Achieving Exact Cluster Recovery Threshold via Semidefinite Programming

Bruce Hajek Yihong Wu Jiaming Xu

***Abstract*—The binary symmetric stochastic block model deals with a random graph of n vertices partitioned into two equal-sized clusters, such that each pair of vertices is connected independently with probability p within clusters and q across clusters. In the asymptotic regime of $p = a\log n/n$ and $q = b\log n/n$ for fixed a, b and $n \to \infty$, we show that the semidefinite programming relaxation of the maximum likelihood estimator achieves the optimal threshold for exactly recovering the partition from the graph with probability tending to one, resolving a conjecture of Abbe et al. [1]. Furthermore, we show that the semidefinite programming relaxation also achieves the optimal recovery threshold in the planted dense subgraph model containing a single cluster of size proportional to n.**

I. Introduction

The community detection problem refers to finding the underlying communities within a network using only knowledge of the network topology [8]. This paper considers the following probabilistic model for generating a network with underlying community structures: Suppose that out of a total of n vertices, rK of them are partitioned into r clusters of size K, and the remaining $n - rK$ vertices do not belong to any clusters (called outlier vertices); a random graph G is generated based on the cluster structure, where each pair of vertices is connected independently with probability p if they are in the same cluster or q otherwise. This random graph ensemble is known as the *planted cluster model* [3] with parameters $n, r, K \in \mathbb{N}$ and $p, q \in [0, 1]$ such that $n \geq rK$. In particular, we call p and q the in-cluster and cross-cluster edge density, respectively. In the special case with no outlier vertices, i.e., $n = rK$, the planted cluster model reduces to the classical stochastic block model [12], also known as the planted partition model [4]. The planted cluster model and its special cases have been widely used for studying the community detection and graph partitioning problem (see, e.g., [14], [5], [15], [2] and the references therein). In this paper, we focus on the following particular cases in the asymptotic regime $n \to \infty$:

- Binary symmetric stochastic block model (assuming n is even):

$$r = 2,\ K = \frac{n}{2},\ p = \frac{a\log n}{n},\ q = \frac{b\log n}{n}, \qquad (1)$$

This research was supported by the National Science Foundation under Grant ECCS 10-28464, IIS-1447879, and CCF-1423088, and Strategic Research Initiative on Big-Data Analytics of the College of Engineering at the University of Illinois. The authors are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. Email:{b-hajek,yihongwu,jxu18}@illinois.edu.

- Planted dense subgraph model:

$$r = 1,\ K = \lfloor \rho n \rfloor,\ p = \frac{a\log n}{n},\ q = \frac{b\log n}{n}, \qquad (2)$$

where $a \neq b$ and $0 < \rho < 1$ are fixed constants, and study the problem of exactly recovering the clusters (up to a permutation of cluster indices) from the observation of the graph G .

Exact cluster recovery under the binary symmetric stochastic block model is studied in [1], [17] and a sharp recovery threshold is found.

Theorem 1 ([1], [17])**.** *Under the binary symmetric stochastic block model (1), if $(\sqrt{a} - \sqrt{b})^2 > 2$, clusters can be exactly recovered up to a permutation of cluster indices with probability converging to 1; if $(\sqrt{a} - \sqrt{b})^2 < 2$, no algorithm can exactly recover the clusters with probability converging to 1.*

The optimal reconstruction threshold in Theorem 1 is achieved by the maximum likelihood (ML) estimator, which entails finding the minimum bisection of the graph, a problem known to be NP-hard in the worst case [9, Theorem 1.3]. Nevertheless, it has been shown that the optimal recovery threshold can be attained in polynomial time using a two-step procedure [1], [17]: First, apply the partial recovery algorithms in [16], [13] to correctly cluster all but $o(n)$ vertices; Second, flip the cluster memberships of those vertices who do not agree with the majority of their neighbors. This two-step procedure has two limitations: a) the partial recovery algorithms used in the first step are sophisticated; b) the original graph needs to be split to implement the two steps to ensure their independence. It remains open to find a simple direct approach to achieve the exact recovery threshold in polynomial time. It was proved in [1] that a semidefinite programming (SDP) relaxation of the ML estimator succeeds if $(a - b)^2 > 8(a + b) + 8/3(a - b)$. Backed by compelling simulation results, it was further conjectured in [1] that the SDP relaxation can achieve the optimal recovery threshold. In this paper, we resolve this conjecture in the positive.

