A Nearly-Linear Time Algorithm for Exact Community Recovery in Stochastic Block Model

Peng Wang [*1] **Zirui Zhou** [*2] **Anthony Man-Cho So** [1]

Abstract

Learning community structures in graphs that are randomly generated by stochastic block models (SBMs) has received much attention lately. In this paper, we focus on the problem of exactly recovering the communities in a binary symmetric SBM, where a graph of n vertices is partitioned into two equal-sized communities and the vertices are connected with probability $p = \alpha \log(n)/n$ within communities and $q = \beta \log(n)/n$ across communities for some $\alpha > \beta > 0$. We propose a two-stage iterative algorithm for solving this problem, which employs the power method with a random starting point in the first stage and turns to a generalized power method that can identify the communities in a finite number of iterations in the second stage. It is shown that for any fixed α and β such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, which is known to be the information-theoretic limit for exact recovery, the proposed algorithm exactly identifies the underlying communities in $\tilde{O}(n)$ time with probability tending to one as $n \to \infty$. We also present numerical results of the proposed algorithm to support and complement our theoretical development.

1. Introduction

Learning community structures in graphs is a fundamental task in machine learning and computer science. It has found broad applications in physics (Fortunato, 2010; Newman & Girvan, 2004), biology (Cline et al., 2007), and social science (Girvan & Newman, 2002), to name a few. In research on community estimation, the stochastic block model (SBM), which is a generative model for random graphs that admits community structures, has become increasingly popular as a platform for validating theoretical ideas and comparing numerical algorithms. In particular, substantial advances have been made in the past decade on understanding the statistical limits for detection and recovery of communities in graphs that are generated by SBMs, and on developing computationally tractable methods that can detect and recover the underlying communities at their corresponding statistical limits; see, for example, Mossel et al. (2014); Abbe & Sandon (2015); Abbe et al. (2016; 2017); Chen & Xu (2016); Gao et al. (2017).

In this paper, we focus on the problem of exactly recovering the communities in the binary symmetric SBM (also known as the planted bisection model). Specifically, given n nodes that are partitioned into two equal-sized clusters, a graph on these n nodes is randomly generated such that each pair of nodes are connected with probability p if they are in the same cluster and with probability q if not, where $p > q > 0$. Then, we aim to recover the underlying clusters using only the adjacency matrix of the generated graph. For this problem, the regime where $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ for some $\alpha > \beta > 0$ is of particular interest as it possesses a sharp threshold for exact recovery. Indeed, from an information-theoretic point of view, it has been proved that recovering the clusters with high probability is impossible if $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$ and is possible if $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$ (Abbe et al., 2016; Mossel et al., 2014). Moreover, a number of computationally tractable algorithms have been shown to achieve this threshold for exact recovery (Abbe et al., 2016; 2017; Hajek et al., 2016; Gao et al., 2017). More precisely, these algorithms have running time that is polynomial in n and for any fixed α and β such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, they successfully identify the underlying clusters with probability tending to one as $n \to \infty$. However, these algorithms rely on solving a semidefinite programming (SDP) or eigenvector computation, whose time complexities are polynomial but generally not linear in n. A major goal of this paper is to propose an algorithm that runs in time nearly linear in n and achieves exact recovery at the information-theoretic limit.

[*]Equal contribution [1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong [2]Huawei Technologies Canada Co., Ltd., Burnaby, Canada. Correspondence to: Zirui Zhou <zirui.zhou@huawei.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1.1. Related Works

For the binary symmetric SBM, whether it is possible to recover the underlying communities with high probability depends on the scaling of p, q, and $p-q$. When $p = a/n$ and $q = b/n$ for some $a > b > 0$ (also known as the sparse regime), it is impossible to recover the communities because the graph is unconnected with high probability (Decelle et al., 2011). In the regime where $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ for some $\alpha > \beta > 0$, Abbe et al. (2016) and Mossel et al. (2014) independently showed that there exists a sharp threshold for exact recovery. Specifically, they proved that from an information-theoretic point of view, exact recovery is impossible if $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$, while it is possible and can be achieved by the maximum likelihood (ML) estimator if $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$. Computing the ML estimator is, however, computationally intractable in general. Indeed, by encoding the underlying clusters into a vector $\boldsymbol{x}^* \in \{\pm 1\}^n$ such that $x_i^* = 1$ if node i is in the first cluster and $x_i^* = -1$ otherwise, the ML estimator of $\boldsymbol{x}^*$ is the solution of the following problem:

$$\max \left\{ \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} : \mathbf{1}_n^T \boldsymbol{x} = 0,\, x_i = \pm 1,\, i = 1, \dots, n \right\}, \quad (1)$$

where $\boldsymbol{A}$ is the adjacency matrix of the graph and $\mathbf{1}_n$ is the all-one vector of dimension n. Problem (1) is equivalent to the problem of finding minimum bisection of a graph, which is known to be NP-hard in the worst case (Garey et al., 1974). Nevertheless, the above threshold provides the information-theoretic limit for exact recovery in the binary symmetric SBM, which is useful for benchmarking recovery algorithms.

Over the past decades, many computationally tractable algorithms have been proposed that can be applied to the problem of exact community recovery in the binary symmetric SBM. Some of them are proposed as heuristic combinatorial algorithms for the minimum bisection problem with an average-case performance guarantee and some others tackle the community recovery problem based on convex or non-convex relaxations of the ML estimation problem (1). In the comparison of these algorithms, of particular interest are the following two factors: (i) computational efficiency and (ii) conditions on p and q that guarantee exact recovery. Moreover, in light of the above discussion, a computationally efficient algorithm that can achieve exact recovery at the information-theoretic limit would be highly desirable. Prior to the work Abbe et al. (2016), there were two polynomial-time algorithms that allow for exact recovery in the regime $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, which are respectively proposed in Boppana (1987) and McSherry (2001). However, Boppana (1987) needs $(\alpha - \beta)^2/(\alpha + \beta) > 72$ and McSherry (2001) needs $(\alpha - \beta)^2/(\alpha + \beta) > 64$ for exact recovery, both of which do not meet the information-theoretic limit $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$. The first computationally tractable algorithm that can achieve exact recovery at this limit was given in Abbe et al. (2016), which employs the polynomial-time algorithm in Massoulié (2014) for obtaining a partial recovery, followed by a local improvement procedure. Later, Hajek et al. (2016) and Bandeira (2018) independently proved that the solution of a semidefinite relaxation of Problem (1) can identify the underlying communities exactly with high probability if $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, confirming a conjecture raised in Abbe et al. (2016). As SDPs can be solved in polynomial time, this also provides a computationally tractable algorithm that achieves exact recovery at the information-theoretic limit. More recently, by conducting an entrywise eigenvector analysis of random matrices with low expected rank, Abbe et al. (2017) proved that the vanilla eigenvector-based algorithm achieves exact recovery at the information-theoretic limit. The dominating computational cost in this algorithm is to compute the eigenvector associated with the second largest eigenvalue of the adjacency matrix $\boldsymbol{A}$, which is known to be polynomial time computable.

We would also like to mention some algorithms for the considered problem with extremely low computational complexity. Condon & Karp (2001) proposed a simple combinatorial algorithm for minimum bisection problem, whose per iteration cost is linear in n. To achieve exact community recovery with high probability, it needs $p - q \geq n^{-1/2+\epsilon}$ with some $\epsilon > 0$. Recently, Bandeira et al. (2016) proposed to apply the Burer-Monteiro decomposition (Burer & Monteiro, 2003) to the semidefinite relaxation of Problem (1), which results in solving an optimization problem defined on a smooth manifold. They showed that all second-order stationary points of the problem, which can be computed efficiently by the Riemannian trust-region method (Boumal et al., 2018), correspond to the underlying communities with high probability as long as $(p-q)/\sqrt{p+q} \geq cn^{-1/6}$ for some constant $c > 0$. Despite their low computational complexity, both approaches require much stronger conditions on p and q for ensuring exact recovery. In particular, they cannot even allow for exact recovery in the regime $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ for some $\alpha > \beta > 0$.

1.2. Our Contribution

In this work, we propose a two-stage iterative algorithm for exact community recovery in the binary symmetric SBM. Our algorithm is based on the following regularized version of the ML estimation problem (1):

$$\max \left\{ \boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x} : x_i = \pm 1,\, i = 1, \dots, n \right\}, \quad (2)$$

where $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ and $\rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$. Specifically, the proposed iterative algorithm starts with a vector that is chosen randomly with uniform distribution over the unit sphere of $\mathbb{R}^n$, employs the standard power method (PM) in the first stage for approximating the dominant eigenvector of $\boldsymbol{B}$, and then applies a generalized power method (GPM) to Problem

(2) in the second stage that can exactly identify the underlying communities in a finite number of iterations. We show that in the regime $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ for some $\alpha > \beta > 0$, the proposed algorithm achieves exact recovery at the information-theoretic limit and enjoys a time complexity that scales nearly linear in n. More precisely, for any fixed α and β such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, we prove that the following event happens with probability tending to one as $n \to \infty$: the proposed algorithm terminates at the ground truth $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$[1] in $O(\log n/\log\log n)$ iterations of the PM and $O(\log n/\log\log n)$ iterations of the GPM, with $O(n \log n)$ computational complexity in each iteration of the PM and the GPM. At the heart of our result is a careful analysis on the performance of the PM and the GPM when applied to our problem. In particular, we show the following results hold with high probability:

(i) The sequence of iterates generated by the first stage of our algorithm converges linearly to a dominant eigenvector of $\boldsymbol{B}$ and the ratio in defining the linear rate of convergence tends to 0 as $n \to \infty$.

