

Community Detection with Secondary Latent Variables

Mohammad Esmaeili and Aria Nosratinia

Department of Electrical and Computer Engineering, The University of Texas at Dallas

Email: {Esmaeili, Aria}@utdallas.edu

Abstract—Community detection refers to recovering a (latent) label on which the distribution of the observed graph depends. Recent work has also investigated the impact of additionally knowing the value of another variable at each vertex that is correlated with the vertex label (side information), while assuming side information is independent of the graph edges conditioned on the label. This work extends the scope of community detection in two ways. First, we consider a side information that does not form a Markov chain with the label and graph, and analyze the detection threshold of semidefinite programming subject to knowledge of this side information, which is a non-label latent variable on which the graph edges also depend. In the second part of the work, we consider aside from vertex labels a second latent variable that is unknown both in realization and in distribution. We then investigate the performance of the semidefinite programming community detection as a function of the (unknown) composition of the nuisance latent variable. In both cases, it is shown that semidefinite programming can achieve exact recovery down to the optimal (information theoretic) threshold.

I. INTRODUCTION

Community detection refers to grouping (clustering) vertices of graph having similar affiliations. Community detection has been investigated widely, see e.g. [1] and references therein. Information-theoretic threshold of community detection has been found for several random graph models, e.g., stochastic block model, censored model, and exponential random graph model.

In many applications, aside from the graph observation, non-graphical side information about communities (clusters) is also available [2]–[4]. The effect of side information on the community detection threshold has been studied for partially revealed labels, noisy labels, and more extensive side information consisting of multiple features [2], [5], [6]. These models have assumed that, conditioned on the graph labels that we wish to recover, the graph and side information are independent. In other words, these early works have assumed that side information provides fresh information about the labels over and above the graph, and does not repeat or overlap the same information in the graph, concerning the desired labels.

In some applications, while the graph depends on both the community label as well as another variable, the second variable may be unknowable and its distribution may also be unknown. In these cases, the second variable is essentially a nuisance variable for the purposes of detecting labels from the graph,

however, its unknown distribution may in fact have an effect on the detection threshold.

In this paper, we consider a model where graph edges depend on two latent random variables, and investigate the following two problems:

- Recovering the label for each vertex, assuming the second latent variable (side information) is revealed. For this part of the proposed work, we relax the Markov assumption mentioned earlier, so that the graph and side information are no longer independent conditioned on the side information.
- Recovering the first latent variable for each vertex, while the second latent variable is unknown. In this case, we investigate the threshold of detection as a function of the composition of the nuisance latent variable.

This paper investigates the semidefinite programming (SDP) detector, showing in the process that it can achieve exact recovery down to the optimal threshold in each case.

The practical significance of the first component of this work is to render more realistic the role of the side information as compared with earlier work. In most practical scenarios of interest, a side information variable may not be independent of the graph subject to the vertex label, i.e., the information available via the connective and individual attributes of the vertices may overlap. This work addresses such practical conditions. The value of the second component of this work is in showing how the performance of community detection is affected by the composition of parameters that are unknown and are not subject to detection.

II. NOTATION AND GRAPHICAL MODEL

$\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is the all-one matrix. $\mathbf{1}$ and $\mathbf{0}$ denote all-one and all-zero vectors, respectively. $S \succeq 0$ indicates a positive semidefinite matrix and $S \geq 0$ a matrix with non-negative entries. $\|S\|$ is the spectral norm and $\lambda_2(S)$ is the second smallest eigenvalue (for a symmetric matrix). $\langle \cdot, \cdot \rangle$ is the inner product and $*$ is the element-wise product. We abbreviate $[n] \triangleq \{1, \dots, n\}$. Probabilities are denoted by $\mathbb{P}(\cdot)$. Random variables with Bernoulli and Binomial distributions are indicated by $\text{Bern}(p)$ and $\text{Bin}(n, p)$ respectively, with n trials and success probability p .

Let n denote the number of vertices in the graph. Each vertex v in the graph is related to two binary independent latent variables $x_v \in \{-1, +1\}$ and $y_v \in \{-1, +1\}$. For each vertex v , $\mathbb{P}\{x_v = +1\} = \rho_{1x}$, $\mathbb{P}\{x_v = -1\} = \rho_{2x}$, $\mathbb{P}\{y_v =$

This work was supported in part by the NSF grant 1711689.

$+1\} = \rho_{1y}$, and $\mathbb{P}\{y_v = -1\} = \rho_{2y}$. Also, for each vertex v , a set of latent variables $\sigma_v = \{x_v, y_v\}$ is assigned independently. Each pair of vertices (u, v) is independently connected to each other with probability Q_{σ_u, σ_v} that is specified by a symmetric 4×4 matrix Q with entries in $[0, 1]$. Also, P is an 1×4 vector with entries in $[0, 1]$ such that $\sum_i P_i = 1$. The elements of P determine the relative size of vertices with the same set of latent variables. This model can be easily generalized to K independent latent variables with various cardinalities.

