Exact Recovery in the Balanced Stochastic Block Model with Side Information

Jin Sima[1], **Feng Zhao**[2], and **Shao-Lun Huang**[3]

[1]Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China 100084
[3]DSIT Research Center, Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China 518055

***Abstract*—The role that side information plays in improving the exact recovery threshold in the stochastic block model (SBM) has been studied in many aspects. This paper studies exact recovery in n node balanced binary symmetric SBM with side information, given in the form of $O(\log n)$ i.i.d. samples at each node. A sharp exact recovery threshold is obtained and turns out to coincide with an existing threshold result, where no balanced constraint is imposed. Our main contribution is an efficient semi-definite programming (SDP) algorithm that achieves the optimal exact recovery threshold. Compared to the existing works on SDP algorithm for SBM with constant number of samples as side information, the challenge in this paper is to deal with the number of samples increasing in n.**

I. Introduction

The stochastic block model (SBM) [1], also known as the planted partition model, is a statistical model that admits neat theoretical analysis and efficient algorithms while capturing some key features exhibited in large data networks, such as social, biological, and computer networks [2]. The SBM describes a graph of n nodes, partitioned into multiple communities. Each edge in the graph exists independently with a probability determined by the communities the two nodes on the edge belong to. The goal is to recover the community each node belongs to, based on one instance of the graph edges. While there are several levels of recovery defined and studied (see [3] for a comprehensive survey), in this paper we focus on the exact recovery, which aims to recover all the communities.

For exact recovery in the SBM, a more interesting regime is when the edge connecting probabilities are in the order of $O(\frac{\log n}{n})$, in which circumstance there are phase transition phenomena for the exact recovery problem. The tight threshold on the exact recovery in this setting was not established until the work of [2], [4], where efficient and optimal algorithms based on semi-definite programming (SDP) were also provided. The results in [2], [4] were generalized in [5] where the sharp threshold for multiple communities with asymmetric edge connecting probabilities is derived, and in [6], where optimal SDP algorithms for multiple communities and two communities with unequal community sizes are given.

While the SBM focuses on graphical data only, it is natural to study the benefit of additional local data at the nodes, referred to as side information, to the exact recovery in the SBM. Such setting arises in applications where multi-modal data are observed. For example, in a social network, not only the interactions among people, but also the profile of each individual can be collected. It has been shown in many studies that side information such as node attributes or features [7], [8], [9] can assist the community detection tasks. For exact recovery in the SBM with side information, the work of [10] generalized [5] and derived a sharp threshold for exact recovery in the SBM with side information, given in the form of $\log n$ data samples at each node, drawn identically and independently according to a probability distribution determined by the community the node belongs to. The community belongings are described by node labels, which are i.i.d. according to a probability distribution. The work of [11], [12], [13], [14], [15] considered variations where side information is given as partial revealed labels, noisy labels, or latent variables. SDP algorithms were presented to achieve the sharp threshold in [11], [14], [15].

In this paper, we consider balanced binary symmetric SBM with side information in the form of $O(\log n)$ i.i.d. node samples, drawn according to a distribution determined by the community the node belongs to. The balanced property in this paper means that the two communities in the graph have equal sizes. The setting in this paper can be regarded as a special case of that in [10], with the difference that each node belongs to any one of the two communities with probability $\frac{1}{2}$ in [3], while a balanced constraint is imposed in this paper.

The contributions of this paper are as follows. First, we derived a sharp threshold for the balanced SBM with side information. It turns out that the threshold in this paper coincides with that of [10] for the special case of SBM with side information and without balanced constraint. Our major contribution is an SDP based algorithm that achieves the threshold with high probability. Different from the SDP algorithms proposed in [11], [14], [15], where the side information consists of a constant number of samples, our SDP algorithm deals with cases where the number of samples is of order $O(\log n)$. Our SDP algorithm is a nontrivial generalization of the SDP algorithm in [2] where a set of row and column constraints are imposed, to deal with the scaling of the number of samples. This poses challenges in analyzing the optimality of the proposed SDP algorithm, since the dual problem of the SDP relaxation becomes more complex.

The paper is organized as follows. In Section II, we introduce the model and present some definitions and lemmas

needed throughout the paper. Section III presents a sharp bound on exact recovery. In Section IV, we provide an SDP algorithm that achives the optimal threshold. Section V provides simulation results for the SDP algorithm. Section VI concludes the paper.