In addition, we prove that the SDP relaxation achieves the optimal recovery threshold for the planted dense subgraph model (2) where the cluster size K scales *linearly* in n. This conclusion is in sharp contrast to the following computational barrier established in [11]: If K grows and p, q decay *sublinearly* in n, attaining the statistical optimal recovery threshold is at least as hard as solving the planted clique problem (See Section III for detailed discussions).

Notation: Let A denote the adjacency matrix of the graph G, $\mathbf{I}$ denote the identity matrix, and $\mathbf{J}$ denote the all-one matrix. We write $X \succeq 0$ if X is positive semidefinite and $X \geq 0$ if all the entries of X are non-negative. Let $\mathcal{S}^n$ denote the set of all $n \times n$ symmetric matrices. For $X \in \mathcal{S}^n$, let $\lambda_2(X)$ denote its second smallest eigenvalue. For any matrix Y, let $\|Y\|$ denote its spectral norm. For any positive integer n, let $[n] = \{1, \ldots, n\}$. For any set $T \subset [n]$, let $|T|$ denote its cardinality and T^c denote its complement. We use standard big O notations, e.g., for any sequences $\{a_n\}$ and $\{b_n\}$, $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if there is an absolute constant $c > 0$ such that $1/c \leq a_n/b_n \leq c$. Let $\mathrm{Bern}(p)$ denote the Bernoulli distribution with mean p and $\mathrm{Binom}(N, p)$ denote the binomial distribution with N trials and success probability p. All logarithms are natural and we use the convention $0 \log 0 = 0$.

II. Stochastic block model

The cluster structure under the binary symmetric stochastic block model can be represented by a vector $\sigma \in \{\pm 1\}^n$ such that $\sigma_i = 1$ if vertex i is in the first cluster and $\sigma_i = -1$ otherwise. Let σ^* correspond to the true clusters. Then the ML estimator of σ^* for the case $a > b$ can be simply stated as

$$\begin{aligned} \max_{\sigma} \ & \sum_{i,j} A_{ij}\sigma_i\sigma_j \\ \text{s.t.} \ & \sigma_i \in \{\pm 1\}, \quad i \in [n] \\ & \sigma^\top \mathbf{1} = 0, \end{aligned} \tag{3}$$

which maximizes the number of in-cluster edges minus the number of out-cluster edges. This is equivalent to solving the NP-hard minimum graph bisection problem. Instead, let us consider its convex relaxation similar to the SDP relaxation studied in [1]. Let $Y = \sigma\sigma^\top$. Then $Y_{ii} = 1$ is equivalent to $\sigma_i = \pm 1$ and $\sigma^\top \mathbf{1} = 0$ if and only if $\langle Y, \mathbf{J} \rangle = 0$. Therefore, (3) can be recast as

$$\begin{aligned} \max_{Y,\sigma} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y = \sigma\sigma^\top \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle \mathbf{J}, Y \rangle = 0. \end{aligned} \tag{4}$$

Notice that the matrix $Y = \sigma\sigma^\top$ is a rank-one positive semidefinite matrix. If we relax this condition by dropping the rank-one restriction, we obtain the following convex relaxation of (4), which is a semidefinite program:

$$\begin{aligned} \widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle \mathbf{J}, Y \rangle = 0. \end{aligned} \tag{5}$$

We remark that (5) does not rely on any knowledge of the model parameters except that $a > b$; for the case $a < b$, we replace $\arg\max$ in (5) by $\arg\min$.

Let $Y^* = \sigma^*(\sigma^*)^\top$ and $\mathcal{Y}_n \triangleq \{\sigma\sigma^\top : \sigma \in \{-1, 1\}^n, \sigma^\top \mathbf{1} = 0\}$. The following result establishes the optimality of the SDP procedure:

Theorem 2. *If* $(\sqrt{a} - \sqrt{b})^2 > 2$, *then* $\min_{Y^* \in \mathcal{Y}_n} \mathbb{P}\{\widehat{Y}_{\mathrm{SDP}} = Y^*\} \to 1$ *as* $n \to \infty$.