Let $x = [x_1, x_2, \dots, x_n]^T$ and $y = [y_1, y_2, \dots, y_n]^T$. If A denote the symmetric adjacency matrix of the graph then for $i \in [n]$ and $i > j$ we have

$$A_{ij} \sim \begin{cases} \text{Bern}(q_{xy}) & \text{if } x_i = x_j, y_i = y_j \\ \text{Bern}(q_x) & \text{if } x_i = x_j, y_i \neq y_j \\ \text{Bern}(q_y) & \text{if } x_i \neq x_j, y_i = y_j \\ \text{Bern}(q) & \text{if } x_i \neq x_j, y_i \neq y_j \end{cases}$$

where $q_{xy} = a \frac{\log n}{n}$, $q_x = b \frac{\log n}{n}$, $q_y = c \frac{\log n}{n}$, $q = d \frac{\log n}{n}$, and $a \geq b \geq c \geq d > 0$. Then Q that is a symmetric 4×4 matrix is written as

$$Q = \begin{bmatrix} q_{xy} & q_x & q_y & q \\ q_x & q_{xy} & q & q_y \\ q_y & q & q_{xy} & q_x \\ q & q_y & q_x & q_{xy} \end{bmatrix}.$$

III. DETECTION BY SEMIDEFINITE PROGRAMMING

Let $W \triangleq yy^T$ and $B \triangleq W * A$, where y is an $n \times 1$ vector consists of second latent variables and A is the adjacency matrix of the graph. This paper considers two different scenarios.

A. One Unknown Binary Latent Variable

In the first scenario, given an observation of the graph and vector y , we exactly recover x that consists of the first latent variable for each vertex in the graph. Here, the vector y is considered as a side information to help estimator to recover the desired latent variables. In this model, it is assumed that the desired latent variables in x satisfy the constraint $x^T \mathbf{1} = 0$ which implies the existence of two equal sized communities in the graph. The log-likelihood function of the graph is written as

$$\log \mathbb{P}(A|x, y) = \frac{T_1}{8} x^T B x + \frac{T_2}{8} x^T A x + c,$$

where $T_1 = \log\left(\frac{ad}{bc}\right)$ and $T_2 = \log\left(\frac{ab}{cd}\right)$, as $n \rightarrow \infty$ and c is a constant. Then the maximum likelihood estimator can be stated as

$$\begin{aligned} \hat{x} = \arg \max_x & T_1 x^T B x + T_2 x^T A x \\ \text{subject to} & x_i \in \{\pm 1\}, \quad i \in [n] \\ & x^T \mathbf{1} = 0, \end{aligned} \quad (1)$$

that is a non-convex optimization problem. Let $Z = xx^T$. The maximum likelihood estimator (1) can be restated as

$$\begin{aligned} \hat{Z} = \arg \max_Z & \langle Z, T_1 B + T_2 A \rangle \\ \text{subject to} & Z = xx^T \\ & Z_{ii} = 1, \quad i \in [n] \\ & \langle Z, \mathbf{J} \rangle = 0. \end{aligned} \quad (2)$$

Relaxing the rank-one constraint, we obtain the following semidefinite programming relaxation of (2):

$$\begin{aligned} \hat{Z} = \arg \max_Z & \langle Z, T_1 B + T_2 A \rangle \\ \text{subject to} & Z \succeq 0 \\ & Z_{ii} = 1, \quad i \in [n] \\ & \langle Z, \mathbf{J} \rangle = 0. \end{aligned} \quad (3)$$

Let $Z^* = x^* x^{*T}$ corresponding to the true labels for desired latent variable and define $\mathcal{Z}_n \triangleq \{xx^T : x \in \{\pm 1\}^n, x^T \mathbf{1} = 0\}$. For convenience, define

$$\begin{aligned} \eta_1 &\triangleq \frac{\rho_{1y}}{2} (\sqrt{a} - \sqrt{c})^2 + \frac{\rho_{2y}}{2} (\sqrt{b} - \sqrt{d})^2, \\ \eta_2 &\triangleq \frac{\rho_{2y}}{2} (\sqrt{a} - \sqrt{c})^2 + \frac{\rho_{1y}}{2} (\sqrt{b} - \sqrt{d})^2. \end{aligned}$$

Theorem 1. For one unknown binary latent variable model with side information, if

- $\eta_1 > 1$ when $\rho_{1y} \leq 0.5$
- $\eta_2 > 1$ when $\rho_{1y} > 0.5$

then as $n \rightarrow \infty$, for semidefinite programming estimator $\hat{Z}_{SDP}$, $\min_{Z^* \in \mathcal{Z}_n} \mathbb{P}(\hat{Z}_{SDP} = Z^*) \geq 1 - o(1)$.