(ii) The iterates generated by the second stage of our algorithm possess a contraction property once their Euclidean distance to the ground truth is no greater than a threshold that scales linearly with $\sqrt{n/\log n}$.

(iii) The second stage of our algorithm exhibits finite termination as it admits a one-step convergence property once the generated iterate has Euclidean distance no greater than a constant threshold to the ground truth.

We also conduct experiments on synthetic data sets and compare the performance of the proposed algorithm to some existing ones. Numerical results support our theoretical development and also show the advantage of the proposed algorithm in terms of computational efficiency.

The rest of this paper is organized as follows. In section 2, we introduce the proposed two-stage algorithm for exact community recovery and present the main results of this paper regarding its recovery performance and computational efficiency. In sections 3 and 4, we respectively analyze the properties of the power method and the generalized power method that are respectively used in the two stages of the proposed algorithm. In section 5, we provide the proofs of the main results. Numerical results of the proposed algorithm are reported in section 6.

Notation. Let $\mathbb{R}^n$ be the n-dimensional Euclidean space and $\|\cdot\|_2$ be the Euclidean norm. We write matrices in capital bold letters like $\boldsymbol{A}$, vectors in bold lower case like $\boldsymbol{a}$, and scalars as plain letters. We denote by a_{ij} the (i, j)-th element of $\boldsymbol{A}$ and x_i the i-th element of $\boldsymbol{x}$. We use $\mathbf{1}_n$ and $\boldsymbol{E}_n$ to denote the all-one vector and all-one matrix of dimension n, respectively. Given a vector $\boldsymbol{d}$, we denote by $\mathbf{Diag}(\boldsymbol{d})$ the diagonal matrix with $\boldsymbol{d}$ on its diagonal. For any positive integer n, let $[n]$ denote the set $\{1, \dots, n\}$; for any discrete set T, let $|T|$ denote the number of elements in T. We use $\mathbf{Bern}(p)$ to denote the Bernoulli random variable with mean p. For any $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v}/|\boldsymbol{v}|$ denotes the vector of $\mathbb{R}^n$ defined as

$$\left(\frac{\boldsymbol{v}}{|\boldsymbol{v}|}\right)_i = \begin{cases} 1, & \text{if } v_i \geq 0, \\ -1, & \text{otherwise}, \end{cases} \qquad i = 1, \dots, n. \tag{3}$$

[1]Note that $\boldsymbol{x}^*$ and $-\boldsymbol{x}^*$ represent the same community structure up to a global flip of the labels.

2. Main Results

Given n nodes that are partitioned into two equal-sized clusters, we denote by $\boldsymbol{x}^*$ the vector that encodes its true community structure, *e.g.*, for every $i \in [n]$, $x_i^* = 1$ if the node i belongs to the first cluster and $x_i^* = -1$ if it belongs to the second one. Given $\boldsymbol{x}^*$, the binary symmetric SBM generates an undirected graph of n vertices, whose adjacency matrix $\boldsymbol{A}$ is randomly generated as follows.

Model 1. *The elements $\{a_{ij} : 1 \leq i \leq j \leq n\}$ of $\boldsymbol{A}$ are generated independently by*

$$a_{ij} \sim \begin{cases} \mathbf{Bern}(p), & \textit{if } x_i^* x_j^* = 1, \\ \mathbf{Bern}(q), & \textit{if } x_i^* x_j^* = -1, \end{cases}$$

where

$$p = \frac{\alpha \log n}{n} \quad \textit{and} \quad q = \frac{\beta \log n}{n}$$

for some constants $\alpha > \beta > 0$. Besides, we have $a_{ij} = a_{ji}$ for all $1 \leq j < i \leq n$.

Then, the problem of community recovery is to develop efficient methods that can find $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$ with high probability, given the adjacency matrix $\boldsymbol{A}$ generated from Model 1. Our proposal for tackling this problem is a two-stage algorithm summarized in Algorithm 1. It starts with a vector $\boldsymbol{y}^0$ that is chosen randomly with uniform distribution over the unit sphere of $\mathbb{R}^n$, employs the standard power iteration N times in the first stage (lines 4-6 of Algorithm 1), and then applies a generalized power iteration to Problem (2) in the second stage until a fixed point of the generalized power iteration is reached (lines 8–13 of Algorithm 1). We use the name generalized power iteration due to the fact that the operation $\boldsymbol{B}\boldsymbol{x}^{k-1}/|\boldsymbol{B}\boldsymbol{x}^{k-1}|$ (see (3) for its definition) is essentially computing a projection of $\boldsymbol{B}\boldsymbol{x}^{k-1}$ to the feasible set of Problem (2).

We next present the main result of this paper, which shows that Algorithm 1 achieves exact recovery at the information-theoretic limit and also provides explicit iteration complexity bounds for Algorithm 1 to exactly recover the underlying communities.

Algorithm 1 A Two-Stage Algorithm for Exact Recovery

Input: adjacency matrix $\boldsymbol{A}$, positive integer N
set $\rho \leftarrow \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$ and $\boldsymbol{B} \leftarrow \boldsymbol{A} - \rho \boldsymbol{E}_n$
choose $\boldsymbol{y}^0$ randomly with uniform distribution over the unit sphere of $\mathbb{R}^n$
for $k = 1, 2, \ldots, N$ **do**
  set $\boldsymbol{y}^k \leftarrow \boldsymbol{B}\boldsymbol{y}^{k-1} / \|\boldsymbol{B}\boldsymbol{y}^{k-1}\|_2$
end for
set $\boldsymbol{x}^0 \leftarrow \sqrt{n}\boldsymbol{y}^N$
for $k = 1, 2, \ldots$ **do**
  set $\boldsymbol{x}^k \leftarrow \boldsymbol{B}\boldsymbol{x}^{k-1} / |\boldsymbol{B}\boldsymbol{x}^{k-1}|$
  if $\boldsymbol{x}^k = \boldsymbol{x}^{k-1}$ **then**
    terminate and return $\boldsymbol{x}^k$
  end if
end for

Theorem 1. *Let $\boldsymbol{A}$ be randomly generated by Model 1. If $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, then the following statement holds with probability at least $1 - n^{-\Omega(1)}$: Algorithm 1 finds $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$ in $O(\log n / \log\log n)$ power iterations and $O(\log n / \log\log n)$ generalized power iterations.*

Note that the computational cost of both the power iteration and the generalized power iteration is dominated by the one for computing the matrix-vector product $\boldsymbol{B}\boldsymbol{v}$ for some $\boldsymbol{v} \in \mathbb{R}^n$. Since $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ and $\boldsymbol{E}_n = \mathbf{1}_n \mathbf{1}_n^T$ is a rank-one matrix, the cost of computing $\boldsymbol{B}\boldsymbol{v}$ is mainly determined by the number of non-zero entries in $\boldsymbol{A}$. Also, as $\boldsymbol{A}$ is randomly generated by Model 1, one can show by a simple concentration argument that the number of non-zero entries in $\boldsymbol{A}$ is, with high probability, in the order of $n \log n$. These, together with the iteration bounds provided in Theorem 1, imply that the time complexity of Algorithm 1 is with high probability nearly linear in n.

Corollary 1. *Let $\boldsymbol{A}$ be randomly generated by Model 1. If $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, then with probability at least $1 - n^{-\Omega(1)}$, Algorithm 1 finds $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$ in $O(n \log^2 n)$ time complexity.*

3. Analysis of the Power Method

In this section, we analyze the performance of the PM with a random initial point that is employed in the first-stage of Algorithm 1. In particular, we shall characterize the convergence rate of the iterates generated by the PM to a dominant eigenvector of $\boldsymbol{B}$. To begin, we present two lemmas that shall be used in studying the spectral property of $\boldsymbol{B}$. The proof of Lemma 1 can be found in the appendix, while the proof of Lemma 2 is omitted because it follows directly from Lei et al. (2015, Theorem 5.2).

Lemma 1. *Suppose that $\boldsymbol{A}$ is generated by Model 1 and $\rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$. Then, it holds with probability at least $1 - 2n^{-\frac{1}{2(\alpha+\beta+1)}}$ that*

$$\left| \rho - \frac{p+q}{2} \right| \leq \frac{\log n}{n^{3/2}}. \tag{4}$$

Lemma 2. *Let $\boldsymbol{A}$ be generated by Model 1. Then, there exist constants $c_1 \geq 1$ and $c_2 > 0$, whose values depend on α and β, such that*

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|_2 \leq c_1 \sqrt{\log n} \tag{5}$$

holds with probability at least $1 - c_2 n^{-3}$.