III. Planted dense subgraph model

In this section we turn to the planted dense subgraph model in the asymptotic regime (2), where there exists a single cluster of size $\lfloor \rho N \rfloor$. To specify the optimal reconstruction threshold, define the following function: For $a, b \geq 0$, let

$$f(a,b) = \begin{cases} a - \tau^* \log \frac{ea}{\tau^*} & \text{if } a, b > 0, a \neq b \\ a & \text{if } b = 0 \\ b & \text{if } a = 0 \\ 0 & \text{if } a = b \end{cases}, \tag{6}$$

where $\tau^* \triangleq \frac{a-b}{\log a - \log b}$ if $a, b > 0$ and $a \neq b$. We show that if $\rho f(a,b) > 1$, exact recovery is achievable in polynomial-time via SDP with probability tending to one; if $\rho f(a,b) < 1$, any estimator fails to recover the cluster with probability tending to one regardless of the computational costs. The sharp threshold $\rho f(a,b) = 1$ is plotted in Fig. 1 for various values of ρ.

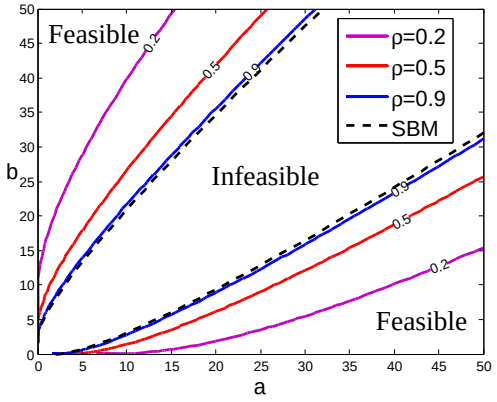

Fig. 1: The recovery threshold: $\rho f(a,b) = 1$ (solid curves) for the planted dense subgraph model (2); $(\sqrt{a} - \sqrt{b})^2 = 2$ (dashed curve) for the stochastic block model (1).

We first introduce the maximum likelihood estimator and its convex relaxation. For ease of notation, in this section we use a vector $\xi \in \{0,1\}^n$, as opposed to $\sigma \in \{\pm 1\}^n$ used in Section II for the SBM, as the indicator function of the cluster, such that $\xi_i = 1$ if vertex i is in the cluster and $\xi_i = 0$ otherwise. Let ξ^* be the indicator of the true cluster. Assuming $a > b$, i.e., the nodes in the cluster are more densely connected, the ML estimation of ξ^* is simply

$$\begin{aligned} \max_{\xi} \ & \sum_{i,j} A_{ij}\xi_i\xi_j \\ \text{s.t.} \ & \xi \in \{0,1\}^n \\ & \xi^\top \mathbf{1} = K, \end{aligned} \tag{7}$$

which maximizes the number of in-cluster edges. Due to the integrality constraints, it is computationally difficult to solve

(7), which prompts us to consider its convex relaxation. Note that (7) can be equivalently[1] formulated as

$$\begin{aligned} \max_{Z,\xi} \ & \langle A, Z\rangle \\ \text{s.t.} \ & Z = \xi\xi^\top \\ & Z_{ii} \le 1, \quad \forall i \in [n] \\ & Z_{ij} \ge 0, \quad \forall i,j \in [n] \\ & \langle \mathbf{I}, Z\rangle = K \\ & \langle \mathbf{J}, Z\rangle = K^2, \end{aligned} \tag{8}$$

where the matrix $Z = \xi\xi^\top$ is positive semidefinite and rank-one. Removing the rank-one restriction leads to the following convex relaxation of (8), which is a semidefinite program.

$$\begin{aligned} \widehat{Z}_{\mathrm{SDP}} = \arg\max_{Z} \ & \langle A, Z\rangle \\ \text{s.t.} \ & Z \succeq 0 \\ & Z_{ii} \le 1, \quad \forall i \in [n] \\ & Z_{ij} \ge 0, \quad \forall i,j \in [n] \\ & \langle \mathbf{I}, Z\rangle = K \\ & \langle \mathbf{J}, Z\rangle = K^2. \end{aligned} \tag{9}$$

We note that, apart from the assumption that $a > b$, the only model parameter needed by the estimator (9) is the cluster size K; for the case $a < b$, we replace $\arg\max$ in (9) by $\arg\min$.