The following Theorem establishes the optimality of semidefinite programming procedure.

Theorem 2. For one unknown binary latent variable model with side information, if

- $\eta_1 < 1$ when $\rho_{1y} \leq 0.5$
- $\eta_2 < 1$ when $\rho_{1y} > 0.5$

then for any sequence of estimators $\hat{Z}_n$, $\mathbb{P}(\hat{Z}_n = Z^*) \rightarrow 0$ as $n \rightarrow \infty$.

B. Two Unknown Binary Latent Variables

In the second scenario, given an observation of the graph, the aim is to exactly recover x while both vectors x and y are unknown variables. In this model, it is assumed that the desired latent variables in x satisfy the constraint $x^T \mathbf{1} = 0$ which implies the existence of two equal sized communities in the graph. Also, it is assumed that the estimator does not know anything about the second latent variables in y and the prior distribution of y is uniform over $\{y : y \in \{\pm 1\}^n\}$. The log-likelihood function of A given x and y is obtained as

$$\begin{aligned} \log \mathbb{P}(A|x, y) &= \frac{T_1}{8} y^T (A * xx^T) y + \frac{T_2}{8} x^T A x \\ &\quad + \frac{T_3}{8} y^T A y + c, \end{aligned}$$

where $T_1 = \log\left(\frac{ad}{bc}\right)$, $T_2 = \log\left(\frac{ab}{cd}\right)$, and $T_3 = \log\left(\frac{ac}{bd}\right)$, as $n \rightarrow \infty$ and c is a constant. Then,

$$\begin{aligned} \log \mathbb{P}(A|x) &\propto \log \sum_{\mathcal{Y}} e^{\mathbb{P}(A|x,y)} \\ &\propto \log \sum_{\mathcal{Y}} e^{\frac{T_1}{T_3} y^T (A^* x x^T) y + \frac{T_2}{T_3} x^T A x + y^T A y} \\ &= \frac{T_1 + T_2}{T_3} x^T A x + \sum_i \sum_j A_{ij} \\ &\quad + \log \sum_{\mathcal{Y}} e^{\frac{T_1}{T_3} y^T (A^* x x^T) y + y^T A y - \frac{T_1}{T_3} x^T A x - \sum_i \sum_j A_{ij}}. \end{aligned}$$

Applying the log-sum-exp approximation, the maximum likelihood estimator is

$$\begin{aligned} \hat{x} &= \arg \max_x x^T A x \\ \text{subject to } & x_i \in \{\pm 1\}, \quad i \in [n] \\ & x^T \mathbf{1} = 0, \end{aligned} \quad (4)$$

that is a non-convex optimization problem. Let $Z = x x^T$. Then (4) can be restated as

$$\begin{aligned} \hat{x} &= \arg \max_x \langle Z, A \rangle \\ \text{subject to } & Z = x x^T \\ & Z_{ii} = 1, \quad i \in [n] \\ & \langle Z, \mathbf{J} \rangle = 0. \end{aligned} \quad (5)$$

Relaxing the rank-one constraint, we obtain the following semidefinite programming relaxation of (5):

$$\begin{aligned} \hat{x} &= \arg \max_x \langle Z, A \rangle \\ \text{subject to } & Z \succeq 0 \\ & Z_{ii} = 1, \quad i \in [n] \\ & \langle Z, \mathbf{J} \rangle = 0. \end{aligned} \quad (6)$$

Let $Z^* = x^* x^{*T}$ corresponding to the true labels for desired latent variables and define $\mathcal{Z}_n = \{x x^T : x \in \{\pm 1\}^n, x^T \mathbf{1} = 0\}$. For convenience, define

$$\begin{aligned} \eta_3 &\triangleq \frac{1}{2} \left(\sqrt{a\rho_{1y} + b\rho_{2y}} - \sqrt{c\rho_{1y} + d\rho_{2y}} \right)^2, \\ \eta_4 &\triangleq \frac{1}{2} \left(\sqrt{a\rho_{2y} + b\rho_{1y}} - \sqrt{c\rho_{2y} + d\rho_{1y}} \right)^2. \end{aligned}$$

Theorem 3. For two unknown binary latent variable model, if

$$\min \{\eta_3, \eta_4\} > 1$$

then as $n \rightarrow \infty$, for semidefinite programming estimator $\hat{Z}_{SDP}$, $\min_{Z^* \in \mathcal{Z}_n} \mathbb{P}(\hat{Z}_{SDP} = Z^*) \geq 1 - o(1)$.