Equipped with Lemmas 1 and 2, and by applying Weyl's inequality, it is not hard to obtain the following result regarding the magnitudes of eigenvalues of $\boldsymbol{B}$. The proof of this result can be found in the appendix.

Lemma 3. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1 and $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ with $\rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\boldsymbol{B}$. Then, for all sufficiently large n, it holds with probability at least $1 - 2n^{-\frac{1}{2(\alpha+\beta+1)}} - c_2 n^{-3}$ that*

$$\lambda_1 \geq \frac{\alpha - \beta}{3} \log n \tag{6}$$

and

$$|\lambda_i| \leq 2c_1 \sqrt{\log n}, \quad i = 2, \ldots, n, \tag{7}$$

where c_1 and c_2 are the constants in Lemma 2.

Recall that the PM in Algorithm 1 starts with a random initial point. To characterize its convergence rate, we need the following result on the uniform distribution over the unit sphere of $\mathbb{R}^n$, whose proof follows similar arguments as those in Kuczyński & Woźniakowski (1992) and is provided in the appendix.

Lemma 4. *Suppose that $\boldsymbol{b} \in \mathbb{R}^n$ is randomly generated from the uniform distribution over the unit sphere of $\mathbb{R}^n$. Then, for all $n \geq 8$, it holds that*

$$\mathbb{P}\left(\sum_{i=2}^{n} \left(\frac{b_i}{b_1} \right)^2 \leq \frac{n^2}{2} \right) \geq 1 - 2n^{-1/2}. \tag{8}$$

Now we are ready to characterize the convergence rate of the PM used in the first-stage of Algorithm 1.

Proposition 1. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1. Let $\{\boldsymbol{y}^k\}_{k \geq 0}$ be the sequence generated in the first-stage of Algorithm 1. Then, for all sufficiently large n, it holds with probability at least $1 - n^{-\Omega(1)}$ that*

$$\min_{s \in \{\pm 1\}} \|\boldsymbol{y}^k - s\boldsymbol{u}_1\|_2 \leq n \cdot \left(\frac{6c_1}{(\alpha - \beta)\sqrt{\log n}} \right)^k \tag{9}$$

for all $k \geq 0$, where $\boldsymbol{u}_1$ is an eigenvector of $\boldsymbol{B}$ associated with the largest eigenvalue and c_1 is the constant in Lemma 2.

Proof. Let $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the eigenvalue decomposition of $\boldsymbol{B}$, where $\boldsymbol{\Lambda} = \mathbf{Diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$ being the eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n]$ are the associated eigenvectors of $\boldsymbol{B}$. Suppose that (6) and (7) hold, which, according to Lemma 3, happens with probability at least $1 - 2n^{-\frac{1}{2(\alpha+\beta+1)}} - c_2 n^{-3}$ for all sufficiently large n. Let us define $\boldsymbol{b} = \boldsymbol{U}^T \boldsymbol{y}^0$. We claim that if $b_1 > 0$, then

$$\|\boldsymbol{y}^k - \boldsymbol{u}_1\|_2 \leq \left(\frac{6c_1}{(\alpha-\beta)\sqrt{\log n}}\right)^k \cdot \sqrt{2\sum_{i=2}^n \left(\frac{b_i}{b_1}\right)^2}. \tag{10}$$

Indeed, since $\{\boldsymbol{y}^k\}_{k\geq 0}$ is generated by the power method and $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, we have

$$\langle \boldsymbol{y}^k, \boldsymbol{u}_1 \rangle = \frac{\boldsymbol{u}_1^T \boldsymbol{B}^k \boldsymbol{y}^0}{\|\boldsymbol{B}^k \boldsymbol{y}^0\|_2} = \frac{\boldsymbol{u}_1^T \boldsymbol{U}\boldsymbol{\Lambda}^k \boldsymbol{b}}{\|\boldsymbol{\Lambda}^k \boldsymbol{b}\|_2} = \frac{\lambda_1^k b_1}{\sqrt{\sum_{i=1}^n \lambda_i^{2k} b_i^2}}. \tag{11}$$

Notice that (6) and $\alpha > \beta$ imply that $\lambda_1 > 0$. Hence, if $b_1 > 0$, then $\lambda_1^k b_1 > 0$ for all $k \geq 0$, which yields that $\langle \boldsymbol{y}^k, \boldsymbol{u}_1 \rangle > 0$ for all $k \geq 0$. This, together with $\|\boldsymbol{y}^k\|_2 = \|\boldsymbol{u}_1\|_2 = 1$, leads to

$$\begin{aligned} 1 = \|\boldsymbol{y}^k\|_2 &\leq \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2 + \|\langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2 \\ &= \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2 + \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle, \end{aligned}$$

where the last equality follows from $\|\boldsymbol{u}_1\|_2 = 1$ and $\langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle > 0$. In addition, by Cauchy-Schwarz inequality, we have $\langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \leq \|\boldsymbol{y}^k\|_2 \|\boldsymbol{u}_1\|_2 = 1$. Thus, we obtain

$$0 \leq 1 - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle = \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2. \tag{12}$$

It then follows that

$$\begin{aligned} &\|\boldsymbol{y}^k - \boldsymbol{u}_1\|_2^2 \\ &= \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1 + \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1 - \boldsymbol{u}_1\|_2^2 \\ &= \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2^2 + \|\langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1 - \boldsymbol{u}_1\|_2^2 \\ &= \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2^2 + (1 - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle)^2 \\ &\leq 2\|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2^2, \end{aligned}$$

where the inequality is due to (12). In addition, by $\|\boldsymbol{y}^k\|_2 = \|\boldsymbol{u}_1\|_2 = 1$ and (11), we have

$$\begin{aligned} \|\boldsymbol{y}^k - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle \boldsymbol{u}_1\|_2^2 &= 1 - \langle \boldsymbol{y}^k, \boldsymbol{u}_1\rangle^2 = \frac{\sum_{i=2}^n \lambda_i^{2k} b_i^2}{\sum_{i=1}^n \lambda_i^{2k} b_i^2} \\ &\leq \frac{\sum_{i=2}^n \lambda_i^{2k} b_i^2}{\lambda_1^{2k} b_1^2} \leq \left(\frac{\bar{\lambda}}{\lambda_1}\right)^{2k} \sum_{i=2}^n \left(\frac{b_i}{b_1}\right)^2, \end{aligned}$$

where $\bar{\lambda} := \max\{|\lambda_i| : i = 2, \dots, n\}$. These, together with (6) and (7), establish the claim in (10). By the same arguments, one can show that if $b_1 < 0$, then

$$\|\boldsymbol{y}^k + \boldsymbol{u}_1\|_2 \leq \left(\frac{6c_1}{(\alpha-\beta)\sqrt{\log n}}\right)^k \cdot \sqrt{2\sum_{i=2}^n \left(\frac{b_i}{b_1}\right)^2}. \tag{13}$$

Since $\boldsymbol{y}^0$ is chosen according to the uniform distribution over the unit sphere of $\mathbb{R}^n$, which is known to be orthogonally invariant, we have that $\boldsymbol{b} = \boldsymbol{U}^T \boldsymbol{y}^0$ also follows the uniform distribution over the unit sphere of $\mathbb{R}^n$. It then follows from Lemma 4 that

$$\sum_{i=2}^n \left(\frac{b_i}{b_1}\right)^2 \leq \frac{n^2}{2} \tag{14}$$

holds with probability at least $1 - 2n^{-1/2}$. The desired result then follows from (10), (13), (14), and the union bound. $\square$

One can observe from Proposition 1 that with high probability, the sequence generated in the first-stage of Algorithm 1 converges at least linearly to the eigenvector of $\boldsymbol{B}$ associated with its largest eigenvalue. Moreover, for any fixed α and β, (9) shows that the ratio in the linear rate of convergence tends to 0 as $n \to \infty$.

4. Analysis of the Generalized Power Method

In this section, we analyze the performance of the GPM that is used in the second-stage of Algorithm 1. In particular, we shall show that with high probability, the sequence of iterates generated by the GPM converges to the ground truth $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$ in a finite number of iterations, provided that its initial iterate is in a suitable neighborhood of $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$. To begin, we present the following two lemmas that will be used for establishing the contraction property of the GPM. Their proofs can be found in the appendix.

Lemma 5. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1 and $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ with $\rho = \boldsymbol{1}_n^T \boldsymbol{A} \boldsymbol{1}_n / n^2$. For any $\boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = \sqrt{n}$, it holds with probability at least $1 - n^{-\Omega(1)}$ that*

$$\begin{aligned} \|\boldsymbol{B}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{x}^*\|_2 \leq &\left(\frac{\log n}{\sqrt{n}} + c_1\sqrt{\log n}\right)\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \\ &+ \frac{(\alpha-\beta)\log n}{4\sqrt{n}}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2. \end{aligned} \tag{15}$$

Lemma 6. *Let $\delta > 0$ be arbitrarily chosen. Then, for any $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}^n$ with $|v_i| \geq \delta$ for all $i = 1, 2, \dots, n$, the following inequality holds:*

$$\left\|\frac{\boldsymbol{u}}{|\boldsymbol{u}|} - \frac{\boldsymbol{v}}{|\boldsymbol{v}|}\right\|_2 \leq \frac{2\|\boldsymbol{u} - \boldsymbol{v}\|_2}{\delta}. \tag{16}$$

We are now ready to establish the following result, which implies that the GPM used in the second-stage of Algorithm 1 possesses a contraction property.