Let $Z^* = \xi^*(\xi^*)^\top$ correspond to the true cluster and define $\mathcal{Z}_n = \{\xi\xi^\top : \xi \in \{0,1\}^n, \xi^\top \mathbf{1} = K\}$. The recovery threshold for the SDP (9) is given as follows.

Theorem 3. *Under the planted dense subgraph model (2), if*

$$\rho f(a,b) > 1, \tag{10}$$

then $\min_{Z^* \in \mathcal{Z}_n} \mathbb{P}\{\widehat{Z}_{\mathrm{SDP}} = Z^*\} \to 1$ *as* $n \to \infty$.

Next we prove a converse for Theorem 3 which shows that the recovery threshold achieved by the SDP relaxation is in fact optimal.

Theorem 4. *Under the planted dense subgraph model (2), if*

$$\rho f(a,b) < 1, \tag{11}$$

and the true cluster is uniformly chosen among all K-subsets of $[n]$, then for any sequence of estimators $\widehat{Z}_n$, $\mathbb{P}\{\widehat{Z}_n = Z^*\} \to 0$ as $n \to \infty$.

Under the planted dense subgraph model, our investigation of the exact cluster recovery problem thus far in this paper has been focused on the regime where the cluster size K grows **linearly** with n and $p, q = \Theta(\frac{\log n}{n})$, where the statistically optimal threshold can be attained by SDP in polynomial time. However, this need *not* be the case if K grows **sublinearly** in n. In fact, the exact cluster recovery problem has been studied in [3], [11] in the following asymptotic regime:

$$K = \Theta(n^\beta), \ p = cq = \Theta(n^{-\alpha}), \quad n \to \infty, \tag{12}$$

where $c > 1$ and $\alpha, \beta \in (0,1)$ are fixed constants. The statistical and computational complexities of the cluster recovery problem depend crucially on the value of α and β (see [11, Figure 2] for an illustration):

- $\beta > \frac{1}{2} + \frac{\alpha}{2}$: the planted cluster can be perfectly recovered in polynomial-time with high probability via the SDP relaxation (9).[2]
- $\frac{1}{2} + \frac{\alpha}{4} < \beta < \frac{1}{2} + \frac{\alpha}{2}$: the planted cluster can be detected in linear time with high probability by thresholding the total number of edges, but it is conjectured to be computationally intractable to exactly recover the planted cluster.
- $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$: the planted cluster can be exactly recovered with high probability via ML estimation; however, no randomized polynomial-time solver exists conditioned on the planted clique hardness hypothesis.[3]
- $\beta < \alpha$: regardless of the computational costs, no algorithm can exactly recover the planted cluster with vanishing probability of error.

Consequently, assuming the planted clique hardness hypothesis, in the asymptotic regime of (12) when $\alpha \in (0, \frac{2}{3})$ (and, quite possibly, the entire range $(0,1)$), there exists a significant gap between the information limit (recovery threshold of the optimal procedure) and the computational limit (recovery threshold for polynomial-time algorithms). In contrast, in the asymptotic regime of (2), the computational constraint imposes no penalty on the statistical performance, in that the optimal threshold can be attained by SDP relaxation in view of Theorem 3.

IV. Proofs

The proofs of our main theorems are sketched. The excluded proofs can be found in the full paper [10]. Our analysis of the SDP relies on two key ingredients: the spectrum of Erdős-Rényi random graphs and tail bounds for the binomial distributions, which we first present.

A. Spectrum of Erdős-Rényi random graph

Let A denote the adjacency matrix of an Erdős-Rényi random graph G, where nodes i and j are connected independently with probability p_{ij}. Then $\mathbb{E}[A_{ij}] = p_{ij}$. Let $p = \max_{ij} p_{ij}$ and assume $p \ge c_0 \frac{\log n}{n}$ for any constant $c_0 > 0$. We aim to show that $\|A - \mathbb{E}[A]\|_2 \le c'\sqrt{np}$ with high probability for some constant $c' > 0$. To this end, we establish the following more general result where the entries need not be binary-valued.

[1] Here (7) and (8) are equivalent in the following sense: for any feasible ξ for (7), $Z = \xi\xi^\top$ is feasible for (8); for any feasible Z, ξ for (8), either ξ or $-\xi$ is feasible for (7).