The following Theorem establishes the optimality of semidefinite programming procedure.

Theorem 4. For two unknown binary latent variable model, if

$$\min \{\eta_3, \eta_4\} < 1$$

then for any sequence of estimators $\hat{Z}_n$, $\mathbb{P}(\hat{Z}_n = Z^*) \rightarrow 0$ as $n \rightarrow \infty$.

IV. PROOF OF THEOREM 1

The following lemma shows the sufficient conditions for the optimality of semidefinite programming (3).

Lemma 1. Suppose there exist $D^* = \text{diag}(d_i^*) \geq 0$ and $\lambda^* \in \mathbb{R}$, such that $S^* \triangleq D^* + \lambda^* \mathbf{J} - T_1 B - T_2 A$ satisfies $S^* \succeq 0$, $\lambda_2(S^*) > 0$, and $S^* x^* = \mathbf{0}$. Then $\hat{Z}_{SDP} = Z^*$ is the unique solution to (3).

Proof. Let $D = \text{diag}(d_i)$, $\lambda \in \mathbb{R}$, and $S \succeq 0$ denote the Lagrangian multipliers of (3). For any Z that satisfies the constraints in (3), we have

$$\begin{aligned} T_1 \langle B, Z \rangle + T_2 \langle A, Z \rangle &\stackrel{(a)}{\leq} L(Z, S^*, D^*, \lambda^*) = \langle D^*, \mathbf{I} \rangle \\ &\stackrel{(b)}{=} \langle S^* - \lambda^* \mathbf{J} + T_1 B + T_2 A, Z^* \rangle \\ &\stackrel{(c)}{=} T_1 \langle B, Z^* \rangle + T_2 \langle A, Z^* \rangle, \end{aligned}$$

where (a) holds because $\langle S^*, Z \rangle \geq 0$, (b) holds because $Z_{ii} = 1$ for all $i \in [n]$ and $S^* = D^* + \lambda^* \mathbf{J} - T_1 B - T_2 A$, and (c) holds because $S^* x^* = \mathbf{0}$ and $x^{*T} \mathbf{1} = 0$. Therefore, $Z^* = x^* x^{*T}$ is an optimal solution of (3). Now, assume $\tilde{Z}$ is another optimal solution. Then,

$$\begin{aligned} \langle S^*, \tilde{Z} \rangle &= \langle D^* + \lambda^* \mathbf{J} - T_1 B - T_2 A, \tilde{Z} \rangle \\ &\stackrel{(a)}{=} \langle D^* + \lambda^* \mathbf{J} - T_1 B - T_2 A, Z^* \rangle = \langle S^*, Z^* \rangle = 0, \end{aligned}$$

where (a) holds because $\langle T_1 B + T_2 A, Z^* \rangle = \langle T_1 B + T_2 A, \tilde{Z} \rangle$, $Z_{ii}^* = \tilde{Z}_{ii} = 1$ for all $i \in [n]$, and $\langle \mathbf{J}, Z^* \rangle = \langle \mathbf{J}, \tilde{Z} \rangle = 0$. Since $\tilde{Z} \succeq 0$ and $S^* \succeq 0$ while its second smallest eigenvalue $\lambda_2(S^*)$ is positive (since $S^* x^* = \mathbf{0}$), $\tilde{Z}$ must be a multiple of Z^* . Also, since $\tilde{Z}_{ii} = Z_{ii}^* = 1$ for all $i \in [n]$, we have $\tilde{Z} = Z^*$. $\square$

We now show that $S^* = D^* + \lambda^* \mathbf{J} - T_1 B - T_2 A$ satisfies other conditions in Lemma 1 with probability $1 - o(1)$. Let

$$d_i^* = T_1 \sum_{j=1}^n B_{ij} x_j^* x_i^* + T_2 \sum_{j=1}^n A_{ij} x_j^* x_i^*. \quad (7)$$

Then $D^* x^* = T_1 B x^* + T_2 A x^*$ and based on the definition of S^* in Lemma 1, S^* satisfies the condition $S^* x^* = \mathbf{0}$. It remains to show that $S^* \succeq 0$ and $\lambda_2(S^*) > 0$ with probability $1 - o(1)$. In other words, we need to show that

$$\mathbb{P} \left\{ \inf_{v \perp x^*, \|v\|=1} v^T S^* v > 0 \right\} \geq 1 - o(1),$$

where v is an $n \times 1$ vector. Then for any v such that $v^T x^* = 0$ and $\|v\| = 1$,

$$\begin{aligned} v^T S^* v &= v^T D^* v + \lambda^* v^T \mathbf{J} v - T_1 v^T (B - \mathbb{E}[B]) v \\ &\quad - T_2 v^T (A - \mathbb{E}[A]) v - T_1 v^T \mathbb{E}[B] v \\ &\quad - T_2 v^T \mathbb{E}[A] v \\ &\geq \min_i d_i^* + \lambda^* v^T \mathbf{J} v - T_1 \|B - \mathbb{E}[B]\| \\ &\quad - T_2 \|A - \mathbb{E}[A]\| - T_1 v^T \mathbb{E}[B] v \\ &\quad - T_2 v^T \mathbb{E}[A] v. \end{aligned}$$