Lemma 7. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1 and $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ with $\rho = \boldsymbol{1}_n^T \boldsymbol{A} \boldsymbol{1}_n / n^2$. For any fixed $\alpha > \beta > 0$ such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, there exists a constant $\gamma > 0$ such that the following event happens with* *probability at least $1 - n^{-\Omega(1)}$: for all $\boldsymbol{x} \in \mathbb{R}^n$ such that $\|\boldsymbol{x}\|_2 = \sqrt{n}$ and*

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq \frac{6\sqrt{2}c_1}{\alpha - \beta} \cdot \sqrt{\frac{n}{\log n}}, \tag{17}$$

it holds that

$$\left\| \frac{\boldsymbol{Bx}}{|\boldsymbol{Bx}|} - \boldsymbol{x}^* \right\|_2 \leq \frac{8c_1}{\gamma\sqrt{\log n}} \cdot \|\boldsymbol{x} - \boldsymbol{x}^*\|_2. \tag{18}$$

Proof. Since $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ and $\boldsymbol{E}_n \boldsymbol{x}^* = \boldsymbol{0}$, we obtain $\boldsymbol{Bx}^* = \boldsymbol{Ax}^*$. By $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, there exists a postive constant γ such that $(\sqrt{\alpha} - \sqrt{\beta})^2/2 - \gamma \log(\alpha/\beta)/2 > 1$. These, together with (Abbe, 2017, Lemma 8), yield that

$$\min \{x_i^*(\boldsymbol{Bx}^*)_i : i = 1, 2, \ldots, n\} \geq \gamma \log n \tag{19}$$

holds with probability at least $1 - n^{-\Omega(1)}$. Notice that (19) implies $\boldsymbol{Bx}^*/|\boldsymbol{Bx}^*| = \boldsymbol{x}^*$ and

$$\min \{|(\boldsymbol{Bx}^*)_i| : i = 1, 2, \ldots, n\} \geq \gamma \log n. \tag{20}$$

It then follows from (20), Lemma 5, and Lemma 6 that for all $\boldsymbol{x} \in \mathbb{R}^n$ satisfying $\|\boldsymbol{x}\|_2 = \sqrt{n}$ and (17),

$$\begin{aligned}
&\left\| \frac{\boldsymbol{Bx}}{|\boldsymbol{Bx}|} - \boldsymbol{x}^* \right\|_2 = \left\| \frac{\boldsymbol{Bx}}{|\boldsymbol{Bx}|} - \frac{\boldsymbol{Bx}^*}{|\boldsymbol{Bx}^*|} \right\|_2 \leq \frac{2\|\boldsymbol{Bx} - \boldsymbol{Bx}^*\|_2}{\gamma \log n} \\
&\leq \left(\frac{(\alpha - \beta)}{2\gamma\sqrt{n}} \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 + \frac{2}{\gamma\sqrt{n}} + \frac{2c_1}{\gamma\sqrt{\log n}} \right) \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \\
&\leq \left(\frac{3\sqrt{2}c_1}{\gamma\sqrt{\log n}} + \frac{2}{\gamma\sqrt{n}} + \frac{2c_1}{\gamma\sqrt{\log n}} \right) \cdot \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \\
&\leq \frac{8c_1}{\gamma\sqrt{\log n}} \|\boldsymbol{x} - \boldsymbol{x}^*\|_2
\end{aligned}$$

for all n sufficiently large. $\square$

From Algorithm 1, we observe that all the iterates $\{\boldsymbol{x}^k\}$ generated by the GPM satisfy $\|\boldsymbol{x}^k\|_2 = \sqrt{n}$. Indeed, we have from line 7 of Algorithm 1 that $\|\boldsymbol{x}^0\|_2 = \sqrt{n}$ and $\|\boldsymbol{x}^k\|_2 = \sqrt{n}$ for all $k \geq 1$ due to the fact that $\boldsymbol{x}^k \in \{\pm 1\}^n$ for all $k \geq 1$. Thus, Lemma 7 implies that for all sufficiently large n, the iterates generated by the GPM possess a contraction property once their Euclidean distance to the ground truth is no greater than a threshold that scales linearly with $\sqrt{n/\log n}$. This is summarized in the following proposition.

Proposition 2. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1. Let $\alpha > \beta > 0$ be fixed such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$. Suppose that the $\boldsymbol{x}^0$ in Algorithm 1 satisfies $\|\boldsymbol{x}^0\|_2 = \sqrt{n}$ and*

$$\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2 \leq \frac{6\sqrt{2}c_1}{\alpha - \beta} \sqrt{\frac{n}{\log n}},$$

where c_1 is the constant in Lemma 2. Then, for all sufficiently large n, it holds with probability at least $1 - n^{-\Omega(1)}$ that

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2 \left(\frac{8c_1}{\gamma\sqrt{\log n}} \right)^k, \tag{21}$$

where $\gamma > 0$ is any constant such that $(\sqrt{\alpha} - \sqrt{\beta})^2/2 - \gamma \log(\alpha/\beta)/2 > 1$.

Furthermore, the following result indicates that the GPM exhibits finite termination. More precisely, it admits a one-step convergence property once the generated iterate has Euclidean distance no greater than a constant threshold to the ground truth.

Lemma 8. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1 and $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ with $\rho = \boldsymbol{1}_n^T \boldsymbol{A} \boldsymbol{1}_n / n^2$. For any fixed $\alpha > \beta > 0$ such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, the following event happens with probability at least $1 - n^{-\Omega(1)}$: for all $\boldsymbol{x} \in \{\pm 1\}^n$ such that $\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq 2$, it holds that*

$$\frac{\boldsymbol{Bx}}{|\boldsymbol{Bx}|} = \boldsymbol{x}^*. \tag{22}$$

Proof. Notice that to prove (22), it suffices to show that $x_i^*(\boldsymbol{Bx})_i > 0$ for all $i = 1, 2, \ldots, n$. Let $\boldsymbol{x} \in \{\pm 1\}^n$ with $\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq 2$ be arbitrarily chosen. Then, one can observe that $\boldsymbol{x} = \boldsymbol{x}^*$ or $\boldsymbol{x} = \boldsymbol{x}^* \pm 2\boldsymbol{e}_l$ for some $l = 1, 2, \ldots, n$, where $\boldsymbol{e}_l$ is the vector of $\mathbb{R}^n$ with all entries being 0 except the l-th entry being 1. If $\boldsymbol{x} = \boldsymbol{x}^*$, then it follows from (19) that $x_i^*(\boldsymbol{Bx})_i > 0$ for all $i = 1, 2, \ldots, n$. If $\boldsymbol{x} = \boldsymbol{x}^* \pm 2\boldsymbol{e}_l$ for some $l = 1, 2, \ldots, n$, then for any $i = 1, 2, \ldots, n$,

$$\begin{aligned}
x_i^*(\boldsymbol{Bx})_i &= x_i^*(\boldsymbol{Bx}^* \pm 2\boldsymbol{Be}_l)_i \\
&= x_i^*(\boldsymbol{Bx}^*)_i \pm 2x_i^*(\boldsymbol{Be}_l)_i \\
&\geq \gamma \log n \pm 2x_i^*(\boldsymbol{Be}_l)_i,
\end{aligned} \tag{23}$$

where the inequality follows from (19). Notice that $\boldsymbol{Be}_l = \boldsymbol{Ae}_l - \rho \boldsymbol{E}_n \boldsymbol{e}_l$. Hence, one has $(\boldsymbol{Be}_l)_i = A_{il} - \rho$. It then follows from Model 1 and (4) that

$$(\boldsymbol{Be}_l)_i \leq 1 - \frac{(\alpha + \beta) \log n}{2n} + \frac{\log n}{n^{3/2}}$$

and

$$(\boldsymbol{Be}_l)_i \geq -\frac{(\alpha + \beta) \log n}{2n} - \frac{\log n}{n^{3/2}}.$$

This, together with $|x_i^*| = 1$ and (23), implies that for all sufficiently large n, $x_i^*(\boldsymbol{Bx})_i > 0$ holds for all $i = 1, 2, \ldots, n$. $\square$

5. Proofs of Main Results

In this section, we provide the proofs of our main results Theorem 1 and Corollary 1. Recall that the iterates of the

PM converge to the eigenvector $\boldsymbol{u}_1$ of $\boldsymbol{B}$ associated with its largest eigenvalue, not to the ground truth $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$. As such, we need the following lemma that bounds the Euclidean distance between these two vectors. The proof of this lemma can be found in the appendix.

Lemma 9. *Suppose that $\boldsymbol{A}$ is randomly generated by Model 1 and $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ with $\rho = \boldsymbol{1}_n^T \boldsymbol{A} \boldsymbol{1}_n / n^2$. Let $\boldsymbol{u}_1$ be the eigenvector of $\boldsymbol{B}$ associated with its largest eigenvalue. Then, it holds with probability at least $1 - c_2 n^{-3}$ that*

$$\min_{s \in \{\pm 1\}} \left\| \boldsymbol{u}_1 - \frac{s\boldsymbol{x}^*}{\sqrt{n}} \right\|_2 \le \frac{3\sqrt{2}c_1}{(\alpha - \beta)\sqrt{\log n}}, \tag{24}$$

where c_1, c_2 are the constants in Lemma 2.