II. Preliminaries

A balanced binary symmetric SBM is defined by a random graph with n nodes $\{1,\dots,n\}$ and edges $Z=\{Z_{i,j}\}_{1\le i<j\le n}$, where $Z_{i,j}=1$ if nodes i and j are connected with an edge and $Z_{i,j}=0$ otherwise. Each node $i\in\{1,\dots,n\}$ is associated with a label $Y_i\in\{\pm 1\}$ such that the label $Y=(Y_1,\dots,Y_n)$ is uniformly distributed over the space $\{Y:\sum_{i=1}^n Y_i=0\}$. The edges $\{Z_{i,j}\}_{1\le i<j\le n}$ are independently distributed Bernoulli random variables, where $Z_{i,j}=1$ with probability $p=a\frac{\log n}{n}$ for nodes i,j with the same labels, i.e., $Y_i=Y_j$, and $Z_{i,j}=1$ with probability $q=b\frac{\log n}{n}$ if $Y_i\neq Y_j$. In this paper, it is assumed that $a>b$.

A balanced binary symmetric SBM with side information (SBMSI) is a generalization of the balanced SBM. In addition to the graph Z and the labels Y, each node i has $m=\gamma\log n$ data samples X_j^i, $i\in\{1,\dots,n\}$, $j\in\{1,\dots,m\}$, that are drawn identically and independently from distribution P_0 if $Y_i=1$ and from distribution P_1 if $Y_i=-1$. Note that the data samples X_j^i, $j\in\{1,\dots,m\}$ are independent from $\{Z_{i,j}\}_{1\le i<j\le n}$ given the label Y_i for any $i\in\{1,\dots,n\}$. Hence, the joint probability distribution of $(\{Z_{i,j}\}_{1\le i<j\le n},\{X_j^i\}_{1\le i\le n,1\le j\le m})$ conditioned on Y is

$$\begin{aligned}&P(x=\{x_j^i\}_{1\le i\le n,1\le j\le m},z=\{z_{i,j}\}_{1\le i<j\le n}|(y_1,\dots,y_n))\\&=\prod_{1\le i,j\le n}P(z_{i,j}|y_i,y_j)\prod_{i=1}^n\prod_{j=1}^m P(x_j^i|y_i),\end{aligned}\tag{1}$$

where

$$P(z_{i,j}=1|y_i,y_j)=\begin{cases}p & \text{if } y_i=y_j\\ q & \text{if } y_i\neq y_j\end{cases},$$

and

$$P(x_j^i|y_i)=\begin{cases}P_0(x_j^i) & y_i=1\\ P_1(x_j^i) & y_i=-1\end{cases}$$

The conditional probability distribution $P(\{x_j^i\}_{1\le i\le n,1\le j\le m},\{z_{i,j}\}_{1\le i<j\le n}|y_1,\dots,y_n)$ is determined by parameters n, p, q, P_0, and P_1. Hence, the SBMSI is denoted as $\mathrm{SBMSI}(n,m,p,q,P_0,P_1)$. In $\mathrm{SBMSI}(n,m,p,q,P_0,P_1)$, the goal is to recover the unknown labels Y, given the graph Z and the data samples X. In this paper, we consider exact recovery of Y, which is defined as follows.

Definition 1 (Exact Recovery for $\mathrm{SBMSI}(n,m,p,q,P_0,P_1)$)**.** *Let $(Z=\{Z_{i,j}\}_{1\le i<j\le n},Y,X=\{X_j^i\}_{1\le i\le n,1\le j\le m})$ be a graph Z, node labels Y, and node data samples X be drawn from the distribution defined by* $\mathrm{SBMSI}(n,m,p,q,P_0,P_1)$*. Exact recovery is solvable if there exists an algorithm that takes (Z,X) as inputs and outputs $\hat{Y}=\hat{Y}(Z,X)$ such that the error probability $P_e:=P(\hat{Y}\neq Y)$ goes to 0 as n increases.*

The following definition will be used throughout this paper and is a special case of the definition in [10]. Define I_+ to be the Chernoff information between $\mathrm{Pois}(\frac{a}{2},\frac{b}{2})\times P_0$ and $\mathrm{Pois}(\frac{b}{2},\frac{a}{2})\times P_1$ where $\mathrm{Pois}(\cdot,\cdot)$ represents the bivariant Poisson distribution.

By computing the Karush–Kuhn–Tucker (KKT) conditions, we have the following lemma.