It can be shown that

$$\begin{aligned} & -T_1 v^T \mathbb{E}[B] v - T_2 v^T \mathbb{E}[A] v \\ &= -\frac{1}{4} [T_1 c_1 + T_2 c_2] v^T W v - \frac{1}{4} [T_1 c_3 + T_2 c_4] v^T (Z * W) v \\ & -\frac{1}{4} [T_1 c_1 + T_2 c_2] v^T \mathbf{J} v + (T_1 + T_2) q_{xy}, \end{aligned}$$

where $c_1 = (a - b + c - d) \frac{\log n}{n}$, $c_2 = (a + b + c + d) \frac{\log n}{n}$, $c_3 = (a - b - c + d) \frac{\log n}{n}$, and $c_4 = (a + b - c - d) \frac{\log n}{n}$.

Lemma 2. For any $c > 0$, there exists $c', c'' > 0$ such that for any $n \geq 1$, $\|A - \mathbb{E}[A]\| \leq c' \sqrt{\log n}$ and $\|B - \mathbb{E}[B]\| \leq c' \sqrt{\log n}$ with probability at least $1 - n^{-c}$.

Proof. The proof is similar to the proofs [7, Theorem 9] and [8, Theorem 5]. $\square$

Lemma 3. With probability at least $1 - n^{-\frac{1}{2}}$, $v^T W v \leq \sqrt{\log n} + (\rho_{1y} - \rho_{2y})^2 v^T \mathbf{J} v + 2|\rho_{1y} - \rho_{2y}| \sqrt{n \log n}$ and $v^T (Z * W) v \leq \sqrt{\log n}$.

Proof. Since $-|v_i| \leq v_i y_i \leq |v_i|$, by applying the Chernoff bound we have

$$\mathbb{P}(v^T y - \mathbb{E}[v^T y] \geq \sqrt{\log n}) \leq n^{-\frac{1}{2}}.$$

Since $\mathbb{E}[v^T y] = (\rho_{1y} - \rho_{2y}) v^T \mathbf{1}$ and $|v^T \mathbf{1}| \leq \|v\|_2 \|\mathbf{1}\|_2 = \sqrt{n}$, with probability converging to one,

$$\begin{aligned} (v^T y)^2 &\leq \log n + (\rho_{1y} - \rho_{2y})^2 v^T \mathbf{J} v \\ & + 2|v^T \mathbf{1}| |\rho_{1y} - \rho_{2y}| \sqrt{\log n} \\ &\leq \log n + (\rho_{1y} - \rho_{2y})^2 v^T \mathbf{J} v + 2|\rho_{1y} - \rho_{2y}| \sqrt{n \log n}. \end{aligned}$$

Similarly, it can be shown that with high probability $v^T (Z * W) v \leq \sqrt{\log n}$ by noticing that $\mathbb{E}[\sum_i x_i y_i v_i] = 0$ and $-|v_i| \leq x_i y_i v_i \leq |v_i|$. $\square$

Lemma 4. Let $\delta = \frac{\log n}{\log \log n}$. Then $\mathbb{P}(\min_i d_i^* \geq \delta) \geq 1 - n^{1-\eta_1+o(1)} - n^{1-\eta_2+o(1)}$.

Proof. The proof is achieved by applying the Chernoff bound and taking the union bound. $\square$

Notice that $\rho_{1y} \leq 0.5$ implies $\eta_1 \leq \eta_2$ and $\rho_{1y} \geq 0.5$ implies $\eta_1 \geq \eta_2$. When $\rho_{1y} \leq 0.5$, if $\eta_1 > 1$ then $\min_i d_i^* \geq \frac{\log n}{\log \log n}$ with probability converging to one. Also, when $\rho_{1y} > 0.5$, if $\eta_2 > 1$ then $\min_i d_i^* \geq \frac{\log n}{\log \log n}$ with high probability.