Now we are ready to provide the proof of Theorem 1.

Proof of Theorem 1. Suppose that (9) and (21) hold simultaneously, which, according to Propositions 1 and 2, happens with probability at least $1 - n^{-\Omega(1)}$. In the first stage, we require the power iterations to output a $\boldsymbol{y} \in \mathbb{R}^n$ such that $\|\boldsymbol{y}\|_2 = 1$ and $\|\boldsymbol{y} - \boldsymbol{u}_1\|_2 \le 3\sqrt{2}c_1 / ((\alpha - \beta)\sqrt{\log n})$. In view of Proposition 1, this can be achieved in N_p power iterations, where N_p is the smallest integer satisfying

$$n \left(\frac{6c_1}{(\alpha - \beta)\sqrt{\log n}} \right)^{N_p} \le \frac{3\sqrt{2}c_1}{(\alpha - \beta)\sqrt{\log n}}. \tag{25}$$

Thus, we have

$$N_p = \left\lceil \frac{\log\left((\alpha - \beta) n \sqrt{\log n}\right) - \log(3\sqrt{2}c_1)}{\log\left((\alpha - \beta)\sqrt{\log n}\right) - \log(6c_1)} \right\rceil,$$

which is roughly $O(\log n / \log\log n)$. Then, the second-stage of Algorithm 1 starts with $\boldsymbol{x}^0 = \sqrt{n}\boldsymbol{y}$, which satisfies

$$\begin{aligned} \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2 &\le \|\boldsymbol{x}_0 - \sqrt{n}\boldsymbol{u}_1\|_2 + \|\sqrt{n}\boldsymbol{u}_1 - \boldsymbol{x}^*\|_2 \\ &\le \frac{6\sqrt{2}c_1\sqrt{n}}{(\alpha - \beta)\sqrt{\log n}}, \end{aligned}$$

where the second inequality follows from (24) and (25). Thus, in the second stage, according to Proposition 2 and Lemma 8, we have that the GPM terminates at $\boldsymbol{x}$ or $-\boldsymbol{x}^*$ in at most N_g iterations, where N_g is the smallest integer satisfying

$$\frac{6\sqrt{2}c_1\sqrt{n}}{(\alpha - \beta)\sqrt{\log n}} \left(\frac{8c_1}{\gamma\sqrt{\log n}} \right)^{N_g} \le 2.$$

Thus, we have

$$N_g = \left\lceil \frac{\log(3\sqrt{2}c_1\sqrt{n}) - \log((\alpha - \beta)\sqrt{\log n})}{\log(\gamma\sqrt{\log n}) - \log(8c_1)} \right\rceil,$$

which is also roughly $O(\log n / \log\log n)$. □

As for the per-iteration cost of Algorithm 1, it mainly depends on the cost of computing the matrix-vector product $\boldsymbol{B}v$ for some $\boldsymbol{v} \in \mathbb{R}^n$ with $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{1}_n \boldsymbol{1}_n^T$. Note that the cost of computing $\boldsymbol{B}\boldsymbol{v}$ is determined by the sparsity of $\boldsymbol{A}$. As such, we need the following lemma that bounds the number of non-zero entries in $\boldsymbol{A}$. The proof of this lemma can be found in the appendix.

Lemma 10. *Let $\boldsymbol{A}$ be randomly generated by Model 1 with $\alpha,\ \beta > 0$. It holds with probability at least $1 - n^{-\Omega(1)}$ that the number of non-zero entries in $\boldsymbol{A}$ is less than $2(\alpha + \beta) n \log n$.*

Armed with the above results, we can prove Corollary 1.

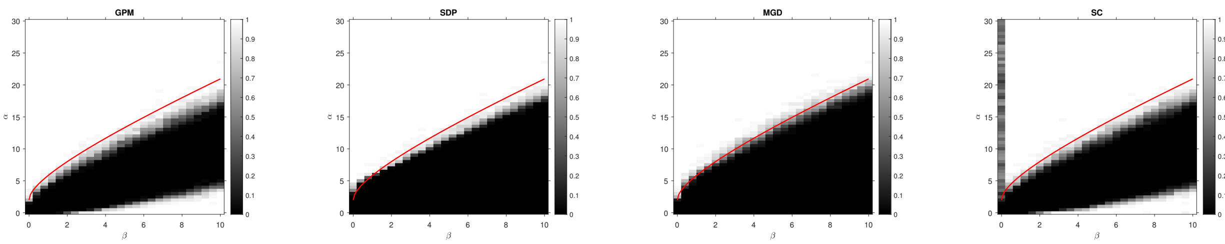
Proof of Corollary 1. According to Algorithm 1 and Lemma 10, the time complexity of each iteration is $2(\alpha + \beta) n \log n$. By Theorem 1, we have the total iteration number of Algorithm 1 is $O(\log n)$. Thus, the total time complexity of Algorithm 1 is $O(n \log^2 n)$ with probability at least $1 - n^{-\Omega(1)}$. □

6. Experimental Results

In this section, we conduct numerical experiments on synthetic data sets to test the performance of the proposed two-stage approach for community recovery in the binary symmetric SBM and compare it with several existing methods. For ease of reference, we denote our two-stage approach simply by GPM in this section. Our codes are implemented in MATLAB R2019a. The experiments are conducted on a PC with 16GB memory and Intel(R) Core(TM) i5-8600 3.10GHz CPU.

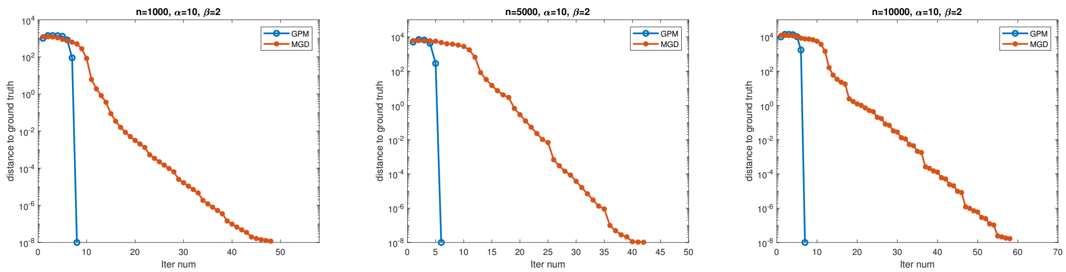
6.1. Phase Transition and Computational Efficiency

We first conduct the experiment of phase transition to test the recovery performance of our approach GPM and compare it with the SDP-based approach in Amini et al. (2018), the manifold optimization (MFO) based approach in Bandeira et al. (2016), and the spectral clustering (SC) approach in Abbe et al. (2017). In the implementation, we use alternating direction method of multipliers (ADMM) to solve the SDP as suggested in Amini et al. (2018), manifold gradient descent (MGD) method to solve the MFO, and the Matlab function eigs for computing the eigenvector that is needed in the SC approach. We choose $n = 300$ in the experiment and let α and β in Model 1 vary from 0 to 30 and 0 to 10, with increments 0.5 and 0.4, respectively. For every pair of α and β, we generate 40 instances and calculate, for all the tested methods, the ratio of exact recovery. The simulation results are presented in Figure 1. We can observe that all the methods admit a phase transition that is close to the information-theoretic limit, with the recovery performance of the GPM and the SC slightly better than



(a) (b) (c) (d)

Figure 1. Phase transition: the x-axis is β, the y-axis is α, and darker pixels represent lower empirical probability of success. The red curve is the information-theoretic threshold $\sqrt{\alpha} - \sqrt{\beta} = \sqrt{2}$.



(a) (b) (c)

Figure 2. Convergence performance: the x-axis is number of iterations, the y-axis for GPM is $\|\boldsymbol{x}^k \boldsymbol{x}^{k^T} - \boldsymbol{x}^* \boldsymbol{x}^{*^T}\|_F$, and the y-axis for MGD is $\|\boldsymbol{Q}^k \boldsymbol{Q}^{k^T} - \boldsymbol{x}^* \boldsymbol{x}^{*^T}\|_F$, where $\boldsymbol{x}^k$ and $\boldsymbol{Q}^k$ are the iterates generated in the k-th iteration of GPM and MGD, respectively.

the other two. This supports our theoretical development that the proposed algorithm can achieve exact recovery at the information-theoretic limit. Besides, as an indicator of computational efficiency, we record in Table 1 the total CPU time consumed by every approach for completing the phase transition experiments. It can be observed that the GPM slightly outperforms the SC in terms of CPU time, and both of them are substantially computationally more efficient than the SDP and the MGD.

Table 1. Total CPU times (in seconds) consumed by the test approaches in the experiment of phase transition.

