and Theorem 10 are stated in terms of the quantity λ_k. We may specialize the results into the parameter spaces defined in (2) and (3). By Proposition 24, $\lambda_k \ge \frac{a-b}{2\beta k}$ for $\Theta_0(n, k, a, b, \beta)$ and $\lambda_k \ge \lambda$ for $\Theta(n, k, a, b, \lambda, \beta; \alpha)$. The implications of Theorem 6 and Theorem 10 and their use as initialization in for Algorithm 1 are discussed in full details in Section 5.

4. Numerical results

In this section we present the performance of the proposed algorithm on simulated datasets. The experiments cover three different scenarios: (1) dense network with communities of equal sizes; (2) dense network with communities of unequal sizes; and (3) sparse network. Recall the definition of $\bar{d}$ in (19). For each setting, we report results of Algorithm 1 initialized with four different approaches: USC(∞), USC($2\bar{d}$), NSC(0) and NSC($\bar{d}$), the description of which can all be found in Section 2.3. For all these spectral clustering methods, Algorithm 2 was used to cluster the leading eigenvectors. The constant μ in the critical radius definition was set to be 0.5 in all the results reported here. For each setting, the results are based on 100 independent draws from the underlying stochastic block model.

To achieve faster running time, we also ran a simplified version of Algorithm 1. Instead of obtaining n different initializers $\{\sigma_u\}_{u \in [n]}$ to refine each node separately, the simplified

algorithm refines all the nodes with a single initialization on the whole network. Thus, the running time can be reduced roughly by a factor of n. Simulation results below suggest that the simplified version achieves similar performances to that of Algorithm 1 in all the settings we have considered. For the precise description of the simplified algorithm, we refer readers to Algorithm 3 in the appendix.

Balanced case In this setting, we generate networks with 2500 nodes and 10 communities, each of which consists of 250 nodes, and we set $B_{ii} = 0.48$ for all i and $B_{ij} = 0.32$ for all $i \neq j$. Figure 1 shows the boxplots of the number of misclassified nodes. The first four boxplots correspond to the four different spectral clustering methods, in the order of USC(∞), USC($2\bar{d}$), NSC(0) and NSC($\bar{d}$). The middle four correspond to the results achieved by applying the simplified refinement scheme with these four initialization methods, and the last four show the results of Algorithm 1 with these four initialization methods. Regardless of the initialization method, Algorithm 1 or its simplified version reduces the number of misclassified nodes from around 30 to around 5.

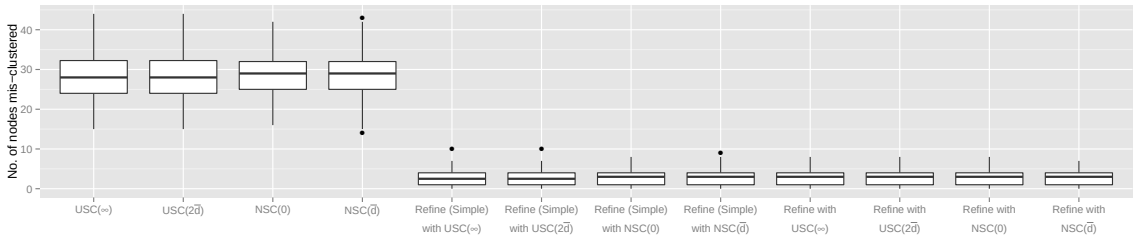

Figure 1: Boxplots of number of misclassified nodes: Balanced case. *Simple* indicates that the simplified version of Algorithm 1 is used instead.

Imbalanced case In this setting, we generate networks with 2000 nodes and 4 communities, the sizes of which are $200, 400, 600$ and 800, respectively. The connectivity matrix is

$$B = \begin{pmatrix} 0.50 & 0.29 & 0.35 & 0.25 \\ 0.29 & 0.45 & 0.25 & 0.30 \\ 0.35 & 0.25 & 0.50 & 0.35 \\ 0.25 & 0.30 & 0.35 & 0.45 \end{pmatrix}.$$

Hence, the within-community edge probability is no smaller than 0.45 while the between-community edge probability is no greater than 0.35, and the underlying SBM is inhomogeneous. Figure 2 shows the boxplots of the number of misclassified nodes obtained by different initialization methods and their refinements, and the boxplots are presented in the same order as those in Figure 1. Similarly, we can see refinement significantly reduces the error.

Sparse case In this setting we consider a much sparser stochastic block model than the previous two cases. In particular, each simulated network has 4000 nodes, divided into 10 communities all of size 400. We set all $B_{ii} = 0.032$ and all $B_{ij} = 0.005$ when

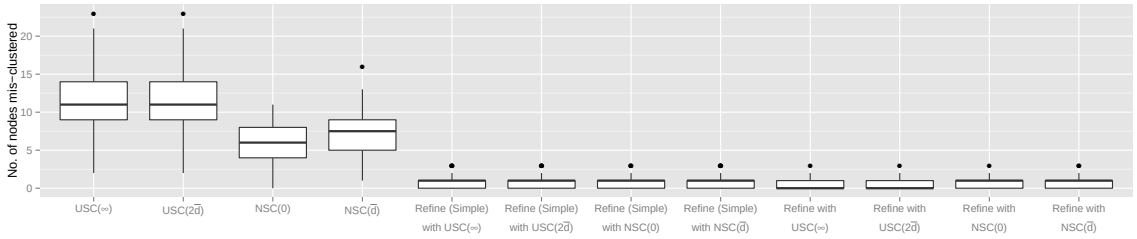

Figure 2: Boxplots of number of misclassified nodes: imbalanced case. *Simple* indicates that the simplified version of Algorithm 1 is used instead.

$i \neq j$. The average degree of each node in the network is around 30. Figure 3 shows the boxplots of the number of misclassified nodes obtained by different initialization methods and their refinements, and the boxplots are presented in the same order as those in Figure 1. Compared with either USC or NSC initialization, refinement reduces the number of misclassified nodes by 50%.

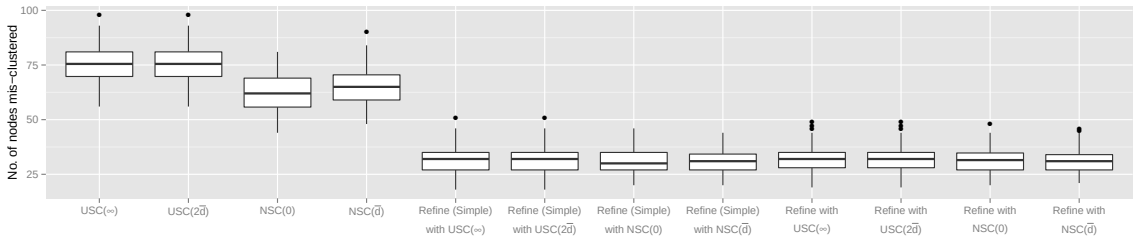

Figure 3: Boxplots of number of misclassified nodes: Sparse case. *Simple* indicates that the simplified version of Algorithm 1 is used instead.

Summary In all three simulation settings, for all four initialization approaches considered, the refinement scheme in Algorithm 1 (and its simplified version) was able to significantly reduce the number of misclassified nodes, which is in agreement with the theoretical properties presented in Section 3.

5. Discussion

In this section, we discuss a few important issues related to the methodology and the theory we have presented in the previous sections.

5.1 Error bounds when $a \asymp b$ may not hold

In Section 3, we established upper bounds on misclassification proportion under the assumption of $a \asymp b$. The following theorem shows that slightly weaker upper bounds can be obtained even when $a \asymp b$ does not hold. To state the result, recall that we assume throughout the paper $\frac{a}{n} \leq 1 - \epsilon$ for some numeric constant $\epsilon \in (0, 1)$.

Theorem 12 *Suppose as $n \to \infty$, $\frac{(a-b)^2}{ak\log k} \to \infty$ and Condition 1 is satisfied for γ satisfying* (16) *and $\Theta = \Theta_0(n, k, a, b, \beta)$. Then for some positive constants c_ϵ and C_ϵ that depend only on ϵ, for any sufficiently small constant $\epsilon_0 \in (0, c_\epsilon)$, if we replace the definition of t_u's in* (11) *with*

$$t_u = \left(\frac{1}{2}\log\frac{\widehat{a}_u(1-\widehat{b}_u/n)}{\widehat{b}_u(1-\widehat{a}_u/n)}\right) \wedge \log\frac{1}{\epsilon_0/2}, \tag{22}$$

then we have

$$\begin{aligned} &\sup_{(B,\sigma)\in\Theta} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma,\widehat{\sigma}) \geq \exp\left(-(1-C_\epsilon\epsilon_0)\frac{nI^*}{2}\right)\right\} \to 0, \qquad \text{if } k = 2, \\ &\sup_{(B,\sigma)\in\Theta} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma,\widehat{\sigma}) \geq \exp\left(-(1-C_\epsilon\epsilon_0)\frac{nI^*}{\beta k}\right)\right\} \to 0, \qquad \text{if } k \geq 3, \end{aligned} \tag{23}$$

where I^* is defined as in (14)*. In particular, we can set $C_\epsilon = \frac{10}{3}\frac{2-\epsilon}{\frac{\epsilon}{2}\log\frac{2}{\epsilon}}$ and $c_\epsilon = \min(\frac{1}{10C_\epsilon}, \frac{\epsilon}{2-\epsilon})$.*

If in addition Condition 1 is satisfied for γ satisfying both (16) *and* (18) *and $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$, then the same conclusion holds for $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$.*

Compared with the conclusion (17) in Theorem 4, the vanish sequence η in the exponent of the upper bound is replaced by $C_\epsilon\epsilon_0$, which is guaranteed to be smaller than $\min(0.1, \frac{2}{\log(2/\epsilon)})$ and can be driven to be arbitrarily small by decreasing ϵ_0. To achieve this, the t_u's used in defining the penalty parameters in the penalized neighbor voting step need to be truncated at the value $\log\frac{1}{\epsilon_0/2}$.

5.2 Implications of the results

We now discuss some implications of the results in Theorems 4 – 12.

When using USC as initialization for Algorithm 1, we obtain the following results by combining Theorem 4, Theorem 6 and Theorem 12. Recall that $\bar{d}$ is the average degree of nodes in A defined in (19).

Theorem 13 *Consider Algorithm 1 initialized by σ^0 with USC(τ) with $\tau = C\bar{d}$ for some sufficiently large constant $C > 0$. If as $n \to \infty$, $a \asymp b$ and*

$$\frac{(a-b)^2}{ak^3\log k} \to \infty, \tag{24}$$

then there is a sequence $\eta \to 0$ such that (17) *holds with $\Theta = \Theta_0(n, k, a, b, \beta)$. If as $n \to \infty$, $a \asymp b$ and*

$$\frac{\lambda^2}{ak(\log k + a/(a-b))} \to \infty, \tag{25}$$

then (17) *holds for $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$. If for either parameter space, $a \asymp b$ may not hold but k is fixed and* (24) *or* (25) *holds respectively, then* (23) *holds as long as t_u is replaced by (22) in Algorithm 1.*

Compared with Theorem 1, the minimax optimal performance is achieved under mild conditions. Take $\Theta = \Theta_0(n, k, a, b, \beta)$ for example. For any fixed k, the minimax optimal misclassification proportion is achieved with high probability only under the additional condition of $a \asymp b$. In addition, weak consistency is achieved for fixed k as long as $\frac{(a-b)^2}{a} \to \infty$, regardless of the behavior of $\frac{a}{b}$. This condition is indeed necessary and sufficient for weak consistency. See, for instance, Mossel et al. (2012, 2013b); Yun and Proutiere (2014b); Zhang and Zhou (2015). To achieve strong consistency for fixed k, it suffices to ensure $\ell(\sigma, \widehat{\sigma}) < \frac{1}{n}$ and Theorem 13 implies that it is sufficient to have

$$\liminf_{n\to\infty} \frac{nI^*}{2\log n} > 1, \quad \text{when } k = 2; \quad \liminf_{n\to\infty} \frac{nI^*}{\beta k \log n} > 1, \quad \text{when } k \geq 3, \tag{26}$$

regardless of the behavior of $\frac{a}{b}$. On the other hand, Theorem 1 shows that it is impossible to achieve strong consistency if

$$\limsup_{n\to\infty} \frac{nI^*}{2\log n} < 1, \quad \text{when } k = 2; \quad \limsup_{n\to\infty} \frac{nI^*}{\beta k \log n} < 1, \quad \text{when } k \geq 3. \tag{27}$$

When $\frac{a}{n} = o(1)$, $nI^* = (1 + o(1))(\sqrt{a} - \sqrt{b})^2$ and so one can replace nI^* in (26) – (27) with $(\sqrt{a} - \sqrt{b})^2$. In the literature, Abbe et al. (2014) and Mossel et al. (2014) obtained comparable strong consistency results via efficient algorithms for the special case of two communities of equal sizes, i.e., $k = 2$ and $\beta = 1$. Abbe and Sandon (2015a) investigated the case of fixed k and $\beta \geq 1$. Their results give necessary and sufficient conditions for each instance of model parameters. In comparison, our result characterizes minimax optimality through worst case analysis and is less general than those of Abbe and Sandon (2015a). On the other hand, compared with Abbe and Sandon (2015a), we allow any fixed k and any $\beta \geq 1$ without assuming $a \asymp b \asymp \log n$. In the weak consistency regime, in terms of misclassification proportion, for the special case of $k = 2$ and $\beta = 1$, Yun and Proutiere (2014a) achieved the optimal rate for $\Theta_0(n, 2, a, b, 1)$ when $a \asymp b \asymp a - b$, while the error bounds in other papers are typically off by a constant multiplier on the exponent. In comparison, Theorem 13 provides optimal results (17) and near optimal results (23) for a much broader class of models under much weaker conditions. Last but not least, our algorithm can provably achieve strong consistency and minimax optimal performance even for growing k, which to our limited knowledge, is the first in the literature.