[2] In fact, an even looser SDP relaxation than (9) has been shown to exactly recover the planted cluster with high probability for $\beta > \frac{1}{2} + \frac{\alpha}{2}$. See [3, Theorem 2.3].

[3] Here the planted clique hardness hypothesis refers to the statement that for any fixed constants $\gamma > 0$ and $\delta > 0$, there exist no randomized polynomial-time tests to distinguish an Erdős-Rényi random graph $\mathcal{G}(n,\gamma)$ and a planted clique model which is obtained by adding edges to $k = n^{1/2-\delta}$ vertices chosen uniformly from $\mathcal{G}(n,\gamma)$ to form a clique. For various hardness results of problems reducible from the planted clique problem, see [11] and the references within.

Theorem 5. *Let A denote a symmetric and zero-diagonal random matrix, where the entries $\{A_{ij} : i < j\}$ are independent and $[0,1]$-valued. Assume that $\mathbb{E}[A_{ij}] \le p$, where $c_0 \log n/n \le p \le 1 - c_1$ for arbitrary constants $c_0 > 0$ and $c_1 > 0$. Then for any $c > 0$, there exists $c' > 0$ such that for any $n \ge 1$, $\mathbb{P}\{\|A - \mathbb{E}[A]\|_2 \le c'\sqrt{np}\} \ge 1 - n^{-c}$.*

Let $\mathcal{G}(n,p)$ denote the Erdős-Rényi random graph model with the edge probability $p_{ij} = p$ for all i, j. Results similar to Theorem 5 have been obtained in [7] for the special case of $\mathcal{G}(n, \frac{c_0 \log n}{n})$ for some *sufficiently large* c_0. In fact, Theorem 5 can be proved by strengthening the combinatorial arguments in [7, Section 2.2]. We provide an alternative proof using results from random matrices and concentration of measures and a seconder-order stochastic comparison argument from [18].

Furthermore, we note that the condition $p = \Omega(\log n/n)$ in Theorem 5 is in fact necessary to ensure that $\|A - \mathbb{E}[A]\|_2 = \Omega_{\mathbb{P}}(\sqrt{np})$ (see [11, Appendix A] for a proof). The condition $p \le 1 - c_1$ can be dropped in the special case of $\mathcal{G}(n,p)$.

B. Tail of the Binomial Distribution

Let $X \sim \mathrm{Binom}\left(m, \frac{a \log n}{n}\right)$ and $R \sim \mathrm{Binom}\left(m, \frac{b \log n}{n}\right)$ for $m \in \mathbb{N}$ and $a, b > 0$, where $m = \rho n + o(n)$ for some $\rho > 0$ as $n \to \infty$. We need the following tail bounds.

Lemma 1 ([1])**.** *Assume that $a > b$ and $k_n \in \mathbb{N}$ such that $k_n = (1 + o(1))\frac{\log n}{\log\log n}$. Then*

$$\mathbb{P}\{X - R \le k_n\} \le n^{-\rho\left(\sqrt{a} - \sqrt{b}\right)^2 + o(1)}.$$

Lemma 2. *Let $k_n, k_n' \in [m]$ be such that $k_n = \tau\rho \log n + o(\log n)$ and $k_n' = \tau'\rho \log n + o(\log n)$ for some $0 \le \tau \le a$ and $\tau' \ge b$. Then*

$$\mathbb{P}\{X \le k_n\} = n^{-\rho\left(a - \tau \log \frac{ea}{\tau} + o(1)\right)} \tag{13}$$

$$\mathbb{P}\{R \ge k_n'\} = n^{-\rho\left(b - \tau' \log \frac{eb}{\tau'} + o(1)\right)}. \tag{14}$$

C. Proof of Theorem 2

The following lemma provides a deterministic sufficient condition for the success of SDP (5) in the case $a > b$.

Lemma 3. *Suppose there exist $D^* = \mathsf{diag}\{d_i^*\}$ and $\lambda^* \in \mathbb{R}$ such that $S^* \triangleq D^* - A + \lambda^* \mathbf{J}$ satisfies $S^* \succeq 0$, $\lambda_2(S^*) > 0$ and*

$$S^* \sigma^* = 0. \tag{15}$$

Then $\widehat{Y}_{\mathrm{SDP}} = Y^*$ is the unique solution to (5).