Methods	GPM	SDP	MGD	SC
Time (s)	28	9193	1063	124

6.2. Convergence Performance

Next we conduct the experiments of convergence performance to test the number of iterations needed by our approach GPM to exactly identify the underlying communities. For comparison, we also test the convergence performance of MGD, which is an iterative algorithm that has similar per-iteration cost to GPM. In the experiments, we use $\alpha = 10$ and $\beta = 2$ and generate graphs of dimensions $n = 1000$, 5000, and 10000 by Model 1. We use $\|\boldsymbol{x}^k \boldsymbol{x}^{k^T} - \boldsymbol{x}^* \boldsymbol{x}^{*^T}\|_F$ (resp. $\|\boldsymbol{Q}^k \boldsymbol{Q}^{k^T} - \boldsymbol{x}^* \boldsymbol{x}^{*^T}\|_F$) as the measure of distance from $\boldsymbol{x}^k$ (resp. $\boldsymbol{Q}^k$) to the ground truth $\boldsymbol{x}^*$ or $-\boldsymbol{x}^*$, where $\boldsymbol{x}^k$ (resp. $\boldsymbol{Q}^k$) is the iterate generated in the k-th iteration of GPM (resp. MGD). In Figure 2, we plot this measure of distance to the ground truth against the iteration number for both GPM and MGD. It is observed that GPM exhibits a finite termination phenomenon and converges to the ground truth much faster than MGD. This also corroborates our theoretical development that our approach identifies the underlying communities in at most $O(\log n / \log\log n)$ number of PM and GPM iterations.

7. Concluding remarks

In this work, we consider the problem of exact community recovery in the binary symmetric SBM. We propose a two-stage iterative algorithm to solve it, which employs the power method firstly and then turns to the generalized power method. We show that the proposed method can achieve exact recovery at the information-theoretic limit in $\tilde{O}(n)$ time as n goes to infinity. Numerical experiments on synthetic data sets demonstrate that the proposed approach has strong recovery performance and is highly efficient.

As suggested in our numerical simulations, applying Algorithm 1 to Problem (2) seems to have a significantly larger one-step convergence region than the constant region in Lemma 8, which can be further improved in future research. Also, it is worthy to study whether the proposed approach can be extended to more general SBMs.

Acknowledgements

The first and third authors are supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) project CUHK 14208117 and in part by the CUHK Research Sustainability of Major RGC Funding Schemes project 3133236. The second author is supported in part by the National Natural Science Foundation of China (NSFC) project 11901490 and in part by an HKBU Start-up Grant. Most of the work of the second author was done when he was affiliated with the Department of Mathematics of Hong Kong Baptist University.

References

Abbe, E. Community detection and stochastic block models: Recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Abbe, E. and Sandon, C. 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*, pp. 670–688. IEEE, 2015.

Abbe, E., Bandeira, A. S., and Hall, G. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Amini, A. A., Levina, E., et al. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1): 149–179, 2018.

Bandeira, A. S. Random Laplacian matrices and convex relaxations. *Foundations of Computational Mathematics*, 18(2):345–379, 2018.

Bandeira, A. S., Boumal, N., and Voroninski, V. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on Learning Theory*, pp. 361–382, 2016.

Boppana, R. B. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science*, pp. 280–285. IEEE, 1987.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2018.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Chen, Y. and Xu, J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *The Journal of Machine Learning Research*, 17(1):882–938, 2016.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. Integration of biological networks and gene expression data using cytoscape. *Nature Protocols*, 2(10):2366, 2007.

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

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18 (1):1980–2024, 2017.

Garey, M. R., Johnson, D. S., and Stockmeyer, L. Some simplified NP-complete problems. In *Proceedings of the 6th annual ACM Symposium on Theory of Computing*, pp. 47–63, 1974.

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

Hajek, B., Wu, Y., and Xu, J. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

Kuczyński, J. and Woźniakowski, H. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992.

Lei, J., Rinaldo, A., et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43 (1):215–237, 2015.

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

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

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

Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*, 69 (2):026113, 2004.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Appendix

In the appendix, we provide proofs of some technical results presented in Sections 3, 4, and 5. To proceed, we introduce some further notation. Given two random variables X and Y, we write $X \overset{d}{=} Y$ if X and Y are equal in distribution.

A. Proofs of the Technical Results in Section 3

1. Proof of Lemma 1

Proof. Recall that $\rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$. Since $\boldsymbol{A}$ is generated by Model 1, one can verify that

$$n^2 \rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n \overset{d}{=} \sum_{i=1}^{m^2-m} 2W_i + \sum_{i=m^2-m+1}^{m^2+m} W_i + \sum_{i=1}^{m^2} 2Z_i, \tag{26}$$

where $m = n/2$, $\{W_i : i = 1, \dots, m^2 + m\}$ are i.i.d. $\mathbf{Bern}(p)$, and $\{Z_i : i = 1, \dots, m^2\}$ are i.i.d. $\mathbf{Bern}(q)$ and independent of $\{W_i : i = 1, \dots, m^2 + m\}$. Thus, we have

$$\mathbb{E}(n^2\rho) = \frac{n^2(p+q)}{2}, \quad \text{Var}(n^2\rho) = (4m^2 - 2m)p(1-p) + 4m^2 q(1-q) \le n^2(p+q).$$

Then, by applying Bernstein's inequality for bounded distributions (see, e.g., (Vershynin, 2018, Theorem 2.8.4)) to (26), we obtain

$$\begin{aligned}
\mathbb{P}\left(\left| n^2\rho - \frac{n^2(p+q)}{2} \right| \ge \sqrt{n}\log n \right) &\le 2\exp\left(-\frac{n\log^2 n/2}{n^2(p+q) + 2\sqrt{n}\log n/3} \right) \\
&= 2\exp\left(-\frac{\log n}{2(\alpha+\beta) + 4/(3\sqrt{n})} \right) \\
&\le 2\exp\left(-\frac{\log n}{2(\alpha+\beta+1)} \right) \\
&= 2n^{-\frac{1}{2(\alpha+\beta+1)}},
\end{aligned}$$

where the second inequality uses $4/(3\sqrt{n}) < 2$ for $n \ge 1$. This implies

$$\mathbb{P}\left(\left| \rho - \frac{p+q}{2} \right| \le \frac{\log n}{n^{3/2}} \right) \ge 1 - 2n^{-\frac{1}{2(\alpha+\beta+1)}}$$

as desired. □

2. Proof of Lemma 3

Proof. Since $\boldsymbol{A}$ is generated by Model 1 and $\rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$, one can verify that

$$\mathbb{E}[\boldsymbol{A}] = \frac{p+q}{2}\boldsymbol{E}_n + \frac{p-q}{2}\boldsymbol{x}^*\boldsymbol{x}^{*T}, \quad \mathbb{E}[\rho] = \frac{\mathbf{1}_n^T \mathbb{E}[\boldsymbol{A}] \mathbf{1}_n}{n^2} = \frac{p+q}{2}.$$

This, together with $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$, yields

$$\mathbb{E}[\boldsymbol{B}] = \mathbb{E}[\boldsymbol{A}] - \mathbb{E}[\rho]\boldsymbol{E}_n = \frac{p-q}{2}\boldsymbol{x}^*\boldsymbol{x}^{*T}. \tag{27}$$

Moreover, by letting $\Delta = \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$, we have

$$\boldsymbol{B} = \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}] + \mathbb{E}[\boldsymbol{A}] - \mathbb{E}[\boldsymbol{B}] + \mathbb{E}[\boldsymbol{B}] - \rho\boldsymbol{E}_n = \Delta + \left(\frac{p+q}{2} - \rho \right)\boldsymbol{E}_n + \mathbb{E}[\boldsymbol{B}]. \tag{28}$$

Besides, by Lemma 1 and Lemma 2, it holds with probability at least $1 - 2n^{-\frac{1}{2(\alpha+\beta+1)}} - c_2 n^{-3}$ that

$$\left\| \Delta + \left(\frac{p+q}{2} - \rho \right) \boldsymbol{E}_n \right\|_2 \le \|\Delta\|_2 + \left| \frac{p+q}{2} - \rho \right| \cdot \|\boldsymbol{E}_n\|_2 \le c_1 \sqrt{\log n} + \frac{\log n}{\sqrt{n}} \le 2c_1 \sqrt{\log n},$$

where the second inequality is due to Lemma 1, Lemma 2, and $\|\boldsymbol{E}_n\|_2 = n$ and the last inequality follows from $n \ge \log n$ and $c_1 \ge 1$. From (27), one can observe that $\mathbb{E}[\boldsymbol{B}]$ is a rank-one matrix with $n(p-q)/2 > 0$ being its non-zero eigenvalue. Then, applying Weyl's inequality to (28) yields that with probability at least $1 - 2n^{-\frac{1}{2(\alpha+\beta+1)}} - c_2 n^{-3}$,

$$\lambda_1 \ge \frac{n(p-q)}{2} - \left\| \Delta + \left(\frac{p+q}{2} - \rho \right) \boldsymbol{E}_n \right\|_2 \ge \frac{\alpha - \beta}{2} \log n - 2c_1 \sqrt{\log n}, \tag{29}$$

and for every $i = 2, \ldots, n$,

$$|\lambda_i| \le \left\| \Delta + \left(\frac{p+q}{2} - \rho \right) \boldsymbol{E}_n \right\|_2 \le 2c_1 \sqrt{\log n}. \tag{30}$$