Lemma 1. *For an* $\mathrm{SBMSI}(n,m,p,q,P_0,P_1)$*, let*

$$I_1=\min_{P_{\widetilde{X}_1}}\gamma D(p_{\widetilde{X}_1}||P_0)+\frac{1}{2}g(a,b,2\epsilon),\ \textit{where}$$
$$\epsilon=\gamma\frac{D(P_{\widetilde{X}_1}||P_1)-D(P_{\widetilde{X}_1}||P_0)}{\log a/b}\tag{2}$$
$$g(a,b,\epsilon)\triangleq a+b-\sqrt{\epsilon^2+4ab}+\epsilon\log\frac{\epsilon+\sqrt{\epsilon^2+4ab}}{2b}\tag{3}$$

and

$$I_2=\min_{P_{\widetilde{X}_2}}\gamma D(p_{\widetilde{X}_2}||P_1)+\frac{1}{2}g(a,b,2\epsilon),\ \textit{where}$$
$$\epsilon=\gamma\frac{D(p_{\widetilde{X}_2}||P_0)-D(P_{\widetilde{X}_2}||P_1)}{\log a/b}.\tag{4}$$

Then, we have that $I_1=I_2=I_+$.

Proof. Using the results from Section 3 of [10], we can write I_+ explicitly as follows

$$\begin{aligned}I_+=&\frac{\lambda}{2}(a^{1-\lambda}b^\lambda-b^{1-\lambda}a^\lambda)\log\frac{b}{a}+\frac{a+b}{2}\\&-\frac{1}{2}(a^{1-\lambda}b^\lambda+b^{1-\lambda}a^\lambda)+\gamma D_{\mathrm{KL}}(p_\lambda||p_0)\end{aligned}\tag{5}$$

where λ is chosen to minimize

$$a^{1-\lambda}b^\lambda+b^{1-\lambda}a^\lambda+2\gamma\log(\sum_{x\in\mathcal{X}}p_0^{1-\lambda}(x)p_1^\lambda(x))\tag{6}$$

and p_λ is defined as

$$p_\lambda=\frac{p_0^{1-\lambda}(x)p_1^\lambda(x)}{\sum_{x\in\mathcal{X}}p_0^{1-\lambda}(x)p_1^\lambda(x)}.\tag{7}$$

We show that $I_1=I_+$ as an example. The other part $I_2=I_+$ can be proved similarly. We use Lagrange multiplier to solve (2). Let

$$\begin{aligned}L(p_{\widetilde{X}_1},\epsilon,\lambda)=&\gamma D(p_{\widetilde{X}_1}||p_0)+\frac{1}{2}g(a,b,2\epsilon)\\&-\lambda(\epsilon\log\frac{a}{b}-\gamma D(p_{\widetilde{X}_1}||P_1)+\gamma D(p_{\widetilde{X}_1}||P_0))\end{aligned}$$

It is equivalent to minimize $(1-\lambda)D(p_{\widetilde{X}_1}||P_0)+\lambda D(p_{\widetilde{X}_1}||P_1)$, from which we get $p_{\widetilde{X}_1}(x)=p_\lambda(x)$. From $\frac{\partial L(p_{\widetilde{X}_1},\epsilon,\lambda)}{\partial\epsilon}=0$ and taking (7) into (2), we get

$$\lambda\log\frac{a}{b}=\log\frac{\epsilon+\sqrt{\epsilon^2+ab}}{b}$$
$$\epsilon\log\frac{a}{b}=\gamma\frac{\sum_{x\in\mathcal{X}}p_0^{1-\lambda}(x)p_1^\lambda(x)\log\frac{p_0(x)}{p_1(x)}}{\sum_{x\in\mathcal{X}}p_0^{1-\lambda}(x)p_1^\lambda(x)}$$

After cancelling ϵ from the above two equations, we can get a single equation for λ:

$$\frac{\log\frac{a}{b}}{2}(a^\lambda b^{1-\lambda}-a^{1-\lambda}b^\lambda)+\gamma\frac{\sum_{x\in\mathcal{X}}p_0^{1-\lambda}(x)p_1^\lambda(x)\log\frac{p_1(x)}{p_0(x)}}{\sum_{x\in\mathcal{X}}p_0^{1-\lambda}(x)p_1^\lambda(x)}=0$$

which is the derivative of (6). Then by simple computation we have $I_+ = I_1$. □

III. Sharp Threshold for Balanced SBMSI

In this section we present a sharp closed form threshold for exact recovery in the balanced SBMSI. The threshold coincides with the result in [10] for SBMSI without balanced constraints.

Theorem 1. *For a balanced* SBMSI$(n, m, p = a\frac{\log n}{n}, q = b\frac{\log n}{n}, P_0, P_1)$*, exact recovery is solvable if*

$$I_+ > 1, \tag{8}$$

and is not solvable if $I_+ < 1$, in which case, the error probability P_e goes to 1.