Let $\lambda^* \geq \frac{1}{4} [T_1 c_1 + T_2 c_2] (\rho_{1y} - \rho_{2y})^2$. Therefore, applying Lemmas 2, 3, and 4, we get that when $\rho_{1y} \leq 0.5$, if $\eta_1 > 1$ and when $\rho_{1y} > 0.5$, if $\eta_2 > 1$, then

$$\begin{aligned} v^T S^* v &\geq \frac{\log n}{\log \log n} - (T_1 c' + T_2 c'') \sqrt{\log n} + (T_1 + T_2) q_{xy} \\ &> 0, \end{aligned}$$

and Theorem 1 follows.

V. PROOF OF THEOREM 2

Since x^* has a uniform distribution over $\{x \in \{\pm 1\}^n : x^T \mathbf{1} = 0\}$, maximum likelihood estimator minimizes the error probability among all estimators. Then we need to find when the maximum likelihood estimator fails. Let $e(i, H) \triangleq \sum_{j \in H} A_{ij} (T_1 y_i y_j + T_2)$. Also, let F_1 and F_2 denote the events that $\min_{i \in C_1^*} (e(i, C_1^*) - e(i, C_2^*)) \leq -2$ and $\min_{i \in C_2^*} (e(i, C_2^*) - e(i, C_1^*)) \leq -2$, respectively, where $C_1^* = \{v : x_v = 1, v \in [n]\}$ and $C_2^* = \{v : x_v = -1, v \in [n]\}$. Then $\mathbb{P}(\text{ML fails}) \geq \mathbb{P}(F_1 \cap F_2)$. Thus it suffices to show that with high probability $\mathbb{P}(F_1) \rightarrow 1$ and $\mathbb{P}(F_2) \rightarrow 1$. Here, we just prove that $\mathbb{P}(F_1) \rightarrow 1$, while $\mathbb{P}(F_2) \rightarrow 1$ is proved similarly. By symmetry, we can condition on C_1^* being the first $\frac{n}{2}$ vertices. Let T denote the set of first $\left\lfloor \frac{n}{\log^2 n} \right\rfloor$ vertices of C_1^* . Then

$$\begin{aligned} \min_{i \in C_1^*} (e(i, C_1^*) - e(i, C_2^*)) &\leq \min_{i \in T} (e(i, C_1^*) - e(i, C_2^*)) \\ &\leq \min_{i \in T} (e(i, C_1^* \setminus T) - e(i, C_2^*)) + \max_{i \in T} e(i, T). \end{aligned}$$

Let E_1 and E_2 denote the events $\max_{i \in T} e(i, T) \leq \delta - 2$, $\min_{i \in T} (e(i, C_1^* \setminus T) - e(i, C_2^*)) \leq -\delta$, respectively. To show that $\mathbb{P}(F_1) \rightarrow 1$, we need to show that $\mathbb{P}(E_1) \rightarrow 1$ and $\mathbb{P}(E_2) \rightarrow 1$.

For $i \in T$, $e(i, T) = (T_1 + T_2)X_1 + (T_1 - T_2)X_2$, where $X_1 \sim \text{Binom}(|T|, a \log n/n)$ and $X_2 \sim \text{Binom}(|T|, b \log n/n)$. Applying the Chernoff bound for binomial distributions, we get

$$\begin{aligned} \mathbb{P}\left(X_1 \geq \frac{\delta - 2}{2(T_1 + T_2)}\right) &\leq \left(\frac{\log n}{2(T_1 + T_2)ea} \left(\frac{\delta}{2} - 1\right)\right)^{\frac{2-\delta}{2(T_1 + T_2)}} \leq n^{-2+o(1)}, \\ \mathbb{P}\left(X_2 \geq \frac{\delta - 2}{2|T_1 - T_2|}\right) &\leq \left(\frac{\log n}{2|T_1 - T_2|eb} \left(\frac{\delta}{2} - 1\right)\right)^{\frac{2-\delta}{2|T_1 - T_2|}} \leq n^{-2+o(1)}. \end{aligned}$$

Since $T_1 - T_2 < 0$ and $T_1 + T_2 > 0$, we get that

$$\begin{aligned} \mathbb{P}(e(i, T) \geq \delta - 2) &\leq \mathbb{P}((T_1 + T_2)X_1 + |T_1 - T_2|X_2 \geq \delta - 2) \leq n^{-2+o(1)}. \end{aligned}$$

Using the union bound yields $\mathbb{P}(E_1) \geq 1 - n^{-1+o(1)}$. Therefore $\mathbb{P}(E_1) \rightarrow 1$ with high probability.