The desired result then follows from (29), (30), and the fact that

$$\frac{\alpha - \beta}{2} \log n - 2c_1 \sqrt{\log n} \ge \frac{\alpha - \beta}{3} \log n$$

for all sufficiently large n. □

3. Proof of Lemma 4

Proof. The proof of Lemma 4 follows similar arguments as those in (Kuczyński & Woźniakowski, 1992). We denote by $\mathbb{B}_n(r)$ the n-dimensional Euclidean ball of radius $r > 0$, i.e., $\mathbb{B}_n(r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \le r\}$, and define $\mathbb{I} : \mathbb{R}^n \to \{0, 1\}$ as

$$\mathbb{I}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \sum_{i=2}^{n} \left(\frac{x_i}{x_1} \right)^2 \le \frac{n^2}{2}, \\ 0, & \text{otherwise}, \end{cases} \qquad \forall \boldsymbol{x} \in \mathbb{B}_n(1). \tag{31}$$

Recall that $\boldsymbol{b} \in \mathbb{R}^n$ is randomly generated from the uniform distribution over the unit sphere of $\mathbb{R}^n$. It follows that

$$\mathbb{P}\left(\sum_{i=2}^{n} \left(\frac{b_i}{b_1} \right)^2 \le \frac{n^2}{2} \right) = \int_{\|\boldsymbol{x}\|_2 = 1} \mathbb{I}(\boldsymbol{x}) \mu(d\boldsymbol{x}), \tag{32}$$

where μ is the probability measure function of the uniform distribution. Notice from (31) that $\mathbb{I}$ satisfies $\mathbb{I}(\alpha \boldsymbol{x}) = \mathbb{I}(\boldsymbol{x})$ for all $\alpha > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$, and $\mathbb{I}$ does not depend on signs of x_i's, i.e., $\mathbb{I}(\boldsymbol{x}) = \mathbb{I}(s_1 x_1, s_2 x_2, \ldots, s_n x_n)$ for any $s_i = \pm 1$ and $i = 1, 2, \ldots, n$. It then follows from (Kuczyński & Woźniakowski, 1992, Remark 7.2) that

$$\int_{\|\boldsymbol{x}\|_2 = 1} \mathbb{I}(\boldsymbol{x}) \mu(d\boldsymbol{x}) = \frac{1}{V_n} \int_{\mathbb{B}_n(1)} \mathbb{I}(\boldsymbol{x}) d\boldsymbol{x}, \tag{33}$$

where V_n is the volume of the n-dimensional unit ball $\mathbb{B}_n(1)$. Moreover, by (31), one has

$$\int_{\mathbb{B}_n(1)} \mathbb{I}(\boldsymbol{x}) d\boldsymbol{x} = 2 \int_0^1 \left(\int_{\sum_{i=2}^n x_i^2 \le \min\{1 - x_1^2,\ x_1^2 n^2/2\}} dx_2 \ldots dx_n \right) dx_1. \tag{34}$$

Notice that

$$\int_{\sum_{i=2}^n x_i^2 \le \min\{1 - x_1^2,\ x_1^2 n^2/2\}} dx_2 \ldots dx_n$$

is the volume of the $(n-1)$-dimensional ball $\mathbb{B}_{n-1}\left(\min\{1 - x_1^2,\ x_1^2 n^2/2\}^{1/2}\right)$ and it equals $V_{n-1} \cdot \min\{1 - x_1^2,\ x_1^2 n^2/2\}^{(n-1)/2}$, where V_{n-1} is the volume of the $(n-1)$-dimensional unit ball $\mathbb{B}_{n-1}(1)$. This, together with (32), (33), and (34), implies that

$$\mathbb{P}\left(\sum_{i=2}^{n} \left(\frac{b_i}{b_1} \right)^2 \le \frac{n^2}{2} \right) = \frac{2V_{n-1}}{V_n} \int_0^1 \min\{1 - x_1^2,\ x_1^2 n^2/2\}^{\frac{n-1}{2}} dx_1. \tag{35}$$

Since $1 - x_1^2 \le x_1^2 n^2/2$ for any $x_1 \ge \sqrt{2/(2+n^2)}$, we have

$$\begin{aligned}
\int_0^1 \min\{1 - x_1^2,\ x_1^2 n^2/2\}^{\frac{n-1}{2}} dx_1 &= \int_0^{\sqrt{\frac{2}{2+n^2}}} \left(\frac{x_1^2 n^2}{2}\right)^{\frac{n-1}{2}} dx_1 + \int_{\sqrt{\frac{2}{2+n^2}}}^1 (1 - x_1^2)^{\frac{n-1}{2}} dx_1 \\
&\ge \int_{\sqrt{\frac{2}{2+n^2}}}^1 (1 - x_1^2)^{\frac{n-1}{2}} dx_1 \\
&= \int_0^1 (1 - x_1^2)^{\frac{n-1}{2}} dx_1 - \int_0^{\sqrt{\frac{2}{2+n^2}}} (1 - x_1^2)^{\frac{n-1}{2}} dx_1 \\
&\ge \frac{V_n}{2V_{n-1}} - \sqrt{\frac{2}{2+n^2}}.
\end{aligned}$$

In addition, by (Kuczyński & Woźniakowski, 1992, Eq. (13)), we have

$$\frac{V_{n-1}}{V_n} \le \frac{192}{105\pi} \cdot \sqrt{\frac{n}{2}} \le 0.412\sqrt{n}$$

for all $n \ge 8$. It then follows from (35) that for all $n \ge 8$,

$$\mathbb{P}\left(\sum_{i=2}^n \left(\frac{b_i}{b_1}\right)^2 \le \frac{n^2}{2}\right) \ge 1 - \frac{2V_{n-1}}{V_n} \cdot \sqrt{\frac{2}{2+n^2}} \ge 1 - 1.165\sqrt{\frac{n}{2+n^2}} \ge 1 - 2n^{-1/2},$$

which completes the proof. □

B. Proofs of the Technical Results in Section 4

The second stage of Algorithm 1 is the generalized power method (GPM), which starts with the point $\boldsymbol{x}^0 = \sqrt{n}\boldsymbol{y}^N$ and updates as follows

$$\boldsymbol{x}^k = \frac{\boldsymbol{B}\boldsymbol{x}^{k-1}}{|\boldsymbol{B}\boldsymbol{x}^{k-1}|}, \qquad \text{where } \left(\frac{\boldsymbol{v}}{|\boldsymbol{v}|}\right)_j = \begin{cases} 1 & \text{if } v_j \ge 0 \\ -1 & \text{otherwise} \end{cases}, \quad \forall \boldsymbol{v} \in \mathbb{R}^n,\ j = 1, \dots, n. \tag{36}$$

Recall that

$$\boldsymbol{B} = \frac{p-q}{2}\boldsymbol{x}^*\boldsymbol{x}^{*T} + \left(\frac{p+q}{2} - \rho\right)\boldsymbol{E}_n + \Delta, \tag{37}$$

where $\boldsymbol{x}^* \in \mathbb{R}^n$ is the ground truth and $\Delta = \boldsymbol{A} - \mathbb{E}(\boldsymbol{A})$.

1. Proof of Lemma 5

Proof. Suppose that (4) and (5) hold simultaneously, which, according to Lemma 1, Lemma 2, and the union bound, happens with probability at least $1 - n^{-\Omega(1)}$. Let $\boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = \sqrt{n}$ be arbitrarily chosen. Since $\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}^*\|_2 = \sqrt{n}$, we have

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x}\|_2 + \|\boldsymbol{x}^*\|_2^2 - 2\boldsymbol{x}^T\boldsymbol{x}^* = 2\boldsymbol{x}^{*T}\boldsymbol{x}^* - 2\boldsymbol{x}^T\boldsymbol{x}^*,$$

which implies

$$\left|\boldsymbol{x}^{*T}\boldsymbol{x}^* - \boldsymbol{x}^T\boldsymbol{x}^*\right| = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2.$$

This, together with (4), (5), (37), and $\|\boldsymbol{E}_n\|_2 = n$, yields that

$$\begin{aligned}
\|\boldsymbol{B}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{x}^*\|_2 &= \left\|\frac{p-q}{2}\left(\boldsymbol{x}^T\boldsymbol{x}^* - \boldsymbol{x}^{*T}\boldsymbol{x}^*\right)\boldsymbol{x}^* + \left(\frac{p+q}{2} - \rho\right)\boldsymbol{E}_n(\boldsymbol{x} - \boldsymbol{x}^*) + \Delta(\boldsymbol{x} - \boldsymbol{x}^*)\right\|_2 \\
&\le \frac{p-q}{2}\sqrt{n}\left|\boldsymbol{x}^T\boldsymbol{x}^* - \boldsymbol{x}^{*T}\boldsymbol{x}^*\right| + \left|\frac{p+q}{2} - \rho\right| \cdot \|\boldsymbol{E}_n\|_2\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 + \|\Delta\|_2\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \\
&\le \frac{p-q}{4}\sqrt{n}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{\log n}{\sqrt{n}} \cdot \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 + c_1\sqrt{\log n} \cdot \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \\
&= \left(\frac{\log n}{\sqrt{n}} + c_1\sqrt{\log n}\right)\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 + \frac{(\alpha - \beta)\log n}{4\sqrt{n}}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2
\end{aligned}$$

as desired. $\square$

2. Proof of Lemma 6

Proof. Notice that it suffices to prove that for any $x,\ y \in \mathbb{R}$ with $|y| \geq \delta$, one has

$$\left| \frac{x}{|x|} - \frac{y}{|y|} \right| \leq \frac{2|x-y|}{\delta}. \tag{38}$$