The performance of Algorithm 1 initialized by NSC can be summarized as the following theorem by combining Theorem 4, Theorem 10 and Theorem 12. In this case, the sufficient condition for achieving minimax optimal performance is slightly stronger than when USC is used for initialization.

Theorem 14 *Consider Algorithm 1 initialized by σ^0 with NSC(τ) with $\tau = C\bar{d}$ for some sufficiently large constant $C > 0$. If as $n \to \infty$, $a \asymp b$ and*

$$\frac{(a-b)^2}{ak^3 \log k \log a} \to \infty, \tag{28}$$

then there is a sequence $\eta \to 0$ such that (17) holds with $\Theta = \Theta_0(n, k, a, b, \beta)$. If as $n \to \infty$, $a \asymp b$ and

$$\frac{\lambda^2}{ak \log a(\log k + a/(a-b))} \to \infty, \tag{29}$$

then (17) *holds for* $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$. *If for either parameter space,* $a \asymp b$ *may not hold but* k *is fixed and* (28) *or* (29) *holds respectively, then* (23) *holds as long as* t_u *is replaced by (22) in Algorithm 1.*

Last but not least, we would like to point out that when the key parameters a and b are known, we can obtain the desired performance guarantee under weaker conditions as summarized in the following theorem.

Theorem 15 (The case of known a, b**)** *Suppose* a, b *are known. Consider Algorithm 1 initialized by* σ^0 *with USC(*τ*) with* $\tau = Ca$ *for some sufficiently large constant* $C > 0$ *and* $\widehat{a}_u = a$, $\widehat{b}_u = b$ *in* (9) *for all* $u \in [n]$. *If as* $n \to \infty$, $a \asymp b$ *and*

$$\frac{(a-b)^2}{ak^3} \to \infty, \tag{30}$$

then there is a sequence $\eta \to 0$ *such that* (17) *holds with* $\Theta = \Theta_0(n, k, a, b, \beta)$. *If as* $n \to \infty$, $a \asymp b$ *and*

$$\frac{\lambda^2}{ak} \to \infty, \tag{31}$$

then (17) *holds with* $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$. *If for either parameter space without assuming* $a \asymp b$, (30) *or* (31) *holds respectively, then* (23) *holds if in addition* t_u *is replaced by* (22).

If instead NSC(τ*) is used for initialization with* $\tau = Ca$ *for some sufficiently large constant* $C > 0$, *then the above conclusions hold if we replace* (30) *with* $\frac{(a-b)^2}{ak^3 \log a} \to \infty$ *and* (31) *with* $\frac{\lambda^2}{ak \log a} \to \infty$, *respectively.*

6. Proofs of main results

The main result of the paper, Theorem 4, is proved in Section 6.1. Theorem 6 and Theorem 10 are proved in Section 6.2 and Section 6.3 respectively. The proofs of the remaining results, together with some auxiliary lemmas, are given in the appendix.

6.1 Proof of Theorem 4

We first state a lemma that guarantees the accuracy of parameter estimation in Algorithm 1.

Lemma 16 *Let* $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$. *Suppose as* $n \to \infty$, $\frac{(a-b)^2}{ak} \to \infty$ *and Condition 1 holds with* γ *satisfying* (16) *and* (18). *Then there is a sequence* $\eta \to 0$ *as* $n \to \infty$ *and a constant* $C > 0$ *such that*

$$\min_{u \in [n]} \inf_{(B,\sigma) \in \Theta} \mathbb{P}\left\{ \min_{\pi \in S_k} \max_{i,j \in [k]} |\widehat{B}^u_{ij} - B_{\pi(i)\pi(j)}| \le \eta \left(\frac{a-b}{n} \right) \right\} \ge 1 - Cn^{-(1+\delta)}. \tag{32}$$

For $\Theta = \Theta_0(n, k, a, b, \beta)$, *the conclusion* (32) *continues to hold even when the assumption* (18) *is dropped.*

Proof 1° Let $\Theta = \Theta(n, k, a, b, \lambda, \beta; \alpha)$. For any community assignments σ_1 and σ_2, define

$$\ell_0(\sigma_1, \sigma_2) = \frac{1}{n} \sum_{u=1}^{n} \mathbf{1}_{\{\sigma_1(u) \neq \sigma_2(u)\}}. \tag{33}$$

Fix any $(B, \sigma) \in \Theta$ and $u \in [n]$. Define event

$$E_u = \left\{ \ell_0(\pi_u(\sigma), \sigma_u^0) \leq \gamma \right\}. \tag{34}$$

To simplify notation, assume that $\pi_u = \mathrm{Id}$ is the identity permutation.

Fix any $i \in [k]$. On E_u,

$$n_i \geq |\widetilde{\mathcal{C}}_i^u \cap \mathcal{C}_i| \geq n_i - \gamma_1 n, \ |\widetilde{\mathcal{C}}_i^u \cap \mathcal{C}_i^c| \leq \gamma_2 n, \quad \text{where } \gamma_1, \gamma_2 \geq 0 \text{ and } \gamma_1 + \gamma_2 \leq \gamma. \tag{35}$$

Let $\mathcal{C}_i'$ be any deterministic subset of $[n]$ such that (35) holds with $\widetilde{\mathcal{C}}_i^u$ replaced by $\mathcal{C}_i'$. By definition, there are at most

$$\begin{aligned}\sum_{l=0}^{\gamma n} \binom{n_i}{l} \sum_{m=0}^{\gamma n} \binom{n - n_i}{m} &\leq (\gamma n + 1)^2 \left(\frac{e n_i}{\gamma n}\right)^{\gamma n} \left(\frac{en}{\gamma n}\right)^{\gamma n} \leq \exp\left\{2 \log(\gamma n + 1) + 2\gamma n \log \frac{e}{\gamma}\right\} \\ &\leq \exp\left\{C_1 \gamma n \log \frac{1}{\gamma}\right\}\end{aligned}$$

different subsets with this property where $C_1 > 0$ is an absolute constant. Let $\mathcal{E}_i'$ be the edges within $\mathcal{C}_i'$. Then $|\mathcal{E}_i'|$ consists of independent Bernoulli random variables, where at least $(1 - \beta\gamma k)^2$ proportion of them follow the $\mathrm{Bern}(B_{ii})$ distribution, at most $(\beta\gamma k)^2$ proportion that are stochastically smaller than $\mathrm{Bern}(\frac{\alpha a}{n})$ and stochastically larger than $\mathrm{Bern}(\frac{a}{n})$, and at most $2\beta\gamma k$ proportion are stochastically smaller than $\mathrm{Bern}(\frac{b}{n})$. Therefore, we obtain that

$$(1 - \beta\gamma k)^2 B_{ii} + (\beta\gamma k)^2 \frac{a}{n} \leq \mathbb{E}\left[\frac{|\mathcal{E}_i'|}{\frac{1}{2}|\mathcal{C}_i'|(|\mathcal{C}_i'| - 1)}\right] \leq \max_{t \in [0, \beta\gamma k]} \left\{(1 - t)^2 B_{ii} + t^2 \frac{\alpha a}{n} + 2t \frac{b}{n}\right\}. \tag{36}$$

Note that the LHS is $(1 - (2 + o(1))\beta\gamma k) B_{ii}$. On the other hand, under condition (18), the RHS is attained at $t = 0$ and equals B_{ii} exactly. Thus, we conclude that

$$\left| \mathbb{E}\left[\frac{|\mathcal{E}_i'|}{\frac{1}{2}|\mathcal{C}_i'|(|\mathcal{C}_i'| - 1)}\right] - B_{ii} \right| \leq C\beta\gamma k \frac{\alpha a}{n} = \eta'\left(\frac{a - b}{n}\right) \tag{37}$$

for some $\eta' \to 0$ that depends only on a, k, α, β and γ, where the last inequality is due to (18).

On the other hand, by Bernstein's inequality, for any $t > 0$,

$$\mathbb{P}\left\{\left||\mathcal{E}_i'| - \mathbb{E}|\mathcal{E}_i'|\right| > t\right\} \leq 2 \exp\left\{-\frac{t^2}{2(\frac{1}{2}(n_i + \gamma n)^2 \frac{\alpha a}{n} + \frac{2}{3}t)}\right\}.$$

Let

$$t^2 = (n_i+\gamma n)^2\frac{\alpha a}{n}(C_1\gamma n\log\gamma^{-1}+(3+\delta)\log n)\vee(2C_1\gamma n\log\gamma^{-1}+2(3+\delta)\log n)^2$$
$$\lesssim \left(\frac{n}{k}\sqrt{a\gamma\log\gamma^{-1}}+\gamma n\log\gamma^{-1}\right)^2,$$

where we the second inequality holds since $\frac{\log x}{x}$ is monotone decreasing as x increases and so $\gamma\log\gamma^{-1}\geq\frac{1}{n}\log n$ for any $\gamma\geq\frac{1}{n}$, which is the case of most interest since $\gamma<\frac{1}{n}$ leads to $\gamma=0$ and so the initialization is already perfect. Even when $\gamma=0$, we can still continue to the following arguments by replacing every γ with $\frac{1}{n}$ and all the steps continue to hold. Thus, we obtain that for positive constant $C_{\alpha,\beta,\delta}$ that depends only on α,β and δ,

$$\mathbb{P}\left\{\left||\mathcal{E}_i'|-\mathbb{E}|\mathcal{E}_i'|\right|>C_{\alpha,\beta,\delta}\left(\frac{n}{k}\sqrt{a\gamma\log\gamma^{-1}}+\gamma n\log\gamma^{-1}\right)\right\}\leq\exp\left\{-C_1\gamma n\log\gamma^{-1}\right\}n^{-(3+\delta)}. \tag{38}$$

Thus, with probability at least $1-\exp\left\{-C_1\gamma n\log\gamma^{-1}\right\}n^{-(3+\delta)}$,

$$\left|\frac{|\mathcal{E}_i'|}{\frac{1}{2}|\mathcal{C}_i'|(|\mathcal{C}_i'|-1)}-\mathbb{E}\frac{|\mathcal{E}_i'|}{\frac{1}{2}|\mathcal{C}_i'|(|\mathcal{C}_i'|-1)}\right|\leq C_{\alpha,\beta,\delta}\left(\frac{k}{n}\sqrt{a\gamma\log\gamma^{-1}}+\frac{k^2\gamma\log\gamma^{-1}}{n}\right)=\eta'\left(\frac{a-b}{n}\right), \tag{39}$$

where $\eta'\to0$ depends only on a,k,α,β,γ and δ. Here, the last inequality holds since

$$k\sqrt{a\gamma\log\gamma^{-1}}=\sqrt{ak}\sqrt{k\gamma\log\gamma^{-1}},$$

where $\sqrt{ak}\ll a-b$ since $\frac{(a-b)^2}{ak}\to\infty$ and $k\gamma\log\gamma^{-1}=O(1)$, and

$$k^2\gamma\log\gamma^{-1}=k\gamma\log\gamma^{-1}\cdot k\lesssim k\ll\frac{(a-b)^2}{a}\lesssim a-b.$$

We combine (37) and (39) and apply the union bound to obtain that for a sequence $\eta\to0$ that depends only on a,k,α,β,γ and δ, with probability at least $1-n^{-(3+\delta)}$

$$\left|\frac{|\widetilde{\mathcal{E}}_i^u|}{\frac{1}{2}|\widetilde{\mathcal{C}}_i^u|(|\widetilde{\mathcal{C}}_i^u|-1)}-B_{ii}\right|\leq\eta\left(\frac{a-b}{n}\right). \tag{40}$$

The proof for B_{ij} estimation is analogous and hence is omitted. A final union bound on $i,j\in[k]$ leads to the desired claim since all the constants and vanishing sequences in the above analysis depend only on $a,b,k,\alpha,\beta,\gamma$ and δ, but not on u, B or σ.