Lemma 5. [2, Lemma 15] Let $X_1, X_2, \dots, X_m$ be a sequence of i.i.d. random variables and $m = n + o(n)$. Then for any $\mu \in \mathbb{R}$ and $\nu \in \mathbb{R}$, $\mathbb{P}(\sum_{i=1}^m X_i \leq \mu - \nu) \geq \min_{t \in \mathbb{R}} e^{-t\mu - |t|\nu} M(t) \left(1 - \frac{\sigma_Y^2}{\nu^2}\right)$, where $M(t)$ is the moment generating function of $Y = \sum_{i=1}^m X_i$ and $\hat{Y}$ is a random variable distributed according to $\frac{e^{ty} \mathbb{P}(y)}{E_Y[e^{ty}]}$ with variance σ_Y^2 .

Lemma 6. Let E_2' denote the event that $e(i, C_1^* \setminus T) - e(i, C_2^*) \leq -\delta$, where $e(i, H) \triangleq \sum_{j \in H} A_{ij} (T_1 y_i y_j + T_2)$. Then $\mathbb{P}(E_2') \geq n^{-\eta_1+o(1)} + n^{-\eta_2+o(1)}$.

Proof. The proof is achieved by applying Lemma 5 and the Chernoff bound. $\square$

Applying Lemma 6 yields

$$\begin{aligned}\mathbb{P}(E_2) &= 1 - \prod_{i \in T} [1 - \mathbb{P}(E'_2)] \\ &\geq 1 - \left[1 - n^{-\eta_2 + o(1)} - n^{-\eta_1 + o(1)}\right]^{|T|} \\ &\geq 1 - e^{-n^{1-\eta_2+o(1)} - n^{1-\eta_1+o(1)}}.\end{aligned}$$

Notice that $\rho_{1y} \leq 0.5$ implies $\eta_1 \leq \eta_2$ and $\rho_{1y} \geq 0.5$ implies $\eta_1 \geq \eta_2$. When $\rho_{1y} \leq 0.5$, if $\eta_1 < 1$ then $\mathbb{P}(E_2) \rightarrow 1$ and the first part of Theorem 2 follows. When $\rho_{1y} \geq 0.5$, if $\eta_2 < 1$ then $\mathbb{P}(E_2) \rightarrow 1$ and the second part of Theorem 2 follows.

VI. PROOF OF THEOREM 3

The following lemma states the sufficient conditions for the optimality of semidefinite programming (6).

Lemma 7. Suppose there exist $D^* = \text{diag}(d_i^*) \geq 0$ and $\lambda^* \in \mathbb{R}$, such that $S^* \triangleq D^* + \lambda^* \mathbf{J} - A$ satisfies $S^* \succeq 0$, $\lambda_2(S^*) > 0$, and $S^* x = 0$. Then $\hat{Z}_{SDP} = Z^*$ is the unique solution to (6).

Proof. The proof is similar to the proof of Lemma 1. $\square$

It suffices to show that $S^* = D^* + \lambda^* \mathbf{J} - A$ satisfies other conditions in Lemma 7 with probability $1 - o(1)$. Let

$$d_i^* = \sum_{j=1}^n A_{ij} x_j^* x_i^*.$$

Then $D^* x^* = A x^*$ and based on the definition of S^* in Lemma 7, S^* satisfies the condition $S^* x^* = 0$. It remains to show that $S^* \succeq 0$ and $\lambda_2(S^*) > 0$ with probability at least $1 - o(1)$, i.e.,

$$\mathbb{P}\left\{\inf_{v \perp x^*, \|v\|=1} v^T S^* v > 0\right\} \geq 1 - o(1),$$

where v is an $n \times 1$ vector. Then for any v such that $v^T x^* = 0$ and $\|v\| = 1$,

$$\begin{aligned}v^T S^* v &= v^T D^* v + \lambda^* v^T \mathbf{J} v - v^T (A - \mathbb{E}[A]) v - v^T \mathbb{E}[A] v \\ &\geq \min_i d_i^* + \lambda^* v^T \mathbf{J} v - \|A - \mathbb{E}[A]\| - v^T \mathbb{E}[A] v.\end{aligned}$$

It can be shown that

$$\begin{aligned}-v^T \mathbb{E}[A] v &= -\frac{1}{4} [c_1 v^T W v - c_2 v^T \mathbf{J} v - c_3 v^T (Z * W) v] \\ &\quad + q_{xy},\end{aligned}$$

where $c_1 = (a - b + c - d) \frac{\log n}{n}$, $c_2 = (a + b + c + d) \frac{\log n}{n}$, and $c_3 = (a - b - c + d) \frac{\log n}{n}$.

Lemma 8. Let $\delta = \frac{\log n}{\log \log n}$. Then $\mathbb{P}(\min_i d_i^* \geq \delta) \geq 1 - n^{1-\eta_3+o(1)} - n^{1-\eta_4+o(1)}$.