Proof: The Lagrangian function is given by

$$L(Y, S, D, \lambda) = \langle A, Y\rangle + \langle S, Y\rangle - \langle D, Y - \mathbf{I}\rangle - \lambda\langle \mathbf{J}, Y\rangle,$$

where the Lagrangian multipliers are denoted by $S \succeq 0$, $D = \mathsf{diag}\{d_i\}$, and $\lambda \in \mathbb{R}$. Then for any Y satisfying the constraints in (5),

$$\begin{aligned}\langle A, Y\rangle &\overset{(a)}{\le} L(Y, S^*, D^*, \lambda^*) = \langle D^*, I\rangle = \langle D^*, Y^*\rangle \\ &= \langle A + S^* - \lambda^* \mathbf{J}, Y^*\rangle \overset{(b)}{=} \langle A, Y^*\rangle,\end{aligned}$$

where (a) holds because $\langle S^*, Y\rangle \ge 0$; (b) holds because $\langle Y^*, S^*\rangle = (\sigma^*)^\top S^* \sigma^* = 0$ by (15). Hence, Y^* is an optimal solution. It remains to establish its uniqueness. To this end, suppose $\widetilde{Y}$ is an optimal solution. Then,

$$\begin{aligned}\langle S^*, \widetilde{Y}\rangle &= \langle D^* - A + \lambda^* \mathbf{J}, \widetilde{Y}\rangle \overset{(a)}{=} \langle D^* - A, \widetilde{Y}\rangle \\ &\overset{(b)}{=} \langle D^* - A, Y^*\rangle = \langle S^*, Y^*\rangle = 0.\end{aligned}$$

where (a) holds because $\langle \mathbf{J}, \widetilde{Y}\rangle = 0$; (b) holds because $\langle A, \widetilde{Y}\rangle = \langle A, Y^*\rangle$ and $\widetilde{Y}_{ii} = Y^*_{ii} = 1$ for all $i \in [n]$. In view of (15), since $\widetilde{Y} \succeq 0$, $S^* \succeq 0$ with $\lambda_2(S^*) > 0$, $\widetilde{Y}$ must be a multiple of $Y^* = \sigma^*(\sigma^*)^\top$. Because $\widetilde{Y}_{ii} = 1$ for all $i \in [n]$, $\widetilde{Y} = Y^*$. ∎

Proof of Theorem 2: The theorem is proved first for $a > b$. Let $D^* = \mathsf{diag}\{d_i^*\}$ with

$$d_i^* = \sum_{j=1}^n A_{ij}\sigma_i^*\sigma_j^* \tag{16}$$

and choose any $\lambda^* \ge \frac{p+q}{2}$. It suffices to show that $S^* = D^* - A + \lambda^* \mathbf{J}$ satisfies the conditions in Lemma 3 with high probability.

By definition, $d_i^*\sigma_i^* = \sum_j A_{ij}\sigma_j^*$ for all i, i.e., $D^*\sigma^* = A\sigma^*$. Since $\mathbf{J}\sigma^* = 0$, (15) holds, that is, $S^*\sigma^* = 0$. It remains to verify that $S^* \succeq 0$ and $\lambda_2(S^*) > 0$ with probability converging to one, which amounts to showing that

$$\mathbb{P}\left\{\inf_{x \perp \sigma^*, \|x\|_2 = 1} x^\top S^* x > 0\right\} \to 1. \tag{17}$$

Note that $\mathbb{E}[A] = \frac{p-q}{2}Y^* + \frac{p+q}{2}\mathbf{J} - p\mathbf{I}$ and $Y^* = \sigma^*(\sigma^*)^\top$. Thus for any x such that $x \perp \sigma^*$ and $\|x\|_2 = 1$,

$$\begin{aligned}&x^\top S^* x \\ &= x^\top D^* x - x^\top \mathbb{E}[A]\, x + \lambda^* x^\top \mathbf{J} x - x^\top (A - \mathbb{E}[A])\, x \\ &= x^\top D^* x - \frac{p-q}{2} x^\top Y^* x + \left(\lambda^* - \frac{p+q}{2}\right) x^\top \mathbf{J} x + p \\ &\quad - x^\top (A - \mathbb{E}[A])\, x \\ &\overset{(a)}{\ge} x^\top D^* x + p - x^\top (A - \mathbb{E}[A])\, x \\ &\ge \min_{i \in [n]} d_i^* + p - \|A - \mathbb{E}[A]\|.\end{aligned} \tag{18}$$