2° If $\Theta=\Theta_0(n,k,a,b,\beta)$, then condition (18) on γ is no longer needed. This is because (36) can be replaced by

$$\min_{t\in[0,\beta\gamma k]}\left\{(1-t)^2\frac{a}{n}+2t(1-t)\frac{b}{n}+t^2\frac{b}{n}\right\}$$
$$\leq\mathbb{E}\left[\frac{|\mathcal{E}_i'|}{\frac{1}{2}|\mathcal{C}_i'|(|\mathcal{C}_i'|-1)}\right]\leq\max_{t\in[0,\beta\gamma k]}\left\{(1-t)^2\frac{a}{n}+t^2\frac{a}{n}+2t(1-t)\frac{b}{n}\right\},$$

where the LHS equals $\frac{a}{n} - (1 - \beta\gamma k(1+o(1)))\frac{a-b}{n} = \frac{a}{n} + o(\frac{a-b}{n})$ and the RHS equals $\frac{a}{n}$. Thus, no additional condition is needed to guarantee (37) in the foregoing arguments. This completes the proof. ■

The next two lemmas establish the desired error bound for the node-wise refinement.

Lemma 17 *Let Θ_0 be defined as in (2) and $k \geq 2$. Suppose as $n \to \infty$, $\frac{(a-b)^2}{ak} \to \infty$ and $a \asymp b$. If there exists two sequences $\gamma = o(1/k)$ and $\eta = o(1)$, constants $C, \delta > 0$ and permutations $\{\pi_u\}_{u=1}^n \subset S_k$ such that*

$$\inf_{(B,\sigma)\in\Theta_0} \min_{u\in[n]} \mathbb{P}\left\{\ell_0(\pi_u(\sigma), \sigma_u^0) \leq \gamma,\ |\widehat{a}_u - a| \leq \eta(a-b),\ |\widehat{b}_u - b| \leq \eta(a-b)\right\} \geq 1 - Cn^{-(1+\delta)}. \tag{41}$$

Then for $\widehat{\sigma}_u(u)$ defined as in (10) with $\rho = \rho_u$ in (12), there is a sequence $\eta' = o(1)$ such that for $k = 2$,

$$\sup_{(B,\sigma)\in\Theta_0} \max_{u\in[n]} \mathbb{P}\left\{\widehat{\sigma}_u(u) \neq \pi_u(\sigma(u))\right\} \leq (k-1)\exp\left\{-(1-\eta')\frac{nI^*}{2}\right\} + Cn^{-(1+\delta)},$$

and for $k \geq 3$,

$$\sup_{(B,\sigma)\in\Theta_0} \max_{u\in[n]} \mathbb{P}\left\{\widehat{\sigma}_u(u) \neq \pi_u(\sigma(u))\right\} \leq (k-1)\exp\left\{-(1-\eta')\frac{nI^*}{\beta k}\right\} + Cn^{-(1+\delta)}.$$

Proof In what follows, let E_u denote the event in (41). For the sake of brevity, we let $p = a/n$, $q = b/n$, $\widehat{p}_u = \widehat{a}_u/n$ and $\widehat{q}_u = \widehat{b}_u/n$. Moreover, let $\sigma_u = \pi_u(\sigma)$, $n_i = |\{v : \sigma_u(v) = i\}|$, $m_i = |\{v : \sigma_u^0(v) = i\}|$ and $m_i' = |\{v : \sigma_u^0(v) = \sigma_u(v) = i\}|$. Without loss of generality, let $\sigma_u(u) = 1$.

Then we have

$$\mathbb{P}\left\{\widehat{\sigma}_u(u) \neq 1 \text{ and } E_u\right\} \leq \sum_{l\neq 1} \mathbb{P}\left\{E_u \text{ and } \sum_{\sigma_u(v)=l} A_{uv} - \sum_{\sigma_u(v)=1} A_{uv} \geq \rho_u(m_l - m_1)\right\} = \sum_{l\neq 1} p_l. \tag{42}$$

Now we bound each p_l. By the independence structure and Chernoff bound, we have

$$\begin{aligned} p_l &\leq \mathbb{E}\Big\{\exp\left(-t_u\rho_u(m_l - m_1)\right)(qe^{t_u} + 1 - q)^{m_l'}(pe^{t_u} + 1 - p)^{m_l - m_l'} \\ &\qquad (pe^{-t_u} + 1 - p)^{m_1'}(qe^{-t_u} + 1 - q)^{m_1 - m_1'}\mathbf{1}_{\{E_u\}}\Big\} \qquad (43) \\ &\leq \mathbb{E}\left\{\exp\left(-t_u\rho_u(m_l - m_1)\right)(qe^{t_u} + 1 - q)^{m_l}(pe^{-t_u} + 1 - p)^{m_1}\mathbf{1}_{\{E_u\}}\right\} \qquad (44) \\ &\quad \times\mathbb{E}\left\{\left(\frac{pe^{t_u} + 1 - p}{qe^{t_u} + 1 - q}\right)^{m_l - m_l'}\left(\frac{qe^{-t_u} + 1 - q}{pe^{-t_u} + 1 - p}\right)^{m_1 - m_1'}\mathbf{1}_{\{E_u\}}\right\}. \qquad (45) \end{aligned}$$

We are going to give bounds for the terms in (44) and (45) respectively. Before doing that, we need some preparatory inequalities. Define t^* through the equation

$$e^{t^*} = \sqrt{\frac{p(1-q)}{q(1-p)}}.$$

Then, on the event E_u,

$$e^{t_u - t^*} + e^{t^* - t_u} \le \exp(C_1 \eta), \tag{46}$$

for some constant $C_1 > 0$. Moreover,

$$|e^{t_u} - 1| \vee |e^{-t_u} - 1| \le C_2 \frac{p-q}{p} = C_2 \frac{a-b}{a}, \tag{47}$$

for some constant $C_2 > 0$. Therefore, for the term in (44), on the event E_u,

$$\begin{aligned} & \exp\left(-t_u \rho_u (m_l - m_1)\right) (qe^{t_u} + 1 - q)^{m_l} (pe^{-t_u} + 1 - p)^{m_1} \\ = \; & \exp\left(-t_u \rho_u (m_l - m_1)\right) \left(qe^{t_u} + 1 - q\right)^{(m_l - m_1)/2} \left(pe^{-t_u} + 1 - p\right)^{(m_1 - m_l)/2} \qquad (48) \\ & \times \left(qe^{t_u} + 1 - q\right)^{(m_1 + m_l)/2} \left(pe^{-t_u} + 1 - p\right)^{(m_1 + m_l)/2}. \qquad (49) \end{aligned}$$

By (46), the term in (49) is upper bounded by

$$\begin{aligned} & \left(pq + (1-p)(1-q) + \sqrt{pq}\sqrt{(1-p)(1-q)}(e^{t_u - t^*} + e^{t^* - t_u})\right)^{\frac{m_1 + m_l}{2}} \\ & \le \exp\left(-(1+o(1))\frac{m_1 + m_l}{2} I^*\right) \le \exp\left(-(1+o(1))\frac{n_1 + n_l}{2} I^*\right). \end{aligned}$$

By (47), the term in (48) is upper bounded by

$$\begin{aligned} & \exp\left(-t_u \rho_u (m_l - m_1)\right) \left(qe^{t_u} + 1 - q\right)^{(m_l - m_1)/2} \left(pe^{-t_u} + 1 - p\right)^{(m_1 - m_l)/2} \\ = \; & \exp\left(\frac{m_1 - m_l}{2}\left(\log \frac{pe^{-t_u} + 1 - p}{qe^{t_u} + 1 - q} - \log \frac{\widehat{p}_u e^{-t_u} + 1 - \widehat{p}_u}{\widehat{q}_u e^{t_u} + 1 - \widehat{q}_u}\right)\right) \\ \le \; & \exp\left(\frac{|m_1 - m_l|}{2}\left(|e^{-t_u} - 1||\widehat{p}_u - p| + |e^{t_u} - 1||\widehat{q}_u - q|\right)\right) \\ \le \; & \exp\left(o\left(\frac{n}{k}\frac{(p-q)^2}{p}\right)\right) \\ = \; & \exp\left(o(1)\frac{n_1 + n_l}{2} I^*\right). \end{aligned}$$

Therefore, we can upper bound (44) as

$$\mathbb{E}\left\{e^{-t_u \rho_u (m_l - m_1)} (qe^{t_u} + 1 - q)^{m_l} (pe^{-t_u} + 1 - p)^{m_1} \mathbf{1}_{\{E_u\}}\right\} \le \exp\left(-(1+o(1))\frac{n_1 + n_l}{2} I^*\right). \tag{50}$$

Now we provide an upper bound for (45). By (47), on E_u,

$$\frac{pe^{t_u} + 1 - p}{qe^{t_u} + 1 - q} = 1 + \frac{(p-q)(e^{t_u} - 1)}{qe^{t_u} + 1 - q} \le 1 + O\left(\frac{(p-q)^2}{p}\right) \le \exp\left(O\left(\frac{(p-q)^2}{p}\right)\right),$$

and

$$\frac{qe^{-t_u} + 1 - q}{pe^{-t_u} + 1 - p} = 1 + \frac{(p-q)(1 - e^{-t_u})}{pe^{-t_u} + 1 - p} \le 1 + O\left(\frac{(p-q)^2}{p}\right) \le \exp\left(O\left(\frac{(p-q)^2}{p}\right)\right).$$

Therefore,

$$\mathbb{E}\left\{\left(\frac{pe^{t_u}+1-p}{qe^{t_u}+1-q}\right)^{m_l-m_l'}\left(\frac{qe^{-t_u}+1-q}{pe^{-t_u}+1-p}\right)^{m_1-m_1'}\mathbf{1}_{\{E_u\}}\right\}\leq\exp\left(o(1)\frac{n_1+n_l}{2}I^*\right). \quad (51)$$

By combining (50) and (51), we have

$$p_l\leq\exp\left(-(1+o(1))\frac{n_1+n_l}{2}I^*\right). \quad (52)$$

Using (42), this implies

$$\mathbb{P}\left\{\widehat{\sigma}_u(u)\neq 1 \text{ and } E_u\right\}\leq(k-1)\exp\left(-(1+o(1))\min_{l\neq 1}\left(\frac{n_1+n_l}{2}\right)I^*\right),$$

and so

$$\mathbb{P}\left\{\widehat{\sigma}_u(u)\neq 1\right\}\leq(k-1)\exp\left(-(1+o(1))\min_{l\neq 1}\left(\frac{n_1+n_l}{2}\right)I^*\right)+Cn^{-(1+\delta)}.$$

When $k=2$, $\min_{l\neq 1}\left(\frac{n_1+n_l}{2}\right)=\frac{n}{2}$, and when $k\geq 3$, $\min_{l\neq 1}\left(\frac{n_1+n_l}{2}\right)\geq\frac{n}{\beta k}$. Thus, the proof is complete. ■

Lemma 18 *Let Θ be defined as in (3) and $k\geq 2$. Suppose as $n\to\infty$, $\frac{(a-b)^2}{ak}\to\infty$ and $a\asymp b$. If there exists two sequences $\gamma=o\left(\frac{a-b}{ak}\right)$ and $\eta=o(1)$, constants $C,\delta>0$ and permutations $\{\pi_u\}_{u=1}^n\subset S_k$ such that (41) holds. Then for $\widehat{\sigma}_u(u)$ defined as in (10) with $\rho=\rho_u$ in (12), the conclusions of Lemma 17 continue to hold.*

Proof The proof is similar to that of Lemma 17 and we use the same notation as there. First, we give a bound for p_l defined in (42). Let $X_j\sim\text{Bern}(q)$, $Y_j\sim\text{Bern}(p)$ and $Z_j\sim\text{Bern}(\alpha p)$, $j\geq 1$, be mutually independent. Then, a stochastic order argument gives

$$\begin{aligned}
p_l &\leq \mathbb{E}\left[\mathbb{P}\left\{\sum_{j=1}^{m_l'}X_j+\sum_{j=1}^{m_l-m_l'}Z_j-\sum_{j=1}^{m_1'}Y_j\geq\rho(m_l-m_1)\text{ and }E_u\Big|A_{-u}\right\}\right]\\
&\leq \mathbb{E}\left\{\exp\left(-t_u\rho_u(m_l-m_1)\right)(qe^{t_u}+1-q)^{m_l}(pe^{-t_u}+1-p)^{m_1}\mathbf{1}_{\{E_u\}}\right\} \quad (53)\\
&\quad\times\mathbb{E}\left\{\left(\frac{1}{qe^{t_u}+1-q}\right)^{m_l-m_l'}\left(\frac{1}{pe^{-t_u}+1-p}\right)^{m_1-m_1'}\right. \quad (54)\\
&\quad\left.(\alpha pe^{t_u}+1-\alpha p)^{m_l-m_l'}\mathbf{1}_{\{E_u\}}\right\}.
\end{aligned}$$

Note that the term in (53) is the same as that in (44), and thus it can be upper bounded by (50) as before. To bound for (54), observe that by (47),

$$\frac{1}{qe^{t_u}+1-q}\leq\exp\left(q|e^{t_u}-1|\right)\leq\exp\left(O(p-q)\right),$$

$$\frac{1}{pe^{-t_u}+1-p} \le \exp\left(Cp|e^{-t_u}-1|\right) \le \exp\left(O(p-q)\right)$$

and

$$\alpha p e^{t_u} + 1 - \alpha p \le \exp\left(\alpha p|e^{t_u}-1|\right) \le \exp\left(O(p-q)\right).$$

Thus, under the assumption $\gamma = o\left(\frac{p-q}{kp}\right)$, the term (54) is bounded by $\exp\left(o(1)\frac{n_1+n_l}{2}I^*\right)$. The remaining proof is the same as that of Lemma 17. ■

Finally, we need a lemma to justify the consensus step in Algorithm 1.