Remark 1. *From Lemma 1 it can be seen that when γ increases, in which case the number of samples becomes larger, I_+ increases. Hence, more samples facilitate the recovery of communities.*

A. Proof of Theorem 1

The achievability part can be proved by our SDP algorithm, the details of which will be given in Section IV. We show that exact recovery is not possible if $I_+ < 1$.

Let S_1 and S_2 be the set of nodes with labels 1 and -1, respectively. For a node i and a node set $S \subset \{1,\dots,n\}$,let $E(i,S)$ be the number of edges between i and nodes in S. By similar arguments to the ones in [2], it can be shown that with high probability, there exists a subset $H_1 \subset S_1$ of size $\frac{n}{\log^3 n}$ and a node $i_1 \in H_1$, such that

$$\sum_{j=1}^m \log\frac{P_1(x_j^{i_1})}{P_0(x_j^{i_1})} \geq \log\frac{p(1-q)}{q(1-p)}\Big(\frac{\log n}{\log\log n}+1 + E(i_1,S_1\backslash H_1)-E(i_1,S_2)\Big), \tag{9}$$

if $I_1 < 1$, where I_1 is defined in (2). Similarly, if $I_2 < 1$, where I_2 is defined in (4), there exists a subset $H_2 \subset S_2$ of size $\frac{n}{\log^3 n}$ and a node $i_2 \in H_2$, such that

$$\sum_{j=1}^m \log\frac{P_0(x_j^{i_2})}{P_1(x_j^{i_2})} \geq \log\frac{p(1-q)}{q(1-p)}\Big(\frac{\log n}{\log\log n}+1 + E(i_2,S_2\backslash H_2)-E(i_2,S_1)\Big), \tag{10}$$

with high probability. Summing up (9) and (10), we obtain

$$\sum_{j=1}^m \log\frac{P_1(x_j^{i_1})}{P_0(x_j^{i_1})}+\sum_{j=1}^m \log\frac{P_0(x_j^{i_2})}{P_1(x_j^{i_2})} \geq \log\frac{p(1-q)}{q(1-p)}\Big(2\frac{\log n}{\log\log n}+2+E(i_1,S_1\backslash H_1) + E(i_2,S_2\backslash H_2)-E(i_1,S_2)-E(i_2,S_1)\Big). \tag{11}$$

Note that if (11) holds, then

$$\sum_{j=1}^m \log\frac{P_1(x_j^{i_1})}{P_0(x_j^{i_1})}+\sum_{j=1}^m \log\frac{P_0(x_j^{i_2})}{P_1(x_j^{i_2})} \geq \log\frac{p(1-q)}{q(1-p)}(E(i_1,S_1\backslash\{i_1\})+E(i_2,S_2\backslash\{i_2\}) - E(i_1,S_2\backslash\{i_2\})-E(i_2,S_1\backslash\{i_1\})). \tag{12}$$

Hence, with high probability, (12) holds if $I_1 < 1$ and $I_2 < 1$. According to Lemma 1, this occurs when $I_+ < 1$. According to (1) and the fact that Y is uniformly distributed, Eq. (12) is equivalent to the fact that the likelihood $P(X,Z|(Y_1,\dots,Y_n))$ (defined in (1)) with $Y_{i_1} = -1, Y_{i_2} = 1$ is larger than the likelihood $P(X,Z|(Y_1,\dots,Y_n))$ with $Y_{i_1} = 1, Y_{i_2} = -1$, which results in recovery failure. Therefore, we conclude that a recovery failure occurs with high probability, if $I_+ < 1$.

IV. SDP Relaxation

In this section, we propose a semi-definite programming based relaxation to find the maximum likelihood estimate of the labels Y. Note that the ML estimate of the labels Y is NP-hard while SDP formulation can be implemented efficiently using interior point method or alternating direction method [16]. Our SDP algorithm finds the true label Y^* with probability approaching 1, if $\gamma D_{1/2}(p_0||p_1)+(\sqrt{a}-\sqrt{b})^2 > 2$. The SDP problem can be solved using various efficient iterative schemes.

Let $S_1(Y)$ and $S_2(Y)$ be the sets of nodes with label 1 and -1, respectively, given Y. According to (1), the log likelihood of Y is given by

$$\begin{aligned}&\sum_{j=1}^m[\sum_{i\in S_1(Y)}\log P_1(X_j^i)+\sum_{i\in S_2(Y)}\log P_0(X_j^i)]\\ &+\sum_{i,j\in S_1(Y),i<j}[z_{i,j}\log p+(1-z_{i,j})\log(1-p)]\\ &+\sum_{i,j\in S_2(Y),i<j}[z_{i,j}\log p+(1-z_{i,j})\log(1-p)]\\ &+\sum_{i\in S_1(Y),j\in S_2(Y)}[z_{i,j}\log q+(1-z_{i,j})\log(1-q)].\end{aligned} \tag{13}$$