where (a) holds since $\lambda^* \ge \frac{p+q}{2}$ and $\langle x, \sigma^*\rangle = 0$. It follows from Theorem 5 that $\|A - \mathbb{E}[A]\| \le c'\sqrt{\log n}$ with high probability for a positive constant c' depending only on a. Moreover, note that each d_i is equal in distribution to $X - R$, where $X \sim \mathrm{Binom}(\frac{n}{2} - 1, \frac{a \log n}{n})$ and $R \sim \mathrm{Binom}(\frac{n}{2}, \frac{b \log n}{n})$ are independent. Hence, Lemma 1 implies that

$$\mathbb{P}\left\{X - R \ge \frac{\log n}{\log\log n}\right\} \ge 1 - n^{-(\sqrt{a} - \sqrt{b})^2/2 + o(1)}.$$

Applying the union bound implies that $\min_{i \in [n]} d_i^* \ge \frac{\log n}{\log\log n}$ holds with probability at least $1 - n^{1 - (\sqrt{a} - \sqrt{b})^2/2 + o(1)}$. It follows from the assumption $(\sqrt{a} - \sqrt{b})^2 > 2$ and (18) that the desired (17) holds, completing the proof in the case $a > b$. For the case $a < b$, we replace the $\arg\max$ by $\arg\min$ in the SDP

(5), which is equivalent to substituting $-A$ for A in the original maximization problem, as well as the sufficient condition in Lemma 3. Set the dual variable d_i^* according to (16) with $-A$ replacing A and choose any $\lambda^* \geq -\frac{p+q}{2}$. Then (15) still holds and (18) changes to $x^\top S^* x \geq \min_{i\in[n]} d_i^* - p - \|A - \mathbb{E}[A]\|$, where $\min_{i\in[n]} d_i^* \geq \frac{\log n}{\log\log n}$ holds with probability at least $1 - n^{1-(\sqrt{a}-\sqrt{b})^2/2+o(1)}$ by Lemma 1 and the union bound. Therefore, in view of Theorem 5 and the assumption $(\sqrt{a} - \sqrt{b})^2 > 2$, the desired (17) still holds, completing the proof for the case $a < b$.

■

D. Proof of Theorem 3

Lemma 4. *Suppose there exist $D^* = \mathsf{diag}\{d_i^*\} \geq 0$, $B^* \in \mathcal{S}^n$ with $B^* \geq 0$, $\lambda^* \in \mathbb{R}$, and $\eta^* \in \mathbb{R}$ such that $S^* \triangleq D^* - B^* - A + \eta^* \mathbf{I} + \lambda^* \mathbf{J}$ satisfies $S^* \succeq 0$, $\lambda_2(S^*) > 0$, and*

$$\begin{aligned} S^*\xi^* &= 0, \\ d_i^*(Z_{ii}^* - 1) &= 0, \quad \forall i, \\ B_{ij}^* Z_{ij}^* &= 0, \quad \forall i, j. \end{aligned} \tag{19}$$

Then $\widehat{Z}_{\mathrm{SDP}} = Z^*$ is the unique solution to (9).

The theorem is proved first for $a > b$. Recall $\tau^* = \frac{a-b}{\log a - \log b}$ if $a, b > 0$ and $a \neq b$. Let $\tau^* = 0$ if $a = 0$ or $b = 0$. Choose $\lambda^* = \tau^* \log n/n$, $\eta^* = \|A - \mathbb{E}[A]\|$, $D^* = \mathsf{diag}\{d_i^*\}$ with

$$d_i^* = \begin{cases} \sum_{j\in C^*} A_{ij} - \eta^* - \lambda^* K & \text{if } i \in C^* \\ 0 & \text{otherwise} \end{cases}.$$

Define $b_i^* \triangleq \lambda^* - \frac{1}{K}\sum_{j\in C^*} A_{ij}$ for $i \notin C^*$. Let $B^* \in \mathcal{S}^n$ be given by

$$B_{ij}^* = b_i^* \mathbf{1}_{\{i\notin C^*, j\in C^*\}} + b_j^* \mathbf{1}_{\{i\in C^*, j\notin C^*\}}.$$

The following claims show that (S^*, D^*, B^*) satisfies the conditions in Lemma 4 with probability tending to one, and hence the theorem follows in the case $a > b$ in view of Lemma 4.