Lemma 19 *For any community assignments σ and σ': $[n] \to [k]$, such that for some constant $C \ge 1$*

$$\min_{l\in[k]} |\{u:\sigma(u)=l\}|, \min_{l\in[k]} |\{u:\sigma'(u)=l\}| \ge \frac{n}{Ck}, \quad \textit{and} \quad \min_{\pi\in S_k} \ell_0(\sigma,\pi(\sigma')) < \frac{1}{Ck}.$$

Define map $\xi : [k] \to [k]$ as

$$\xi(i) = \operatorname*{argmax}_l \left|\{u:\sigma(u)=l\} \cap \{u:\sigma'(u)=i\}\right|, \quad \forall i \in [k]. \tag{55}$$

Then $\xi \in S_k$ and $\ell_0(\sigma,\xi(\sigma')) = \min_{\pi\in S_k} \ell_0(\sigma,\pi(\sigma'))$.

Proof By the definition in (55), we obtain

$$\xi = \operatorname*{argmin}_{\xi':[k]\to[k]} \ell_0(\sigma,\xi'(\sigma')), \quad \text{and} \quad \ell_0(\sigma,\xi(\sigma')) \le \min_{\pi\in S_k} \ell_0(\sigma,\pi(\sigma')) < \frac{1}{Ck}.$$

Thus, what remains to be shown is that $\xi \in S_k$, i.e., $\xi(l_1) \ne \xi(l_2)$ for any $l_1 \ne l_2$. To this end, note that if for some $l_1 \ne l_2$, $\xi(l_1) = \xi(l_2)$, then there would exist some $l_0 \in [k]$ such that for any $l \in [k]$, $\xi(l) \ne l_0$, and so

$$\ell_0(\sigma,\xi(\sigma')) \ge \frac{1}{n}\sum_{u:\sigma(u)=l_0} \mathbf{1}_{\{\sigma(u)\ne\xi(\sigma'(u))\}} = \frac{|\{u:\sigma(u)=l_0\}|}{n} \ge \frac{1}{Ck}.$$

This is in contradiction to the second last display, and hence $\xi \in S_k$. This completes the proof. ■

Proof [Proof of Theorem 4] Let $\Theta = \Theta(n,k,a,b,\lambda,\beta;\alpha)$, and fix any $(B,\sigma) \in \Theta$. For any $u \in [n]$, by Condition 1 and the fact that σ_u^0 and $\widehat{\sigma}_u$ differ only at the community assignment of u, for $\gamma' = \gamma + 1/n$, there exists some $\pi_u \in S_k$ such that

$$\mathbb{P}\left\{\ell_0(\sigma,\pi_u^{-1}(\widehat{\sigma}_u)) \le \gamma_n'\right\} \ge 1 - C_0 n^{-(1+\delta)}. \tag{56}$$

Without loss of generality, we assume $\pi_1 = \mathrm{Id}$ is the identity map. Now for any fixed $u \in \{2,\dots,n\}$, define map $\xi_u : [k] \to [k]$ as in (55) with σ and σ' replaced by $\widehat{\sigma}_1$ and $\widehat{\sigma}_u$. Then by definition

$$\widehat{\sigma}(u) = \xi_u(\widehat{\sigma}_u(u)). \tag{57}$$

In addition, (56) implies with probability at least $1 - Cn^{-(1+\delta)}$, we have

$$\ell_0(\sigma, \widehat{\sigma}_1) \le \gamma' \quad \text{and} \quad \ell_0(\sigma, \pi_u^{-1}(\widehat{\sigma}_u)) \le \gamma'.$$

So the triangle inequality implies $\ell_0(\widehat{\sigma}_1, \pi_u^{-1}(\widehat{\sigma}_u)) \le 2\gamma'$ and hence the condition of Lemma 19 is satisfied. Thus, Lemma 19 implies

$$\mathbb{P}\left\{\xi_u = \pi_u^{-1}\right\} \ge 1 - Cn^{-(1+\delta)}. \tag{58}$$

When $k \ge 3$, Lemma 16, (16) and (18) imply that the condition of Lemma 18 is satisfied, which in turn implies that for a sequence $\eta' = o(1)$,

$$\begin{aligned}
\mathbb{P}\left\{\widehat{\sigma}(u) \ne \sigma(u)\right\} &= \mathbb{P}\left\{\xi_u(\widehat{\sigma}_u(u)) \ne \sigma(u)\right\} \\
&\le \mathbb{P}\left\{\xi_u(\widehat{\sigma}_u(u)) \ne \sigma(u),\, \xi_u = \pi_u^{-1}\right\} + \mathbb{P}\left\{\xi_u \ne \pi_u^{-1}\right\} \\
&\le \mathbb{P}\left\{\widehat{\sigma}_u(u) \ne \pi_u(\sigma(u))\right\} + \mathbb{P}\left\{\xi_u \ne \pi_u^{-1}\right\} \\
&\le Cn^{-(1+\delta)} + (k-1)\exp\left\{-(1-\eta')\frac{nI^*}{\beta k}\right\}.
\end{aligned}$$

Set

$$\eta = \eta' + \beta\sqrt{\frac{k}{nI^*}} = o(1) \tag{59}$$

where the last inequality holds since $\frac{nI^*}{k} \asymp \frac{(a-b)^2}{ak} \to \infty$. Thus, Markov's inequality leads to

$$\begin{aligned}
&\mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > (k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\}\right\} \\
&\qquad \le \frac{1}{(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\}} \frac{1}{n}\sum_{u=1}^{n} \mathbb{P}\left\{\widehat{\sigma}(u) \ne \sigma(u)\right\} \\
&\qquad \le \exp\left\{-(\eta - \eta')\frac{nI^*}{\beta k}\right\} + \frac{Cn^{-(1+\delta)}}{(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\}} \\
&\qquad \le \exp\left\{-\sqrt{\frac{nI^*}{k}}\right\} + \frac{Cn^{-(1+\delta)}}{(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\}}.
\end{aligned}$$

If $(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\} \ge n^{-(1+\delta/2)}$, then

$$\mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > (k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\}\right\} \le \exp\left\{-\sqrt{\frac{nI^*}{k}}\right\} + Cn^{-\delta/2} = o(1).$$

If $(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\} < n^{-(1+\delta/2)}$, then

$$\begin{aligned}
\mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > (k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\}\right\} &= \mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > 0\right\} \le \sum_{u=1}^{n} \mathbb{P}\left\{\widehat{\sigma}(u) \ne \sigma(u)\right\} \\
&\le n(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\} + Cn^{-\delta} \le Cn^{-\delta/2} = o(1).
\end{aligned}$$

Here, the second last inequality holds since $\eta > \eta'$ and so $(k-1)\exp\{-(1-\eta')nI^*/(\beta k)\} < (k-1)\exp\{-(1-\eta)nI^*/(\beta k)\} < n^{-(1+\delta/2)}$. We complete the proof for the case of $\Theta(n,k,a,b,\lambda,\beta;\alpha)$ and $k \geq 3$ by noting that $(k-1)\exp\left\{-(1-\eta)\frac{nI^*}{\beta k}\right\} = \exp\left\{-(1-\eta'')\frac{nI^*}{\beta k}\right\}$ for another sequence $\eta'' = o(1)$ under the assumption $\frac{(a-b)^2}{ak\log k} \to \infty$ and no constant or sequence in the foregoing arguments involves B, σ or u. When $\Theta = \Theta(n,k,a,b,\lambda,\beta;\alpha)$ and $k = 2$, the foregoing arguments continue to hold with β and k replaced with 1 and 2 respectively.

When $\Theta = \Theta_0(n,k,a,b,\beta)$, we can run the foregoing arguments with Lemma 18 replaced by Lemma 17 to reach the conclusion in (17), which does not require condition (18). This completes the proof. ■

6.2 Proof of Theorem 6

The following lemma is critical to establish the result of Theorem 6. Its proof is given in the appendix. Let us introduce the notation $O(k_1,k_2) = \{V \in \mathbb{R}^{k_1\times k_2} : V^TV = I_{k_2}\}$ for $k_1 \geq k_2$.

Lemma 20 *Consider a symmetric adjacency matrix $A \in \{0,1\}^{n\times n}$ and a symmetric matrix $P \in [0,1]^{n\times n}$ satisfying $A_{uu} = 0$ for all $u \in [n]$ and $A_{uv} \sim$ Bernoulli(P_{uv}) independently for all $u > v$. For any $C' > 0$, there exists some $C > 0$ such that*

$$\|T_\tau(A) - P\|_{\mathrm{op}} \leq C\sqrt{np_{\max}+1},$$

with probability at least $1-n^{-C'}$ uniformly over $\tau \in [C_1(np_{\max}+1), C_2(np_{\max}+1)]$ for some sufficiently large constants C_1, C_2, where $p_{\max} = \max_{u\geq v} P_{uv}$.

Lemma 21 *For $P = (P_{uv}) = (B_{\sigma(u)\sigma(v)})$, we have SVD $P = U\Lambda U^T$, where*

$$U = Z\Delta^{-1}W,$$

with $\Delta = diag(\sqrt{n_1}, ..., \sqrt{n_k})$, $Z \in \{0,1\}^{n\times k}$ is a matrix with exactly one nonzero entry in each row at $(i,\sigma(i))$ taking value 1 and $W \in O(k,k)$.

Proof Note that

$$P = ZBZ^T = Z\Delta^{-1}\Delta B\Delta(Z\Delta^{-1})^T,$$

and observe that $Z\Delta^{-1} \in O(n,k)$. Apply SVD to the matrix $\Delta B\Delta^T = W\Lambda W^T$ for some $W \in O(k,k)$, and then we have $P = U\Lambda U^T$ with $U = Z\Delta^{-1}W \in O(k,k)$. ■

Proof [Proof of Theorem 6] Under the current assumption, $\mathbb{E}\tau \in [C_1'a, C_2'a]$ for some large C_1' and C_2'. Using Bernstein's inequality, we have $\tau \in [C_1a, C_2a]$ for some large C_1 and C_2 with probability at least $1-e^{-C'n}$. When (20) holds, by Lemma 20, we deduce that the k^{th} eigenvalue of $T_\tau(A)$ is lower bounded by $c_1\lambda_k$ with probability at least $1-n^{-C'}$ for some small constant $c_1 \in (0,1)$. By Davis–Kahan's sin-theta theorem (Davis and Kahan, 1970),

we have $\|\widehat{U} - UW_1\|_{\mathrm{F}} \leq C\frac{\sqrt{k}}{\lambda_k}\|T_\tau(A) - P\|_{\mathrm{op}}$ for some $W_1 \in O(k,k)$ and some constant $C > 0$. Applying Lemma 21, we have

$$\|\widehat{U} - V\|_{\mathrm{F}} \leq C\frac{\sqrt{k}}{\lambda_k}\|T_\tau(A) - P\|_{\mathrm{op}}, \tag{60}$$

where $V = Z\Delta^{-1}W_2 \in O(n,k)$ for some $W_2 \in O(k,k)$. Combining (60), Lemma 20 and the conclusion $\tau \in [C_1 a, C_2 a]$, we have

$$\|\widehat{U} - V\|_{\mathrm{F}} \leq \frac{C\sqrt{k}\sqrt{a}}{\lambda_k}, \tag{61}$$

with probability at least $1 - n^{-C'}$. The definition of V implies that

$$\|V_{u*} - V_{v*}\| = \sqrt{\frac{1}{n_u} + \frac{1}{n_v}}\mathbf{1}_{\{\sigma(u)\neq\sigma(v)\}}. \tag{62}$$

In other words, define $Q = \Delta^{-1}W_2 \in \mathbb{R}^{k\times k}$ and we have $V_{u*} = Q_{\sigma(u)*}$ for each $u \in [n]$. Hence, for $\sigma(u) \neq \sigma(v)$, $\left\|Q_{\sigma(u)*} - Q_{\sigma(v)*}\right\| = \|V_{u*} - V_{v*}\| \geq \sqrt{\frac{2k}{\beta n}}$. Recall the definition $r = \mu\sqrt{\frac{k}{n}}$ in Algorithm 2. Define the sets

$$T_i = \left\{u \in \sigma^{-1}(i) : \left\|\widehat{U}_{u*} - Q_{i*}\right\| < \frac{r}{2}\right\}, \quad i \in [k].$$

By definition, $T_i \cap T_j = \varnothing$ when $i \neq j$, and we also have

$$\cup_{i\in[k]} T_i = \left\{u \in [n] : \left\|\widehat{U}_{u*} - V_{u*}\right\| < \frac{r}{2}\right\}. \tag{63}$$