It can be verified that maximizing (13) over $Y \in \{\pm 1\}^n$ is equivalent to solving the following optimization problem

$$\begin{aligned}&\max_v h^T v+\frac{1}{4}v^T Bv\\ &s.t.\ \mathbb{1}_n^T v = 0 \text{ and } v_i \in \{\pm 1\}\end{aligned} \tag{14}$$

where h is an n-dimensional vector with entry $h_i = \frac{1}{\log\frac{p(1-q)}{q(1-p)}}\sum_{j=1}^m \log\frac{P_0(x_j^i)}{P_1(x_j^i)}$ for $i \in \{1,\dots,n\}$ and the $n\times n$ matrix B is defined as

$$B_{ij} = \begin{cases}1, & \text{if } i \text{ is connected to } j,\\ -1, & \text{otherwise.}\end{cases} \tag{15}$$

for $i, j \in \{1,\dots,n\}$.

Let v^* be the optimal solution to (14) and $V^* = (1, v^*)(1, v^*)^T$, where $(1, v^*)$ is a $(n+1)$-dimensional vector obtained by concatenating 1 and v^*. Then, the optimal value of (14) equals $\frac{1}{2}Tr(\widetilde{B}V^*)$, where

$$\widetilde{B} = \begin{pmatrix} 0 & h^T \\ h & \frac{1}{2}B \end{pmatrix}. \tag{16}$$

We wish to show that V^* is the unique optimal solution to the following problem.

$$\begin{aligned} \max_{V} \; & Tr(\tilde{B}V) \\ s.t. \; & V_{ii} = 1, \\ & V \succeq 0, \\ & \sum_{j=2}^{n+1}(V_{ij} + V_{ji}) = 0, \; \forall i \in \{1, \ldots, n+1\}. \end{aligned} \tag{17}$$

Note that here we use $n+1$ constraints in the last two lines of (17) to describe the balanced property of the labels. This is important in deriving the optimality and uniqueness of V, which will be proved in the following theorem. The theorem shows that our SDP relaxation achieves exact recovery with high probability, as long as the recovery condition is met.

Theorem 2. *If $I_+ > 1$, then with high probability, the optimal solution V^* to* (17) *is unique and given by $(1, y^*)(1, y^*)^T$, where y^* is the true labeling of the nodes.*

Proof of Theorem 2. Consider the dual problem of (17)

$$\begin{aligned} \min_{a_1, \ldots, a_{n+1}} \; & \sum_{i=1}^{n+1} a_i \\ s.t. \; & \operatorname{diag}\{a_1, \ldots, a_{n+1}\} + \Xi - \widetilde{B} \succeq 0, \end{aligned} \tag{18}$$

where the $(n+1) \times (n+1)$ symmetric matrix Ξ is defined as

$$\Xi_{ij} = \begin{cases} \lambda_1, & i = 1 \text{ or } j = 1, \text{ and } i \neq j \\ \lambda_i + \lambda_j, & i, j \in \{2, \ldots, n+1\} \end{cases}, \tag{19}$$

Let $g = y^* = (1, \ldots, 1, -1, \ldots, -1)$ be the true labels of the nodes, i.e., the first and second half of the nodes are labeled by 1 and -1, respectively. As similarly mentioned in [2], the optimality and uniqueness of the solution $(1, g)(1, g)^T$ to (17) is guaranteed by the following conditions:

(a) $(1, g)(1, g)^T$ is a feasible solution to the primal problem (14).
(b) There exists a feasible solution $(a_1, \ldots, a_{n+1}, \lambda_1, \ldots, \lambda_{n+1})$ to (18), such that $\sum_{i=1}^{n+1} a_i = Tr((1, g)(1, g)^T \widetilde{B})$.
(c) $(\operatorname{diag}\{a_1, \ldots, a_{n+1}\} + \Xi - \widetilde{B})(1, g) = 0$.
(d) The second smallest eigenvalue of $\operatorname{diag}\{a_1, \ldots, a_{n+1}\} + \Xi - \widetilde{B}$ is greater than 0.