Proof. The proof is achieved by applying the Chernoff bound and taking the union bound. $\square$

If $\min\{\eta_3, \eta_4\} > 1$ then $\min_i d_i^* \geq \frac{\log n}{\log \log n}$ with probability converging to one. Let $\lambda^* \geq \frac{1}{4} [c_1 (\rho_{1y} - \rho_{2y})^2 + c_2]$. Applying Lemmas 2, 3, and 8, we get that when $\min\{\eta_3, \eta_4\} > 1$, then

$$v^T S^* v \geq \frac{\log n}{\log \log n} - c' \sqrt{\log n} + q_{xy} > 0,$$

and Theorem 3 follows.

VII. PROOF OF THEOREM 4

The proof is similar to the proof of Theorem 2. Here, we need to find when the maximum likelihood estimator fails. The events F_1, F_2, E_1, E_2 , and E'_2 are the same as we have defined them in the proof of Theorem 2. Also, the definitions for C_1^*, C_2^* , and T remain valid for this part. Then $\mathbb{P}(\text{ML fails}) \geq \mathbb{P}(F_1 \cap F_2)$. Here we just prove that $\mathbb{P}(F_1) \rightarrow 1$, while $\mathbb{P}(F_2) \rightarrow 1$ is proved similarly.

By symmetry, we can condition on C_1^* being the first $\frac{n}{2}$ vertices. Then,

$$\begin{aligned}\min_{i \in C_1^*} (e(i, C_1^*) - e(i, C_2^*)) &\leq \min_{i \in T} (e(i, C_1^*) - e(i, C_2^*)) \\ &\leq \min_{i \in T} (e(i, C_1^* \setminus T) - e(i, C_2^*)) + \max_{i \in T} e(i, T),\end{aligned}$$

where $e(i, H) \triangleq \sum_{j \in H} A_{ij}$.

For $i \in T$, $e(i, T) = X_1 + X_2$, where $X_1 \sim \text{Binom}(|T|, a \log n/n)$ and $X_2 \sim \text{Binom}(|T|, b \log n/n)$. Applying the Chernoff bound for binomial distributions,

$$\begin{aligned}\mathbb{P}\left(X_1 \geq \frac{\delta}{2} - 1\right) &\leq \left(\frac{\log n}{2ea} \left(\frac{\delta}{2} - 1\right)\right)^{1-\frac{\delta}{2}} \leq n^{-2+o(1)}, \\ \mathbb{P}\left(X_2 \geq \frac{\delta}{2} - 1\right) &\leq \left(\frac{\log n}{2eb} \left(\frac{\delta}{2} - 1\right)\right)^{1-\frac{\delta}{2}} \leq n^{-2+o(1)}.\end{aligned}$$

Then $\mathbb{P}(e(i, T) \geq \delta - 2) \leq n^{-2+o(1)}$. Using the union bound, $\mathbb{P}(E_1) \geq 1 - n^{-1+o(1)}$. Therefore $\mathbb{P}(E_1) \rightarrow 1$ with high probability.

Lemma 9. Let E'_2 denote the event that $e(i, C_1^* \setminus T) - e(i, C_2^*) \leq -\delta$, where $e(i, H) \triangleq \sum_{j \in H} A_{ij}$. Then $\mathbb{P}(E'_2) \geq n^{-\eta_3+o(1)} + n^{-\eta_4+o(1)}$.

Proof. The proof is achieved by applying Lemma 5 and the Chernoff bound. $\square$

Applying Lemma 9 yields

$$\begin{aligned}\mathbb{P}(E_2) &= 1 - \prod_{i \in T} [1 - \mathbb{P}(E'_2)] \\ &\geq 1 - \left[1 - n^{-\eta_3+o(1)} - n^{-\eta_4+o(1)}\right]^{|T|} \\ &\geq 1 - e^{-n^{1-\eta_3+o(1)} - n^{1-\eta_4+o(1)}}.\end{aligned}$$

Therefore, if $\min\{\eta_3, \eta_4\} < 1$ then $\mathbb{P}(E_2) \rightarrow 1$ and Theorem 4 follows.

REFERENCES

- [1] E. Abbe, "Community detection and stochastic block models: recent developments," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.
- [2] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.
- [3] —, "Exact recovery in community detection with continuous-valued side information," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 332–336, 2018.
- [4] —, "Side information in recovering a single community: Information theoretic limits," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2107–2111.
- [5] M. Esmaili, H. Saad, and A. Nosratinia, "Exact recovery by semidefinite programming in the binary stochastic block model with partially revealed side information," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3477–3481.
- [6] —, "Community detection with side information via semidefinite programming," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 420–424.
- [7] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming: Extensions," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5918–5937, Oct 2016.
- [8] —, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, May 2016.