Therefore,

$$\left|\left(\cup_{i\in[k]}T_i\right)^c\right|\frac{r^2}{4} \leq \sum_{u\in[n]}\left\|\widehat{U}_{u*} - V_{u*}\right\|^2 \leq \frac{C^2 ka}{\lambda_k^2},$$

where the last inequality is by (61). After rearrangement, we have

$$\left|\left(\cup_{i\in[k]}T_i\right)^c\right| \leq \frac{4C^2 na}{\mu^2\lambda_k^2}. \tag{64}$$

In other words, most nodes are close to the centers and are in the set (63). Note that the sets $\{T_i\}_{i\in[k]}$ are disjoint. Suppose there is some $i \in [k]$ such that $|T_i| < |\sigma^{-1}(i)| - \left|\left(\cup_{i\in[k]}T_i\right)^c\right|$, we have $\left|\cup_{i\in[k]}T_i\right| = \sum_{i\in[k]}|T_i| < n - \left|\left(\cup_{i\in[k]}T_i\right)^c\right| = \left|\cup_{i\in[k]}T_i\right|$, which is impossible. Thus, the cardinality of T_i for each $i \in [k]$ is lower bounded as

$$|T_i| \geq |\sigma^{-1}(i)| - \left|\left(\cup_{i\in[k]}T_i\right)^c\right| \geq \frac{n}{\beta k} - \frac{4C^2 na}{\mu^2\lambda_k^2} > \frac{n}{2\beta k}, \tag{65}$$

where the last inequality above is by the assumption (20). Intuitively speaking, except for a negligible proportion, most data points in $\{\widehat{U}_{u*}\}_{u\in[n]}$ are very close to the population

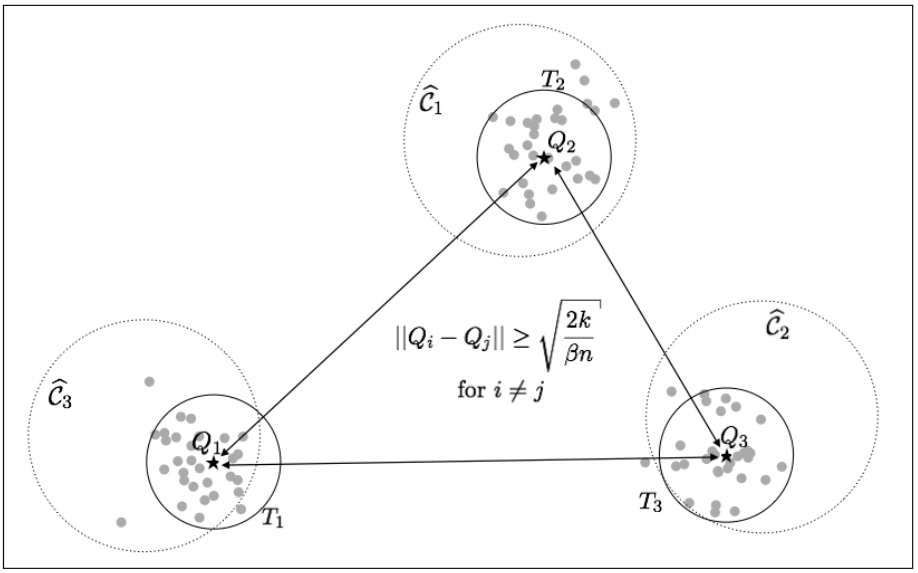

Figure 4: The schematic plot for the proof of Theorem 6. The balls $\{T_i\}_{i\in[k]}$ are centered at $\{Q_i\}_{i\in[k]}$, and the centers are at least $\sqrt{\frac{2k}{\beta n}}$ away from each other. The balls $\{\widehat{\mathcal{C}}_i\}_{i\in[k]}$ intersect with large proportions of $\{T_i\}_{i\in[k]}$, and their subscripts do not need to match due to some permutation.

centers $\{Q_{i*}\}_{i\in[k]}$. Since the centers are at least $\sqrt{\frac{2k}{\beta n}}$ away from each other and $\{T_i\}_{i\in[k]}$ and $\{\widehat{C}_i\}_{i\in[k]}$ are both defined through the critical radius $r = \mu\sqrt{\frac{k}{n}}$ for a small μ, each $\widehat{C}_i$ should intersect with only one T_i (see Figure 4). We claim that there exists some permutation π of the set $[k]$, such that for $\widehat{C}_i$ defined in Algorithm 2,

$$\widehat{\mathcal{C}}_i \cap T_{\pi(i)} \neq \varnothing \quad \text{and} \quad |\widehat{\mathcal{C}}_i| \geq |T_{\pi(i)}| \quad \text{for each } i \in [k]. \tag{66}$$

In what follows, we first establish the result of Theorem 6 by assuming (66). The proof of (66) will be given in the end. Note that for any $i \neq j$, $T_{\pi(i)} \cap \widehat{\mathcal{C}}_j = \varnothing$, which is deduced from the fact that $\widehat{\mathcal{C}}_j \cap T_{\pi(j)} \neq \varnothing$ and the definition of $\widehat{\mathcal{C}}_j$. Therefore, $T_{\pi(i)} \subset \widehat{\mathcal{C}}_j^c$ for all $j \neq i$. Combining with the fact that $T_{\pi(i)} \cap \widehat{\mathcal{C}}_i^c \subset \widehat{\mathcal{C}}_i^c$, we get $T_{\pi(i)} \cap \widehat{\mathcal{C}}_i^c \subset (\cup_{i\in[k]}\widehat{\mathcal{C}}_i)^c$. Therefore,

$$\cup_{i\in[k]} \left(T_{\pi(i)} \cap \widehat{\mathcal{C}}_i^c\right) \subset \left(\cup_{i\in[k]}\widehat{\mathcal{C}}_i\right)^c. \tag{67}$$

Since $T_i \cap T_j = \varnothing$ for $i \neq j$, we deduce from (67) that

$$\sum_{i\in[k]} \left|T_{\pi(i)} \cap \widehat{\mathcal{C}}_i^c\right| \leq \left|\left(\cup_{i\in[k]}\widehat{\mathcal{C}}_i\right)^c\right|. \tag{68}$$

By definition, $\widehat{\mathcal{C}}_i \cap \widehat{\mathcal{C}}_j = \varnothing$ for $i \neq j$, we deduce from (66) that

$$\left|\left(\cup_{i\in[k]}\widehat{\mathcal{C}}_i\right)^c\right| = n - \sum_{i\in[k]} |\widehat{\mathcal{C}}_i| \leq n - \sum_{i\in[k]} |T_i| = \left|\left(\cup_{i\in[k]} T_i\right)^c\right|. \tag{69}$$

Combining (68), (69) and (64), we have

$$\sum_{i\in[k]} \left|T_{\pi(i)} \cap \widehat{\mathcal{C}}_i^c\right| \leq \frac{4C^2na}{\mu^2\lambda_k^2}. \tag{70}$$

Since for any $u \in \cup_{i\in[k]}(\widehat{\mathcal{C}}_i \cap T_{\pi(i)})$, we have $\widehat{\sigma}(u) = i$ when $\sigma(u) = \pi(i)$, the mis-classification rate is bounded as

$$\begin{aligned}
\ell_0(\widehat{\sigma}, \pi^{-1}(\sigma)) &\leq \frac{1}{n}\left|\left(\cup_{i\in[k]}(\widehat{\mathcal{C}}_i \cap T_{\pi(i)})\right)^c\right| \\
&\leq \frac{1}{n}\left(\left|\left(\cup_{i\in[k]}(\widehat{\mathcal{C}}_i \cap T_{\pi(i)})\right)^c \cap \left(\cup_{i\in[k]}T_i\right)\right| + \left|\left(\cup_{i\in[k]}T_i\right)^c\right|\right) \\
&\leq \frac{1}{n}\left(\sum_{i\in[k]}\left|T_{\pi(i)} \cap \widehat{\mathcal{C}}_i^c\right| + \left|\left(\cup_{i\in[k]}T_i\right)^c\right|\right) \\
&\leq \frac{8C^2a}{\mu^2\lambda_k^2},
\end{aligned}$$

where the last inequality is from (70) and (64). This proves the desired conclusion.

Finally, we are going to establish the claim (66) to close the proof. We use mathematical induction. For $i = 1$, it is clear that $|\widehat{\mathcal{C}}_1| \geq \max_{i\in[k]}|T_i|$ holds by the definition of $\widehat{\mathcal{C}}_1$. Suppose $\widehat{\mathcal{C}}_1 \cap T_i = \varnothing$ for all $i \in [k]$, and then we must have

$$\left|\left(\cup_{i\in[k]}T_i\right)^c\right| \geq |\widehat{\mathcal{C}}_1| \geq \max_{i\in[k]}|T_i| \geq \frac{n}{2\beta k},$$

where the last inequality is by (65). This contradicts (64) under the assumption (20). Therefore, there must be a $\pi(1)$ such that $\widehat{\mathcal{C}}_1 \cap T_{\pi(1)} \neq \varnothing$ and $|\widehat{\mathcal{C}}_1| \geq |T_{\pi(1)}|$. Moreover,

$$\begin{aligned}
|\widehat{\mathcal{C}}_1^c \cap T_{\pi(1)}| &= |T_{\pi(1)}| - |T_{\pi(1)} \cap \widehat{\mathcal{C}}_1| \\
&\leq |\widehat{\mathcal{C}}_1| - |T_{\pi(1)} \cap \widehat{\mathcal{C}}_1| \\
&= |\widehat{\mathcal{C}}_1 \cap T_{\pi(1)}^c| \\
&\leq \left|\left(\cup_{i\in[k]}T_i\right)^c\right|,
\end{aligned}$$

where the last inequality is because $T_{\pi(1)}$ is the only set in $\{T_i\}_{i\in[k]}$ that intersects $\widehat{\mathcal{C}}_1$ by the definitions. By (64), we get

$$|\widehat{\mathcal{C}}_i^c \cap T_{\pi(i)}| \leq \frac{4C^2na}{\mu^2\lambda_k^2}, \tag{71}$$

for $i = 1$.

Now suppose (66) and (71) are true for $i = 1, ..., l-1$. Because of the sizes of $\{\widehat{\mathcal{C}}_i\}_{i\in[l-1]}$ and the fact that $\{T_i\}_{i\in[k]}$ are mutually exclusive, we have

$$\left(\cup_{i=1}^{l-1}\widehat{\mathcal{C}}_i\right) \cap \left(\cup_{i\in[k]\setminus\cup_{i=1}^{l-1}\{\pi(i)\}}T_i\right) = \varnothing.$$

Therefore, for the set S in the current step, $\cup_{i\in[k]\setminus\cup_{i=1}^{l-1}\{\pi(i)\}}T_i \subset S$. By the definition of $\widehat{\mathcal{C}}_l$, we have $|\widehat{\mathcal{C}}_l| \geq \max_{i\in[k]\setminus\cup_{i=1}^{l-1}\{\pi(i)\}}|T_i| \geq \frac{n}{2\beta k}$. Suppose $\widehat{\mathcal{C}}_l \cap T_{\pi(i)} \neq \varnothing$ for some $i = 1, ..., l-1$. Then, this $T_{\pi(i)}$ is the only set in $\{T_i\}_{i\in[k]}$ that intersects $\widehat{\mathcal{C}}_l$ by their definitions. This implies that

$$|\widehat{\mathcal{C}}_l| \leq |\widehat{\mathcal{C}}_l \cap T_{\pi(i)}| + \left|\left(\cup_{i\in[k]}T_i\right)^c\right|.$$

Since $\widehat{\mathcal{C}}_l \cap \widehat{\mathcal{C}}_{\pi(i)} = \varnothing$, $|\widehat{\mathcal{C}}_l \cap T_{\pi(i)}| \leq |\widehat{\mathcal{C}}_i^c \cap T_{\pi(i)}|$ is bounded by (71). Together with (64), we have

$$|\widehat{\mathcal{C}}_l| \leq \frac{8C^2na}{\mu^2\lambda_k^2},$$

which contradicts $|\widehat{\mathcal{C}}_l| \geq \frac{n}{2\beta k}$ under the assumption (20). Therefore, we must have $\widehat{\mathcal{C}}_l \cap T_{\pi(i)} = \varnothing$ for all $i = 1, ..., l-1$. Now suppose $\widehat{\mathcal{C}}_l \cap T_{\pi(i)} = \varnothing$ for all $i \in [k]$, we must have

$$\left|\left(\cup_{i\in[k]}T_i\right)^c\right| \geq |\widehat{\mathcal{C}}_l| \geq \frac{n}{2\beta k},$$

which contradicts (64). Hence, $\widehat{C}_l \cap T_{\pi(l)} \neq \varnothing$ for some $\pi(l) \in [k]\setminus \cup_{i=1}^{l-1}\{\pi(i)\}$, and (66) is established for $i = l$. Moreover, (71) can also be established for $i = l$ by the same argument that is used to prove (71) for $i = 1$. The proof is complete. ■

6.3 Proof of Theorem 10

Define $P_\tau = P + \frac{\tau}{n}\mathbf{1}\mathbf{1}^T$. The proof of the following lemma is given in the appendix.