Condition (a) holds by definition. It suffices to choose $(a_1, \ldots, a_{n+1}, \lambda_1, \ldots, \lambda_{n+1})$ that satisfies conditions (b), (c) and (d). Let $\mu = \frac{1}{n}\mathbb{1}_n^T h$ and $\lambda = -\mu/n$. Specify $(a_1, \ldots, a_{n+1}, \lambda_1, \ldots, \lambda_{n+1})$ as follows

$$\lambda_1 = \mu + \lambda, \text{ and } \lambda_{i+1} = g_i \lambda + \lambda, \; i \in \{1, \ldots, n\}$$

Then, from condition (c) we have that

$$\begin{aligned} a_1 &= h^T g \\ a_{i+1} &= (h_i - \lambda) g_i + \frac{1}{2}\operatorname{diag}\{B g g^T\}_i, i = 1, \ldots, n \end{aligned} \tag{20}$$

It can be verified that $\sum_{i=1}^{n+1} a_i = 2h^T g + \frac{1}{2} g^T B g$, and hence condition (b) holds. Hence it suffices to prove that

$$\begin{aligned} &\operatorname{diag}(\{a_1, \ldots, a_{n+1}\}) + \Xi - \widetilde{B} \\ &= \begin{pmatrix} h^T g & -h^T + (\mu + \lambda)\mathbb{1}_n^T \\ -h + (\mu + \lambda)\mathbb{1}_n & \Xi_n \end{pmatrix} \succeq 0, \end{aligned} \tag{21}$$

where

$$\begin{aligned} \Xi_n = &\operatorname{diag}(h g^T + A g g^T - \lambda \mathbb{1}_n g^T) \\ &+ (\frac{1}{2} + 2\lambda) J_n - A + 2\lambda \Xi', \end{aligned} \tag{22}$$

$\mathbb{1}_n$ is the all 1's vector, J_n is the all 1's matrix, and I_n is the identity matrix. The matrix $A = (B + J_n - I_n)/2$ is the adjacency matrix of the graph. The matrix Ξ' is given by $\Xi'_{ij} = g_i + g_j$.

In the following we show that Ξ_n is positive definite. Then according to condition (c) and the Cauchy's Interlacing Theorem [17], condition (d) holds. Then, the proof is done.

We show that $x^T \Xi_n x > 0$ for any vector $x \in \mathbb{R}^n$ satisfying $\|x\| = 1$. Decompose x as $x = \frac{\beta}{\sqrt{n}} g + \sqrt{1 - \beta^2} g^\perp$, where $g^T g^\perp = 0, \beta \in [0, 1]$, and $\|g^\perp\| = 1$. Then, we have that

$$\begin{aligned} x^T \Xi_n x = &\frac{\beta^2}{n} g^T \Xi_n g + \frac{\beta}{\sqrt{n}}\sqrt{1 - \beta^2} g^T \Xi_n g^\perp \\ &+ (1 - \beta^2)(g^\perp)^T \Xi_n g^\perp. \end{aligned} \tag{23}$$

In the next we derive lower bounds to the three terms in the right hand side of (23). For the first term, we have that

$$g^T \Xi_n g = g^T (h - \lambda \mathbb{1}_n) = g^T h$$

Since $\mathbb{E}[g^T h]$ is a positive number of order $O(n \log n)$, by Sanov's theorem or Chernoff bound, the probability $P(g^T h < 0)$ decreases exponentially in n. Hence, with high probability, the first term $g^T \Xi_n g$ is positive.

For the second term, let $\tilde{h} = \Xi_n g = (n-1)\lambda \mathbb{1}_n + h$, then $g^T \Xi_n g^\perp = \tilde{h}^T g^\perp \geq -\|\tilde{h} - \frac{1}{n}(\tilde{h}^T g) g\|$, where the norm $\|\tilde{h} - \frac{1}{n}(\tilde{h}^T g) g\|$ satisfies

$$\|\tilde{h} - \frac{1}{n}(\tilde{h}^T g) g\|^2 = \|\tilde{h}\|^2 - \frac{1}{n}(\tilde{h}^T g)^2.$$

Let $\hat{g}_1 = \frac{1}{2}(g + \mathbb{1}_n)$ and $\hat{g}_2 = \frac{1}{2}(-g + \mathbb{1}_n)$. Using the fact that $\tilde{h}^T \mathbb{1}_n = \mu$, we have

$$\begin{aligned} &\frac{\|\tilde{h}\|^2}{n} - (\frac{1}{n}\tilde{h}^T g)^2 \\ &= \frac{\|\tilde{h}\|^2}{n} - 2\frac{(\tilde{h}^T \hat{g}_1)^2}{n^2} - 2\frac{(\tilde{h}^T \hat{g}_2)^2}{n^2} + \frac{(\tilde{h}^T \mathbb{1}_n)^2}{n^2} \\ &= 2\frac{\sum_{i<j, i,j \in S_1}(\tilde{h}_i - \tilde{h}_j)^2 + \sum_{i<j, i,j \in S_2}(\tilde{h}_i - \tilde{h}_j)^2}{n^2} + \frac{\mu^2}{n^2} \\ &= I_1 + I_2 + \frac{\mu^2}{n^2}. \end{aligned}$$