Claim 1. (S^*, D^*, B^*) satisfies (19).

Claim 2. With probability converging to 1, $D^* \geq 0$, $B^* \geq 0$.

Claim 3. With probability converging to 1, $S^* \succeq 0$ with $\lambda_2(S^*) > 0$.

For the case $a < b$, the proof is similar and can be bound in [10].

E. Proof of Theorem 4

If $a = b$, then the cluster is unidentifiable from the graph. Next, we prove the theorem first for the case $a > b$. If $b = 0$, then perfect recovery is possible if and only if the subgraph formed by the nodes in cluster, which is $\mathcal{G}(K, a\log n/n)$, contains no isolated node.[4] This occurs with high probability if $\rho a < 1$ [6]. Next we consider $a > b > 0$.

[4]To be more precise, if there is an isolated node in the cluster C^*, then the likelihood has at least $n - K$ maximizers, which, in turn, implies that the probability of exact recovery for any estimator is at most $\frac{1}{n-K}$.

Since the prior distribution of the true cluster C^* is uniform, the ML estimator minimizes the error probability among all estimators and thus we only need to find when the ML estimator fails. Let $e(i, S) \triangleq \sum_{j\in S} A_{ij}$ denote the number of edges between node i and nodes in $S \subset [n]$. Let F denote the event that $\min_{i\in C^*} e(i, C^*) < \max_{j\notin C^*} e(j, C^*)$, which implies the existence of $i \in C^*$ and $j \notin C^*$, such that the set $C^*\backslash\{i\} \cup \{j\}$ achieves a strictly higher likelihood than C^*. Hence $\mathbb{P}\{\text{ML fails}\} \geq \mathbb{P}\{F\}$. Next we bound $\mathbb{P}\{F\}$ from below.

By symmetry, we can condition on C^* being the first K nodes. Let T denote the set of first $\lfloor \frac{\rho n}{\log^2 n} \rfloor$ nodes. Then

$$\min_{i\in C^*} e(i, C^*) \leq \min_{i\in T} e(i, C^*) \leq \min_{i\in T} e(i, C^*\backslash T) + \max_{i\in T} e(i, T). \tag{20}$$

Let E_1, E_2, E_3 denote the event that $\max_{i\in T} e(i, T) < \frac{\log n}{\log\log n}$, $\min_{i\in T} e(i, C^*\backslash T) + \frac{\log n}{\log\log n} \leq \tau^*\rho\log n$ and $\max_{j\notin C^*} e(j, C^*) \geq \tau^*\rho\log n$, respectively. In view of (20), we have $F \supset E_1 \cap E_2 \cap E_3$. Then, $\mathbb{P}\{F\} \to 1$ due to Claim 4, completing the proof in the case $a > b > 0$.

Claim 4. $\mathbb{P}\{E_i\} \to 1$ for $i = 1, 2, 3$.

The proof of the theorem for the case $a < b$ is similar and can be found in [10].

References

[1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv:1405.3267*, 2014.

[2] Y. Chen, S. Sanghavi, and H. Xu. Clustering sparse graphs. *arXiv:1210.3335*, 2012.

[3] Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. 2014, available at http://arxiv.org/abs/1402.1267.

[4] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms*, 18(2):116–140, Mar 2001.

[5] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physics Review E*, 84:066106, 2011.

[6] P. Erdös and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[7] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, Sept. 2005.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[9] M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theoret. Comput. Sci.*, 1(3):237–267, 1976.

[10] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv:1412.6156*, Nov. 2014.

[11] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. *arXiv:1406.6625*, 2014.

[12] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[13] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *STOC 2014: 46th Annual Symposium on the Theory of Computing*, pages 1–10, New York, United States, June 2014.

[14] F. McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529 – 537, 2001.

[15] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *available at: http://arxiv.org/abs/1202.1499*, 2012.

[16] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arxiv:1311.4115*, 2013.

[17] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv:1407.1591*, 2014.

[18] D.-C. Tomozei and L. Massouli. Distributed user profiling via spectral methods. *Stochastic Systems*, 4(1):1–43, 2014.