Lemma 22 *Consider a symmetric adjacency matrix $A \in \{0,1\}^{n\times n}$ and a symmetric matrix $P \in [0,1]^{n\times n}$ satisfying $A_{uu} = 0$ for all $u \in [n]$ and $A_{uv} \sim$ Bernoulli(P_{uv}) independently for all $u > v$. For any $C' > 0$, there exists some $C > 0$ such that*

$$\|L(A_\tau) - L(P_\tau)\|_{\text{op}} \leq C\sqrt{\frac{\log(e(np_{\max}+1))}{np_{\max}+1}},$$

with probability at least $1 - n^{-C'}$ uniformly over $\tau \in [C_1(np_{\max}+1), C_2(np_{\max}+1)]$ for some sufficiently large constants C_1, C_2, where $p_{\max} = \max_{u\geq v}P_{uv}$.

Lemma 23 *Consider $P = (P_{uv}) = (B_{\sigma(u)\sigma(v)})$. Let the SVD of the matrix $L(P_\tau)$ be $L(P_\tau) = U\Sigma U^T$, with $U \in O(n,k)$ and $\Sigma = diag(\sigma_1, ..., \sigma_k)$. For $V = UW$ with any $W \in O(r,r)$, we have $\|V_{u*} - V_{v*}\| = \sqrt{\frac{1}{n_u} + \frac{1}{n_v}}$ when $\sigma(u) \neq \sigma(v)$ and $V_{u*} = V_{v*}$ when $\sigma(u) = \sigma(v)$. Moreover, $\sigma_k \geq \frac{\lambda_k}{2\tau}$ as long as $\tau \geq np_{\max}$.*

Proof The first part is Lemma 1 in Joseph and Yu (2013). Define $\bar{d}_v = \sum_{u\in[n]}P_{uv}$ and $\bar{D}_\tau = \text{diag}(\bar{d}_1 + \tau, ..., \bar{d}_n + \tau)$. Then, we have $L(P_\tau) = \bar{D}_\tau^{-1/2}P_\tau\bar{D}_\tau^{-1/2}$. Note that P_τ has an SBM structure so that it has rank at most k, and the k^{th} eigenvalue of P_τ is lower bounded by λ_k. Thus, we have

$$\sigma_k \geq \frac{\lambda_k}{\max_{u\in[n]}\bar{d}_u + \tau}.$$

Observe that $\max_{u\in[n]} \bar{d}_u \le np_{\max} \le \tau$, and the proof is complete. ■

Proof [Proof of Theorem 10] As is shown in the proof of Theorem 6, $\tau \in [C_1 a, C_2 a]$ for some large C_1, C_2 with probability at least $1-e^{-C'n}$. By Davis–Kahan's sin-theta theorem (Davis and Kahan, 1970), we have $\|\widehat{U} - UW\|_{\mathrm{F}} \le C_1 \frac{\sqrt{k}}{\sigma_k} \|L(A_\tau) - L(P_\tau)\|_{\mathrm{op}}$ for some $W \in O(r,r)$ and some constant $C_1 > 0$. Let $V = UW$ and apply Lemma 22 and Lemma 23, we have

$$\|\widehat{U} - V\|_{\mathrm{F}} \le \frac{C\sqrt{k}\sqrt{a\log a}}{\lambda_k}, \tag{72}$$

with probability at least $1 - n^{-C'}$. Note that by Lemma 23, V satisfies (62). Replace (61) by (72), and follow the remaining proof of Theorem 6, the proof is complete. ■

Appendix A. A simplified version of Algorithm 1

We give in Algorithm 3 the precise description of the simplified version of Algorithm 1 that we have used in Section 4.

A.1 Additional simulation results

We report some additional simulation results comparing the performances of Algorithm 1 and Algorithm 3. In particular, we consider networks with 400 nodes and 2 communities with four different sets of community sizes, within and between community connection probabilities and initialization methods, and the simulation results are reported in Figure 5 on page 34. From Figure 5 on page 34 we can see the performances of both algorithms remain similar across all these different settings.

Appendix B. Proofs of Theorem 12

Proof [Proof of Theorem 12] Let us consider $\Theta = \Theta_0(n,k,a,b,\beta)$ and the case of $\Theta(n,k,a,b,\lambda,\beta;\alpha)$ is similar except that the condition (18) is needed to establish the counterpart of Lemma 18. The proof essentially follows the same steps as those in the proof of Theorem 4. First, we note that Lemma 16 continues to hold since it does not need the assumption of a/b being bounded. Thus, the first job is to establish the counterpart of Lemma 17 with η' replaced with $C_\epsilon \frac{2\epsilon_0}{3}$. As before, let $p = a/n$ and $q = b/n$.

To this end, we first proceed in the same way to obtain (42) – (52). Without loss of generality, let us consider the case where $t^* > \log\frac{2}{\epsilon_0}$ and $t_u = \log\frac{2}{\epsilon_0}$ since otherwise we can essentially repeat the proof of Theorem 4. Note that this implies $\frac{a}{b} > (\frac{2}{\epsilon_0})^2$. In this case, with the new t_u in (22), we have on the event E_u,

$$(qe^{t_u} + 1 - q)(pe^{-t_u} + 1 - p) = e^{-I'}$$

Algorithm 3: A simplified refinement scheme for community detection

Input: Adjacency matrix $A \in \{0,1\}^{n\times n}$,
number of communities k,
initial community detection method σ^0.

Output: Community assignment $\widehat{\sigma}$.

Initialization:

Apply σ^0 on A to obtain $\sigma^0(u)$ for all $u \in [n]$;

Define $\widetilde{\mathcal{C}}_i = \{v : \sigma^0(v) = i\}$ for all $i \in [k]$; let $\widetilde{\mathcal{E}}_i$ be the set of edges within $\widetilde{\mathcal{C}}_i$, and $\widetilde{\mathcal{E}}_{ij}$ the set of edges between $\widetilde{\mathcal{C}}_i$ and $\widetilde{\mathcal{C}}_j$ when $i \neq j$;

Define

$$\widehat{B}_{ii} = \frac{|\widetilde{\mathcal{E}}_i|}{\frac{1}{2}|\widetilde{\mathcal{C}}_i|(|\widetilde{\mathcal{C}}_i|-1)}, \quad \widehat{B}_{ij} = \frac{|\widetilde{\mathcal{E}}_{ij}|}{|\widetilde{\mathcal{C}}_i||\widetilde{\mathcal{C}}_j|}, \quad \forall i \neq j \in [k],$$

and let

$$\widehat{a} = n \min_{i\in[k]} \widehat{B}_{ii} \quad \text{and} \quad \widehat{b} = n \max_{i\neq j\in[k]} \widehat{B}_{ij}.$$

Penalized neighbor voting:

For

$$t = \frac{1}{2}\log\frac{\widehat{a}(1-\widehat{b}/n)}{\widehat{b}(1-\widehat{a}/n)},$$

define

$$\rho = -\frac{1}{2t}\log\left(\frac{\frac{\widehat{a}}{n}e^{-t} + 1 - \frac{\widehat{a}}{n}}{\frac{\widehat{b}}{n}e^{t} + 1 - \frac{\widehat{b}}{n}}\right),$$

For each $u \in [n]$, set

$$\widehat{\sigma}(u) = \operatorname*{argmax}_{l\in[k]} \sum_{\sigma^0(v)=l} A_{uv} - \rho \sum_{v\in[n]} \mathbf{1}_{\{\sigma^0(v)=l\}}.$$

where

$$\begin{aligned} I' &= -\log\left((1-p)(1-q) + pq + \left(e^{t_u - t^*} + e^{t^* - t_u}\right)\sqrt{(1-p)(1-q)pq}\right) \\ &\geq \left(1 - C_\epsilon \frac{3\epsilon_0}{5}\right) I^*. \end{aligned} \tag{73}$$

To see this, we first note that for any $x, y \in (0,1)$ and sufficient small constant $c_0 > 0$, if $y \geq x \geq (1-c_0)y$ and $\frac{y-x}{1-y} \leq 1$, then

$$-\log(1-x) = -\log(1-y) - \log\left(1 + \frac{y-x}{1-y}\right) \geq -\log(1-y) - 2\frac{y-x}{1-y} \geq -\left(1 - C'_y c_0\right)\log(1-y),$$

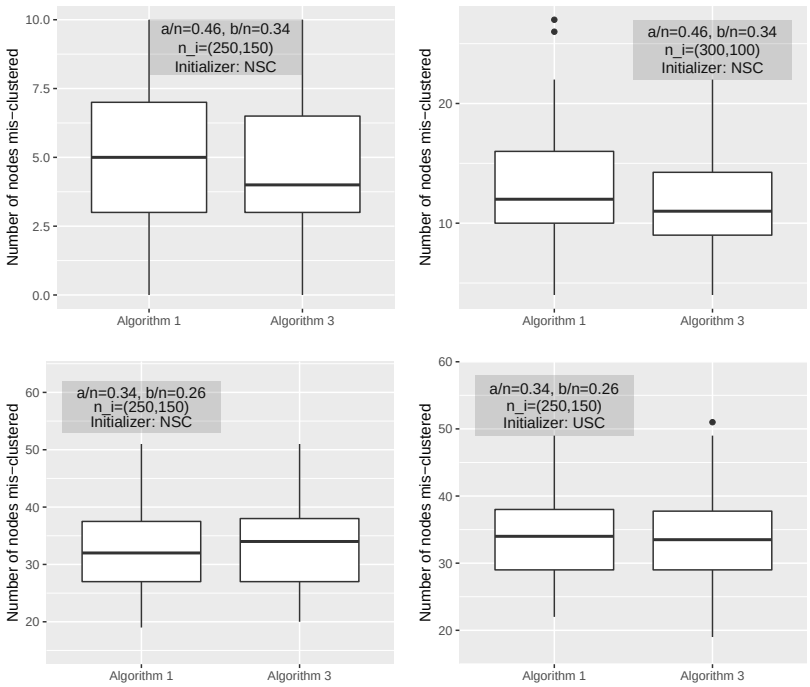

Figure 5: Comparison of Algorithm 1 and its simplified version (Algorithm 3) under various parameters for two-community networks. The parameters are displayed inside each boxplot. Here n_i gives the community sizes and B matrix satisfies $B_{ii} = a/n, \forall i = 1, 2$ and $B_{ij} = b/n, \forall i \neq j$. Each pair of boxplots are based on 100 runs.

where $C'_y = \frac{2y}{-(1-y)\log(1-y)}$. When $\frac{a}{b} > (\frac{2}{\epsilon_0})^2$ and $t_u = \log\frac{2}{\epsilon_0}$, we have $I' = -\log(1-x)$ for

$$x = p + q - 2pq - (e^{t_u - t^*} + e^{t^* - t_u})\sqrt{(1-p)(1-q)pq} \geq p - 2pq - qe^{t_u} - pe^{-t_u} \geq p(1 - \epsilon_0 - \frac{\epsilon_0^2}{2}),$$

while $I^* = -\log(1-y)$ for

$$y = p + q - 2pq - 2\sqrt{(1-p)(1-q)pq} \leq p + q \leq p(1 + (\frac{\epsilon_0}{2})^2).$$

Thus, for any $\epsilon_0 \in (0, c_\epsilon)$, $1 - \frac{\epsilon}{2} \geq y \geq x \geq (1 - 2\epsilon_0)y$ and $\frac{y-x}{1-y} \leq 1$, and we apply the inequality in the third last display to obtain (73).

Thus, the term in (49) is upper bounded by

$$\exp\left(-\big(1 - C_\epsilon \frac{3\epsilon_0}{5}\big)\frac{n_1 + n_l}{2} I^*\right).$$

On the other hand, since $|e^{-t_u} - 1| \leq 1$, $|e^{t_u} - 1|$ is bounded and $\frac{p-q}{p} \asymp 1$, the term in (48) continues to be bounded by

$$\exp\left(-o(1)\frac{n_1 + n_l}{2} I^*\right).$$

Moreover, by the same argument as in Lemma 17, (51) continues to hold. Thus, we can replace (52) as

$$p_l \le \exp\left(-\big(1 - C_\epsilon \frac{2\epsilon_0}{3}\big)\frac{n_1 + n_l}{2} I^*\right),$$

and so when $k \ge 3$,

$$\mathbb{P}\left\{\widehat{\sigma}_u(u) \ne \pi_u(\sigma(u))\right\} \le (k-1)\exp\left\{-\left(1 - C_\epsilon \frac{2\epsilon_0}{3}\right)\frac{nI^*}{\beta k}\right\} + Cn^{-(1+\delta)} \tag{74}$$

and when $k = 2$, we can replace β by 1 in the last display.