where $I_j = \frac{\sum_{i\in S_j} h_i^2}{n} - 2\frac{(h^T\hat{g}_j)^2}{n^2}$ for $j \in \{1,2\}$, and S_1 and S_2 denotes the sets of nodes with label 1 and -1, respectively. Denote $\mathbb{E}[h_i] = mD_j, \mathrm{Var}[h_i] = m\bar{D}_j, \mathbb{E}[h_i^2] = \mathrm{Var}[h_i] + \mathbb{E}^2[h_i] = m\bar{D}_j + m^2D_j^2, \mathrm{Var}[h_i^2] \le m^4\bar{C}_j$ for $i \in S_j$ and $j \in \{1,2\}$. Then $D_j, C_j, j \in \{1,2\}$ are constants. Using Chebyshev's inequality, we have that

$$P\Big(\Big|\frac{\sum_{i\in S_j} h_i^2}{n} - \frac{1}{2}(m\bar{D}_j + m^2D_j^2)\Big| \ge \log n\Big) \le \frac{m^4\bar{C}_j}{2n\log^2 n}$$

$$P\Big(\Big|\frac{h^T\hat{g}_j}{n} - \frac{m}{2}D_j\Big| \ge \log^{-1} n\Big) \le \frac{m\log^2 n\bar{D}_j}{2n}$$

for $j \in \{1,2\}$. Hence, with probability $1 - n^{-1-o(1)}$,

$$\begin{aligned} I_j \le& \frac{1}{2}m\bar{D}_j + \frac{1}{2}m^2D_j^2 + \log n - 2(\frac{1}{2}mD_j - \log^{-1} n)^2 \\ =& O(\log n) \end{aligned}$$

for $j \in \{1,2\}$. Therefore, we have that

$$\frac{1}{\sqrt{n}}g^T\Xi_n g^\perp \ge -\sqrt{\frac{\|\tilde{h}\|^2}{n} - (\frac{1}{n}\tilde{h}^Tg)^2} = O(\sqrt{\log n})$$

For the last term $(g^\perp)^T\Xi_n g^\perp$, by (22) and the fact that $(g^\perp)^Tg = 0$, we have that

$$\begin{aligned} &(g^\perp)^T\Xi_n g^\perp \\ =&(g^\perp)^T \operatorname{diag}(-\lambda\mathbb{1}_n g^T + hg^T + Agg^T)g^\perp + p \\ &+ \frac{1}{2}(1-p-q+2\lambda)(g^\perp)^TJ_ng^\perp - (g^\perp)^T(A - \mathbb{E}[A])g^\perp, \end{aligned}$$

where we use the fact that $\mathbb{E}[A] = \frac{p-q}{2}gg^T + \frac{p+q}{2}J_n - pI_n$. Note that p, q, and λ are $o(1)$ terms and that $(g^\perp)^TJ_ng^\perp \ge 0$. By Theorem 5.2 of [18], with probability $1 - n^{-r}$, $\lambda_{\max}(A - \mathbb{E}[A]) \le c\sqrt{\log n}$ for some positive constant r and c. Hence, we have that

$$(g^\perp)^T\Xi_n g^\perp \ge \min_i\{(-\lambda + h_i)g_i + \lambda + g_i(Ag)_i\} - c\sqrt{\log n} \tag{24}$$

We now show that with high probability,

$$(-\lambda + h_i)g_i + \lambda + g_i(Ag)_i - c\sqrt{\log n} \ge 0 \tag{25}$$

for $i \in \{1,\dots,n\}$. For $g_i = 1$, the term $g_i(Ag)_i$ can be written as $Z_1 - Z_2$, where $Z_1 \sim \mathrm{Binom}(\frac{n}{2} - 1, \frac{a\log n}{n})$ and $Z_2 \sim \mathrm{Binom}(\frac{n}{2}, \frac{b\log n}{n})$ follow binomial distributions. The term $h_i = m\frac{D(P_{\widetilde{X}_i}||P_1) - D(P_{\widetilde{X}_i}||P_0)}{\log a/b}$, where $P_{\widetilde{X}_i}$ is the empirical distribution of samples $\{X_j^i\}_{j=1}^m$ at node i. Then, we have the following lemma, which is a slight generalization of Lemma 8 in [2] and can be proved using Chernoff bound.