When $k \ge 3$, given the last display and (58), we have

$$\begin{aligned}
\mathbb{P}\left\{\widehat{\sigma}(u) \ne \sigma(u)\right\} &= \mathbb{P}\left\{\xi_u(\widehat{\sigma}_u(u)) \ne \sigma(u)\right\} \\
&\le \mathbb{P}\left\{\xi_u(\widehat{\sigma}_u(u)) \ne \sigma(u),\, \xi_u = \pi_u^{-1}\right\} + \mathbb{P}\left\{\xi_u \ne \pi_u^{-1}\right\} \\
&\le \mathbb{P}\left\{\widehat{\sigma}_u(u) \ne \pi_u(\sigma(u))\right\} + \mathbb{P}\left\{\xi_u \ne \pi_u^{-1}\right\} \\
&\le Cn^{-(1+\delta)} + (k-1)\exp\left\{-\left(1 - C_\epsilon \frac{2\epsilon_0}{3}\right)\frac{nI^*}{\beta k}\right\}.
\end{aligned} \tag{75}$$

Thus, the assumption that $\frac{(a-b)^2}{ak\log k} \to \infty$ and Markov's inequality leads to

$$\begin{aligned}
&\mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > \exp\left\{-(1 - C_\epsilon\epsilon_0)\frac{nI^*}{\beta k}\right\}\right\} \\
&\quad\le \mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > (k-1)\exp\left\{-(1 - C_\epsilon\frac{5\epsilon_0}{6})\frac{nI^*}{\beta k}\right\}\right\} \\
&\quad\le \frac{1}{(k-1)\exp\left\{-(1 - C_\epsilon\frac{5\epsilon_0}{6})\frac{nI^*}{\beta k}\right\}} \frac{1}{n}\sum_{u=1}^{n} \mathbb{P}\left\{\widehat{\sigma}(u) \ne \sigma(u)\right\} \\
&\quad\le \exp\left\{-\frac{C_\epsilon\epsilon_0}{6}\frac{nI^*}{\beta k}\right\} + \frac{Cn^{-(1+\delta)}}{(k-1)\exp\left\{-(1 - C_\epsilon\frac{5\epsilon_0}{6})\frac{nI^*}{\beta k}\right\}}.
\end{aligned} \tag{76}$$

If $(k-1)\exp\left\{-(1 - C_\epsilon\frac{5\epsilon_0}{6})\frac{nI^*}{\beta k}\right\} \ge n^{-(1+\delta/2)}$, then

$$\mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > \exp\left\{-(1 - C_\epsilon\epsilon_0)\frac{nI^*}{\beta k}\right\}\right\} \le \exp\left\{-\frac{C_\epsilon\epsilon_0}{6}\frac{nI^*}{\beta k}\right\} + Cn^{-\delta/2} = o(1). \tag{77}$$

If $(k-1)\exp\left\{-(1 - C_\epsilon\frac{5\epsilon_0}{6})\frac{nI^*}{\beta k}\right\} < n^{-(1+\delta/2)}$, then

$$\begin{aligned}
\mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > \exp\left\{-(1 - C_\epsilon\epsilon_0)\frac{nI^*}{\beta k}\right\}\right\} &\le \mathbb{P}\left\{\ell_0(\sigma, \widehat{\sigma}) > 0\right\} \le \sum_{u=1}^{n} \mathbb{P}\left\{\widehat{\sigma}(u) \ne \sigma(u)\right\} \\
&\le n(k-1)\exp\left\{-(1 - C_\epsilon\frac{2\epsilon_0}{3})\frac{nI^*}{\beta k}\right\} + Cn^{-\delta} \le Cn^{-\delta/2} = o(1).
\end{aligned} \tag{78}$$

Here, the second last inequality holds since $(k-1)\exp\left\{-(1-C_\epsilon\frac{2\epsilon_0}{3})\frac{nI^*}{\beta k}\right\} < \exp\left\{-(1-C_\epsilon\frac{5\epsilon_0}{6})\frac{nI^*}{\beta k}\right\} < n^{-(1+\delta/2)}$. We complete the proof for the case of $k \geq 3$ by noting that no constant or sequence in the foregoing arguments involves B, σ or u. When $k = 2$, we run the foregoing arguments with β replaced by 1 to obtain the desired claim. ■

Appendix C. Proofs of Theorems 13, 14 and 15

Proposition 24 *For SBM in the space $\Theta_0(n, k, a, b, \beta)$ satisfying $n \geq 2\beta k$, we have $\lambda_k \geq \frac{a-b}{\beta k}$.*

Proof Since the eigenvalues of P are invariant with respect to permutation of the community labels, we consider the case where $\sigma(u) = i$ for $u \in \left\{\sum_{j=1}^{i-1} n_j - 1, \sum_{j=1}^{i} n_j\right\}$ without loss of generality, where $\sum_{j=1}^{0} n_j = 0$. Let us use the notation $\mathbf{1}_d \in \mathbb{R}^d$ and $\mathbf{0}_d \in \mathbb{R}^d$ to denote the vectors with all entries being 1 and 0 respectively. Then, it is easy to check that

$$P - \frac{b}{n}\mathbf{1}_n\mathbf{1}_n^T = \frac{a-b}{n}\sum_{i=1}^{k} v_i v_i^T,$$

where $v_1 = (\mathbf{1}_{n_1}^T, \mathbf{0}_{n_2}^T, ..., \mathbf{0}_{n_k}^T)^T$, $v_1 = (\mathbf{0}_{n_1}^T, \mathbf{1}_{n_2}^T, \mathbf{0}_{n_3}^T, ..., \mathbf{0}_{n_k}^T)^T$,..., $v_k = (\mathbf{0}_{n_1}^T, ..., \mathbf{0}_{n_{k-1}}^T, \mathbf{1}_{n_k}^T)^T$. Note that $\{v_i\}_{i=1}^k$ are orthogonal to each other, and therefore

$$\lambda_k\left(\sum_{i=1}^{k} v_i v_i^T\right) \geq \min_{i\in[k]} n_i \geq \frac{n}{\beta k} - 1 \geq \frac{n}{2\beta k}.$$

By Weyl's inequality (Theorem 4.3.1 of Horn and Johnson (2012)),

$$\lambda_k(P) \geq \frac{a-b}{n}\lambda_k\left(\sum_{i=1}^{k} v_i v_i^T\right) + \lambda_n\left(\frac{b}{n}\mathbf{1}_n\mathbf{1}_n^T\right) \geq \frac{a-b}{2\beta k}.$$

This completes the proof. ■

Proof [Proof of Theorem 13] Let us first consider $\Theta_0(n, k, a, b, \beta)$. By Theorem 6 and Proposition 24, the misclassification proportion is bounded by $C\frac{k^2 a}{(a-b)^2}$ under the condition $\frac{k^3 a}{(a-b)^2} \leq c$ for some small c. Thus, Condition 1 holds when $\frac{k^3 a}{(a-b)^2} = o(1)$, which leads to the desired conclusion in view of Theorem 4 and Theorem 12. The proof of the space $\Theta(n, k, a, b, \lambda, \beta; \alpha)$ follows the same argument. ■

Proof [Proof of Theorem 14] The proof is the same as that of Theorem 13. ■

Proof [Proof of Theorem 15] When the parameters a and b are known, we can use $\tau = Ca$ for some sufficiently large $C > 0$ for both USC(τ) and NSC(τ). Then, the results of Theorem 6 and Theorem 10 hold without assuming $a \leq C_1 b$ or fixed k. Moreover, $\widehat{a}_u$ and $\widehat{b}_u$ in

(11) and (22) can be replaced by a and b. Then, the conditions (16) and (18) in Theorem 4 and Theorem 12 can be weakened as $\gamma = o(k^{-1})$ because the we do not need to establish Lemma 16 anymore. Combining Theorem 4, Theorem 6, Theorem 10 and Theorem 12, we obtain the desired results. ■

Appendix D. Proofs of Lemma 20 and Lemma 22

The following lemma is Corollary A.1.10 in Alon and Spencer (2004).

Lemma 25 *For independent Bernoulli random variables $X_u \sim Bern(p_u)$ and $p = \frac{1}{n}\sum_{u\in[n]} p_u$, we have*

$$\mathbb{P}\left(\sum_{u\in[n]}(X_u - p_u) \geq t\right) \leq \exp\left(t - (pn+t)\log\left(1+\frac{t}{pn}\right)\right),$$

for any $t \geq 0$.

The following result is Lemma 3.5 in Chin et al. (2015).

Lemma 26 *Consider any adjacency matrix $A \in \{0,1\}^{n\times n}$ for an undirected graph. Suppose $\max_{u\in[n]}\sum_{v\in[n]} A_{uv} \leq \gamma$ and for any $S, T \subset [n]$, one of the following statements holds with some constant $C > 0$:*

1. $\frac{e(S,T)}{|S||T|\frac{\gamma}{n}} \leq C$,
2. $e(S,T)\log\left(\frac{e(S,T)}{|S||T|\frac{\gamma}{n}}\right) \leq C|T|\log\frac{n}{|T|}$,

where $e(S,T)$ is the number of edges connecting S and T. Then, $\sum_{(u,v)\in H} x_u A_{uv} y_v \leq C'\sqrt{\gamma}$ uniformly over all unit vectors x, y, where $H = \{(u,v) : |x_u y_v| \geq \sqrt{\gamma}/n\}$ and $C' > 0$ is some constant.

The following lemma is critical for proving both theorems.

Lemma 27 *For any $\tau > C(1 + np_{\max})$ with some sufficiently large $C > 0$, we have*

$$|\{u \in [n] : d_u \geq \tau\}| \leq \frac{n}{\tau}$$

with probability at least $1 - e^{-C'n}$ for some constant $C' > 0$.

Proof Let us consider any fixed subset of nodes $S \subset [n]$ such that it has degree at least τ and $|S| = l$ for some $l \in [n]$. Let $e(S)$ be the number of edges in the subgraph S and $e(S, S^c)$ be the number of edges connecting S and S^c. By the requirement on S, either $e(S) \geq C_1 l\tau$ or $e(S, S^c) \geq C_1 l\tau$ for some universal constant $C_1 > 0$. We are going to show that both $\mathbb{P}(e(S) \geq C_1 l\tau)$ and $\mathbb{P}(e(S, S^c) \geq C_1 l\tau)$ are small. Note that $\mathbb{E}e(S) \leq C_2 l^2 p_{\max}$

and $\mathbb{E}e(S, S^c) \le C_2 l n p_{\max}$ for some universal $C_2 > 0$. Then, when $\tau > C(np_{\max} + 1)$ for some sufficiently large $C > 0$, Lemma 25 implies

$$\mathbb{P}\left(e(S) \ge C_1 l\tau\right) \le \exp\left(-\frac{1}{4}C_1 l\tau \log\left(1 + \frac{C_1\tau}{2C_2 l p_{\max}}\right)\right),$$

and

$$\mathbb{P}\left(e(S, S^c) \ge C_1 l\tau\right) \le \exp\left(-\frac{1}{4}C_1 l\tau \log\left(1 + \frac{C_1\tau}{2C_2 n p_{\max}}\right)\right).$$

Applying union bound, the probability that the number of nodes with degree at least τ is greater than ξn is

$$\begin{aligned}
& \mathbb{P}\Big(|\{u \in [n] : d_u \ge \tau\}| > \xi n\Big) \\
\le & \sum_{l > \xi n} \mathbb{P}\Big(|\{u \in [n] : d_u \ge \tau\}| = l\Big) \\
\le & \sum_{l > \xi n} \sum_{|S| = l} \left(\mathbb{P}\left(e(S) \ge C_1 l\tau\right) + \mathbb{P}\left(e(S, S^c) \ge C_1 l\tau\right)\right) \\
\le & \sum_{l > \xi n} \exp\left(l \log \frac{en}{l}\right) \left(\exp\left(-\frac{1}{4}C_1 l\tau \log\left(1 + \frac{C_1\tau}{2C_2 l p_{\max}}\right)\right)\right. \\
& \left. + \exp\left(-\frac{1}{4}C_1 l\tau \log\left(1 + \frac{C_1\tau}{2C_2 n p_{\max}}\right)\right)\right) \\
\le & \sum_{l > \xi n} 2\exp\left(l \log \frac{en}{l} - \frac{1}{4}C_1 l\tau \log\left(1 + \frac{C_1\tau}{2C_2 n p_{\max}}\right)\right) \\
\le & \exp(-C'n),
\end{aligned}$$

where the last inequality is by choosing $\xi = \tau^{-1}$. Therefore, with probability at least $1 - e^{-C'n}$, the number of nodes with degree at least τ is bounded by $\tau^{-1}n$. ■

Lemma 28 *Given $\tau > 0$, define the subset $J = \{u \in [n] : d_u \le \tau\}$. Then for any $C' > 0$, there is some $C > 0$ such that*

$$\|A_{JJ} - P_{JJ}\|_{\text{op}} \le C\left(\sqrt{np_{\max}} + \sqrt{\tau} + \frac{np_{\max}}{\sqrt{\tau} + \sqrt{np_{\max}}}\right),$$

with probability at least $1 - n^{-C'}$.

Proof The idea of the proof follows the argument in Friedman et al. (1989); Feige and Ofek (2005). By definition,

$$\|A_{JJ} - P_{JJ}\|_{\text{op}} = \sup_{x,y \in S^{n-1}} \sum_{(u,v) \in J \times J} x_u (A_{uv} - P_{uv}) y_v.$$

Define $L = \{(u,v) : |x_u y_v| \le (\sqrt{\tau} + \sqrt{p_{\max} n})/n\}$ and $H = \{(u,v) : |x_u y_v| \ge (\sqrt{\tau} + \sqrt{p_{\max} n})/n\}$, then we have

$$\|A_{JJ} - P_{JJ}\|_{\mathrm{op}} \le \sup_{x,y \in S^{n-1}} \sum_{(u,v) \in L \cap J \times J} x_u (A_{uv} - P_{uv}) y_v + \sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u (A_{uv} - P_{uv}) y_v.$$

A discretization argument in Chin et al. (2015) implies that

$$\begin{aligned} \sup_{x,y \in S^{n-1}} \sum_{(u,v) \in L \cap J \times J} x_u (A_{uv} - P_{uv}) y_v \lesssim & \max_{x,y \in \mathcal{N}} \max_{S \subset [n]} \sum_{(u,v) \in L \cap S \times S} x_u (A_{uv} - \mathbb{E} A_{uv}) y_v \\ & + \max_{x,y \in \mathcal{N}} \max_{S \subset [n]} \sum_{(u,v) \in L \cap S \times S} x_u (\mathbb{E} A_{uv} - P_{uv}) y_v, \end{aligned}$$

where $\mathcal{N} \subset S^{n-1}$ and $|\mathcal{N}| \le 5^n$. Then, Bernstein's inequality and union bound imply that $\max_{x,y \in \mathcal{N}} \max_{S \subset [n]} \sum_{(u,v) \in L \cap S \times S} x_u (A_{uv} - \mathbb{E} A_{uv}) y_v \le C(\sqrt{\tau} + \sqrt{n p_{\max}})$ with probability at least $1 - e^{-C' n}$. We also have $\max_{x,y \in \mathcal{N}} \max_{S \subset [n]} \sum_{(u,v) \in L \cap S \times S} x_u (\mathbb{E} A_{uv} - P_{uv}) y_v \le \|\mathbb{E} A - P\|_{\mathrm{op}} \le 1$. This completes the first part.

To bound the second part $\sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u (A_{uv} - P_{uv}) y_v$, we are going to bound $\sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u A_{uv} y_v$ and $\sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u P_{uv} y_v$ separately. By the definition of H,

$$\sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u P_{uv} y_v = \sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} \frac{x_u^2 y_v^2}{|x_u y_v|} P_{uv} \le \frac{n p_{\max}}{\sqrt{\tau} + \sqrt{p_{\max} n}}.$$

To bound $\sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u A_{uv} y_v$, it is sufficient to check the conditions of Lemma 26 for the graph A_{JJ}. By definition, its degree is bounded by τ. Following the argument of Lei and Rinaldo (2014), the two conditions of Lemma 26 hold with $\gamma = \tau + n p_{\max}$ with probability at least $1 - n^{-C'}$. Thus, $\sup_{x,y \in S^{n-1}} \sum_{(u,v) \in H \cap J \times J} x_u A_{uv} y_v \le C(\sqrt{\tau} + \sqrt{n p_{\max}})$ with probability at least $1 - n^{-C'}$. Hence, the proof is complete. ■

Proof [Proof of Lemma 20] By triangle inequality,

$$\|T_\tau(A) - P\|_{\mathrm{op}} \le \|T_\tau(A) - T_\tau(P)\|_{\mathrm{op}} + \|T_\tau(P) - P\|_{\mathrm{op}},$$

where $T_\tau(P)$ is the matrix obtained by zeroing out the u^{th} row and column of P with $d_u \ge \tau$. Let $J = \{u \in [n] : d_u \le \tau\}$, and then $\|T_\tau(A) - T_\tau(P)\|_{\mathrm{op}} = \|A_{JJ} - P_{JJ}\|_{\mathrm{op}}$, whose bound has been established in Lemma 28. By Lemma 27, $|J^c| \le n/\tau$ with high probability. This implies $\|T_\tau(P) - P\|_{\mathrm{op}} \le \|T_\tau(P) - P\|_{\mathrm{F}} \le \sqrt{2n|J^c| p_{\max}^2} \le \frac{\sqrt{2} n p_{\max}}{\sqrt{\tau}}$. Taking $\tau \in [C_1(1 + n p_{\max}), C_2(1 + n p_{\max})]$, the proof is complete. ■

Now let us prove Lemma 22. The following lemma, which controls the degree, is Lemma 7.1 in Le et al. (2015).

Lemma 29 *For any $C' > 0$, there exists some $C > 0$ such that with probability at least $1 - n^{-C'}$, there exists a subset $J \subset [n]$ satisfying $n - |J| \le \frac{n}{2e(np_{\max}+1)}$ and*

$$|d_v - \mathbb{E}d_v| \le C\sqrt{(np_{\max}+1)\log(e(np_{\max}+1))}, \quad \text{for all } v \in J,$$

where $d_v = \sum_{u\in[n]} A_{uv}$.

Using this lemma, together with Lemma 27 and Lemma 28, we are able to prove the following result, which improves the bound in Theorem 7.2 of Le et al. (2015).

Lemma 30 *For any $C' > 0$, there exists some $C > 0$ such that with probability at least $1 - n^{-C'}$, there exists a subset $J \subset [n]$ satisfying $n - |J| \le n/d$ and*

$$\|(L(A_\tau) - L(P_\tau))_{J\times J}\|_{\text{op}} \le C\left(\frac{\sqrt{d\log d}(d+\tau)}{\tau^2} + \frac{\sqrt{d}}{\tau}\right),$$

where $d = e(np_{\max}+1)$.

Proof Let us use the notation $d_v = \sum_{u\in[n]} A_{uv}$ in the proof. Define the set $J_1 = \{v \in [n] : d_v \le C_1 d\}$ for some sufficiently large constant $C_1 > 0$. Using Lemma 27 and Lemma 28, with probability at least $1 - n^{-C'}$, we have

$$n - |J_1| \le \frac{n}{2d}, \tag{79}$$

and

$$\|(A-P)_{J_1J_1}\|_{\text{op}} \le C\sqrt{d}. \tag{80}$$

Let J_2 be the subset in Lemma 29, and then with probability at least $1 - n^{-C'}$, J_2 satisfies

$$n - |J_2| \le \frac{n}{2d}, \tag{81}$$

and

$$|d_v - \mathbb{E}d_v| \le C\sqrt{d\log d}, \quad \text{for all } v \in J_2. \tag{82}$$

Define $J = J_1 \cap J_2$. By (79) and (81), we have

$$n - |J| = |(J_1 \cap J_2)^c| \le |J_1^c| + |J_2^c| = n - |J_1| + n - |J_2| \le \frac{n}{d}, \tag{83}$$

and

$$\|(A-P)_{JJ}\|_{\text{op}} \le \|(A-P)_{J_1J_1}\|_{\text{op}} \le C\sqrt{d}. \tag{84}$$

Moreover, (82) implies

$$\max_{v\in J} |d_v - \mathbb{E}d_v| \le C\sqrt{d\log d}.$$

Define $\bar{d}_v = \sum_{u\in[n]} P_{uv}$. Then,

$$\max_{v\in J} |d_v - \bar{d}_v| \le \max_{v\in J} |d_v - \mathbb{E}d_v| + 1 \le C\sqrt{d\log d}. \tag{85}$$

Define $D_\tau = \mathrm{diag}(d_1+\tau, ..., d_n+\tau)$ and $\bar{D}_\tau = \mathrm{diag}(\bar{d}_1+\tau, ..., \bar{d}_n+\tau)$. We introduce the notation

$$R = (A_\tau)_{JJ}, \quad B = (D_\tau)^{-1/2}_{JJ}, \quad \bar{R} = (P_\tau)_{JJ}, \quad \bar{B} = (\bar{D}_\tau)^{-1/2}_{JJ}.$$

Using (85), we have

$$\|B - \bar{B}\|_{\mathrm{op}} \le \max_{v\in[n]} \left| \frac{1}{\sqrt{d_v+\tau}} - \frac{1}{\sqrt{\bar{d}_v+\tau}} \right| \le C \frac{\sqrt{d\log d}}{\tau^{3/2}},$$

for some constant $C > 0$. The definitions of B and $\bar{B}$ implies $\|B\|_{\mathrm{op}} \vee \|\bar{B}\|_{\mathrm{op}} \le \frac{1}{\sqrt{\tau}}$. We rewrite the bound (84) as $\|R-\bar{R}\|_{\mathrm{op}} \le C\sqrt{d}$. Since all entries of $\mathbb{E}A_\tau$ is bounded by $(\tau+d)/n$, we have $\|\bar{R}\|_{\mathrm{op}} \le \|\mathbb{E}A_\tau\|_{\mathrm{op}} \le d+\tau$. Therefore, $\|R\|_{\mathrm{op}} \le \|\bar{R}\|_{\mathrm{op}} + \|R-\bar{R}\|_{\mathrm{op}} \le C(d+\tau)$. Finally,

$$\begin{aligned} & \|(L(A_\tau) - L(P_\tau))_{J\times J}\|_{\mathrm{op}} \\ \le\ & \|B\|_{\mathrm{op}}\|R\|_{\mathrm{op}}\|B-\bar{B}\|_{\mathrm{op}} + \|B\|_{\mathrm{op}}\|R-\bar{R}\|_{\mathrm{op}}\|\bar{B}\|_{\mathrm{op}} + \|B-\bar{B}\|_{\mathrm{op}}\|\bar{R}\|_{\mathrm{op}}\|\bar{B}\|_{\mathrm{op}} \\ \le\ & C\left(\frac{\sqrt{d\log d}(d+\tau)}{\tau^2} + \frac{\sqrt{d}}{\tau}\right). \end{aligned}$$

The proof is complete. ■

Proof [Proof of Lemma 22] Recall that $d = np_{\max}+1$. Following the proof of Theorem 8.4 in Le et al. (2015), it can be shown that with probability at least $1-n^{-C'}$, for any $J \subset [n]$ such that $n - |J| \le n/d$,

$$\|L(A_\tau) - L(P_\tau)\|_{\mathrm{op}} \le \|(L(A_\tau) - L(P_\tau))_{JJ}\|_{\mathrm{op}} + C\left(\frac{1}{\sqrt{d}} + \sqrt{\frac{\log d}{\tau}}\right),$$

where the first term on the right side of the inequality above is bounded in Lemma 30 by choosing an appropriate J. Hence, with probability at least $1-2n^{-C'}$,

$$\|L(A_\tau) - L(P_\tau)\|_{\mathrm{op}} \le C\left(\frac{\sqrt{d\log d}(d+\tau)}{\tau^2} + \frac{\sqrt{d}}{\tau}\right) + C\left(\frac{1}{\sqrt{d}} + \sqrt{\frac{\log d}{\tau}}\right).$$

Choosing $\tau \in [C_1(1+np_{\max}), C_2(1+np_{\max})]$, the proof is complete. ■

Acknowledgments

The research of AY. Z. and HH. Z. was supported in part by the NSF grant DMS-1507511. The research of Z.M. was supported in part by the NSF CAREER grant DMS-1352060 and a Sloan Research Fellowship.

References

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015a.

Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015b.

Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in Neural Information Processing Systems*, pages 676–684, 2015c.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Noga Alon and Joel H Spencer. *The Probabilistic Method*. John Wiley & Sons, 2004.

Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*, 2014.

Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106 (50):21068–21073, 2009.

T Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *arXiv preprint arXiv:1404.6000*, 2014.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021*, 2015.

David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

Joel Friedman, Jeff Kahn, and Endre Szemeredi. On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 587–598. ACM, 1989.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 11(9):1074–1085, 1992.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.

Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

Varun Jog and Po-Ling Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *arXiv preprint arXiv:1312.1733*, 2013.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings-Annual Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

Can M Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: Regularization and concentration of the Laplacian. *arXiv preprint arXiv:1502.03049*, 2015.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2014.

Jing Lei and Lingxue Zhu. A generic sample splitting approach for refined community recovery in stochastic block models. *arXiv preprint arXiv:1411.1469*, 2014.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. In *WALCOM: Algorithms and Computation*, pages 274–285. Springer, 2009.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

Elchanan Mossel and Jiaming Xu. Density evolution in the degree-correlated stochastic block model. *arXiv preprint arXiv:1509.03281*, 7, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction, and optimal recovery of block models. *arXiv preprint arXiv:1309.1380*, 2013a.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013b.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.

Mark EJ Newman and Elizabeth A Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569, 2007.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

Alfred Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Purnamrita Sarkar and Peter J Bickel. Role of normalization in spectral clustering for stochastic blockmodels. *arXiv preprint arXiv:1310.1495*, 2013.

Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

Van Vu. A simple SVD algorithm for finding hidden partitions. *arXiv preprint arXiv:1404.3918*, 2014.

Stanley Wasserman. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, 1994.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014a.

Se-Young Yun and Alexandre Proutiere. Community detection via random and adaptive sampling. In *COLT*, pages 138–175, 2014b.

Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. *ArXiv e-prints*, 5, 2015.

Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block models. *arXiv preprint arXiv:1507.05313*, 2015.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40 (4):2266–2292, 2012.