Lemma 2. *Suppose* $m > n, Z \sim \mathrm{Binom}(m, \frac{b\log n}{n}), X \sim \mathrm{Binom}(m, \frac{a\log n}{n})$. *For* $t \ge \frac{m}{n}(b-a)$, *we have*

$$\begin{aligned} &P(Z - X \ge t\log n) \\ &\le \exp(-\frac{m}{n}\log n \cdot (g(a,b,\frac{n}{m}t) + O(\frac{\log n}{n}))) \end{aligned} \tag{26}$$

where $g(a,b,\epsilon)$ *is defined in* (3).

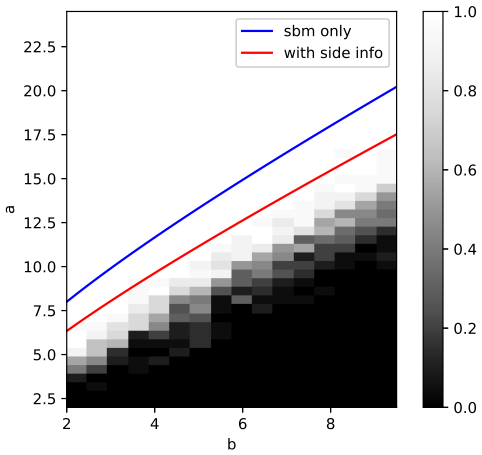

Fig. 1. Comparison of different thresholds and with empirical recovery result by SDP

According to Lemma 2 and Sanov's theorem, if $g_i = 1$, the probability that (25) does not hold is upper bounded by $n^{-I_1+o(1)}$ for any i, where I_1 is defined in (2). By the union bound, the probability that there exists an i that violates (25) is at most $n^{-I_1+1+o(1)}$, which fades to zero as $I_1 = I_+ > 1$. Similarly, the probability that an i satisfying $g_i = -1$ violates (25) decays if $I_2 = I_+ > 1$. Therefore, with high probability, $(g^\perp)^T\Xi_n g^\perp \ge 0$ when $I_+ > 1$, and Theorem 2 holds. □

V. Simulation Results

Fig. 1 shows the empirical probability of successful recovery of the SDP algorithm for the SBM with side information. We fix $n = 300$ and the number of trials to be 20. The number of samples is chosen as $m = 10$ for $\mathrm{Bern}(0.2)$ versus $\mathrm{Bern}(0.8)$. Then at each trial and for fixed a and b, we check how many times each method succeeds. Dividing by the number of trials, we obtain the empirical probability of success with respect to the exact recovery metric. The blue curve corresponds to the threshold $\sqrt{a} - \sqrt{b} = \sqrt{2}$ while the red curve corresponds to the threshold $I_+ > 1$. It can be seen from the figure that the recovery threshold for SBM with side information matches the red line $I_+ > 1$.

VI. Conclusion

In this paper, we obtained a sharp close-form exact recovery condition for a balanced two-community symmetric SBM with side information. Our result provides insight on the number of node samples to achieve exact recovery. We also proposed a semidefinite programming based algorithm that achieves the threshold with high probability. It will be interesting to see if the SDP algorithm in this paper can be extended to more general cases.

References

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[2] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[3] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018. [Online]. Available: http://jmlr.org/papers/v18/16-480.html

[4] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for the planted bisection model," *Electron. J. Probab.*, vol. 21, p. 24 pp., 2016. [Online]. Available: https://doi.org/10.1214/16-EJP4185

[5] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, 2015, pp. 670–688.

[6] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.

[7] Y. Zhang, E. Levina, J. Zhu *et al.*, "Community detection in networks with node features," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3153–3178, 2016.

[8] E. Mossel and J. Xu, "Local algorithms for block models with side information," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 2016, pp. 71–80.

[9] H. Saad and A. Nosratinia, "Recovering a single community with side information," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7939–7966, 2020.

[10] A. R. Asadi, E. Abbe, and S. Verdú, "Compressing data on graphs with clusters," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1583–1587.

[11] M. Esmaeili, 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.

[12] H. Saad, A. Abotabl, and A. Nosratinia, "Exact recovery in the binary stochastic block model with binary side information," in *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 822–829.

[13] 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.

[14] M. Esmaeili, H. Saad, and A. Nosratinia, "Community detection with side information via semidefinite programming," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 420–424.

[15] M. Esmaeili and A. Nosratinia, "Community detection with secondary latent variables," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1355–1360.

[16] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[17] S.-G. Hwang, "Cauchy's interlace theorem for eigenvalues of hermitian matrices," *The American Mathematical Monthly*, vol. 111, no. 2, pp. 157–159, 2002.

[18] J. Lei, A. Rinaldo *et al.*, "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, vol. 43, no. 1, pp. 215–237, 2015.