This, together with the property of $\|\cdot\|_2$, would imply the desired result in (16). To prove (38), we assume without loss of generality that $y \geq \delta$ and consider the two cases $x \geq 0$ and $x < 0$ separately. If $x \geq 0$, then by (36) and $y \geq \delta$, we have $x/|x| = y/|y| = 1$ and (38) holds trivially. If $x < 0$, then by $y \geq \delta > 0$, we have

$$\left| \frac{x}{|x|} - \frac{y}{|y|} \right| = 2 \leq \frac{2y}{\delta} \leq \frac{2(y-x)}{\delta} = \frac{2|x-y|}{\delta}.$$

The proof is then completed. $\square$

3. Proof of (19) in Lemma 7

The following lemma is to establish (19), which is used in the proof of Lemma 7.

Lemma 11. *Let $\boldsymbol{A}$ be randomly generated by Model 1 and $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ with $\rho = \mathbf{1}_n^T \boldsymbol{A} \mathbf{1}_n / n^2$. For any fixed $\alpha > \beta > 0$ such that $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, let $\gamma > 0$ be arbitrarily chosen such that*

$$\frac{(\sqrt{\alpha} - \sqrt{\beta})^2}{2} - \frac{\gamma \log(\alpha/\beta)}{2} > 1.$$

Then, it holds with probability at least $1 - n^{1-(\sqrt{\alpha}-\sqrt{\beta})^2/2 + \gamma \log(\alpha/\beta)/2}$ that

$$\min \left\{ x_i^* (\boldsymbol{B}\boldsymbol{x}^*)_i : i = 1, 2, \ldots, n \right\} \geq \gamma \log n. \tag{39}$$

Proof. Since $\boldsymbol{A}$ is generated by Model 1, one can verify that for every $i = 1, \ldots, n$,

$$x_i^* \left(\boldsymbol{A}\boldsymbol{x}^* \right)_i \stackrel{d}{=} \sum_{i=1}^{m} W_i - \sum_{i=1}^{m} Z_i, \tag{40}$$

where $\{W_i\}_{i=1}^m$ are i.i.d. $\mathbf{Bern}(p)$, and $\{Z_i\}_{i=1}^m$ are i.i.d. $\mathbf{Bern}(q)$, independent of $\{W_i\}_{i=1}^m$. By $p = \alpha \log n / n$ and $q = \beta \log n / n$, it follows from Lemma 8 of Abbe et al. (2017) that for any $\gamma \in \mathbb{R}$,

$$\mathbb{P}\left(\sum_{i=1}^{m} W_i - \sum_{i=1}^{m} Z_i \geq \gamma \log n \right) \geq 1 - n^{-(\sqrt{\alpha}-\sqrt{\beta})^2/2 + \gamma \log(\alpha/\beta)/2}.$$

This, together with (40) and the union bound, yields

$$\mathbb{P}\left(\min \left\{ x_i^* (\boldsymbol{A}\boldsymbol{x}^*)_i : i = 1, 2, \ldots, n \right\} \geq \gamma \log n \right) \geq 1 - n^{1-(\sqrt{\alpha}-\sqrt{\beta})^2/2 + \gamma \log(\alpha/\beta)/2}.$$

In addition, since $\boldsymbol{B} = \boldsymbol{A} - \rho \boldsymbol{E}_n$ and $\boldsymbol{E}_n \boldsymbol{x}^* = \mathbf{0}$, we have $\boldsymbol{B}\boldsymbol{x}^* = \boldsymbol{A}\boldsymbol{x}^*$. The desired result in Lemma 11 then follows from this and the above inequality. $\square$

C. Proofs of the Technical Results in Section 5

1. Proof of Lemma 9

Proof. Suppose that (4) and (5) hold, which, according to Lemma 1, Lemma 2, and the union bound, happens with probability at least $1 - n^{-\Omega(1)}$. Recall from (28) that $\boldsymbol{B} = \mathbb{E}[\boldsymbol{B}] + \tilde{\Delta}$, where $\tilde{\Delta} = \Delta + ((p+q)/2 - \rho) \boldsymbol{E}_n$ and $\Delta = \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$. Also, $\boldsymbol{u}_1$ is an eigenvector of $\boldsymbol{B}$ associated with the largest eigenvalue of $\boldsymbol{B}$ and $\|\boldsymbol{u}_1\|_2 = 1$. In addition, recall from (27) that

$\mathbb{E}[\boldsymbol{B}]$ is a rank-one matrix with $n(p-q)/2 > 0$ being its non-zero eigenvalue and $1/\sqrt{n}\boldsymbol{x}^*$ being the associated eigenvector. Then, upon applying Davis-Kahan $\sin\Theta$ theorem, we obtain

$$\min_{s\in\{\pm 1\}} \left\| \boldsymbol{u}_1 - \frac{s\boldsymbol{x}^*}{\sqrt{n}} \right\|_2 \leq \frac{\sqrt{2}\|\tilde{\Delta}\boldsymbol{x}^*\|_2}{\sqrt{n}\left(\frac{n(p-q)}{2} - \|\tilde{\Delta}\|_2\right)},$$

provided that $\|\tilde{\Delta}\|_2 < n(p-q)/2$. By $\tilde{\Delta} = \Delta + ((p+q)/2 - \rho)\boldsymbol{E}_n$, $\boldsymbol{E}_n\boldsymbol{x}^* = \boldsymbol{0}$, and (5), we have

$$\|\tilde{\Delta}\boldsymbol{x}^*\|_2 = \|\Delta\boldsymbol{x}^*\|_2 \leq \|\Delta\|_2\|\boldsymbol{x}^*\|_2 \leq c_1\sqrt{n\log n}.$$

Besides, by $p = \alpha\log(n)/n$, $q = \beta\log(n)/n$, (4), (5), and $\|\boldsymbol{E}_n\|_2 = n$, we obtain

$$\frac{n(p-q)}{2} - \|\tilde{\Delta}\|_2 \geq \frac{(\alpha-\beta)\log n}{2} - \|\Delta\|_2 - n\cdot\left|\frac{p+q}{2} - \rho\right| \geq \frac{(\alpha-\beta)\log n}{2} - c_1\sqrt{\log n} - \frac{\log n}{\sqrt{n}} \geq \frac{(\alpha-\beta)\log n}{3} > 0$$

for all sufficiently large n. Combining the above three inequalities together, we obtain

$$\min_{s\in\{\pm 1\}} \left\| \boldsymbol{u}_1 - \frac{s\boldsymbol{x}^*}{\sqrt{n}} \right\|_2 \leq \frac{3\sqrt{2}c_1}{(\alpha-\beta)\sqrt{\log n}}$$

as desired. □

2. Proof of Lemma 10

Proof. First, let us compute the number of non-zero entries of some column $\boldsymbol{a} \in \mathbb{R}^n$ of $\boldsymbol{A}$. According to Model 1, we have

$$\|\boldsymbol{a}\|_0 \overset{d}{=} \sum_{i=1}^m W_i + \sum_{i=1}^m Z_i,$$

where $\{W_i\}_{i=1}^m$ are i.i.d. $\mathbf{Bern}(p)$, and $\{Z_i\}_{i=1}^m$ are i.i.d. $\mathbf{Bern}(q)$, independent of $\{W_i\}_{i=1}^m$. It then follows that

$$\mathbb{E}[\|\boldsymbol{a}\|_0] = m(p+q), \qquad \mathrm{Var}[\|\boldsymbol{a}\|_0] = mp(1-p) + mq(1-q) \leq m(p+q).$$

Applying the Bernstein's inequality for bounded distribution yields that

$$\begin{aligned}
\mathbb{P}\left(|\|\boldsymbol{a}\|_0 - m(p+q)| \geq 3m(p+q)\right) &\leq 2\exp\left(-\frac{9m^2(p+q)^2/2}{m(p+q)+m(p+q)}\right) \\
&= 2\exp\left(-\frac{9}{4}m(p+q)\right) \\
&= 2\exp\left(-\frac{9}{8}(\alpha+\beta)\log n\right) \\
&= 2n^{-\frac{9}{8}(\alpha+\beta)}.
\end{aligned}$$

This implies

$$\mathbb{P}\left(\|\boldsymbol{a}\|_0 < 2n(p+q)\right) \geq 1 - 2n^{-\frac{9}{8}(\alpha+\beta)} \geq 1 - 2n^{-\frac{9}{4}},$$

where the second inequality is due to $\alpha+\beta > 2$, which follows from $\alpha > \beta > 0$ and $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$. Finally, upon applying the union bound to the n columns of $\boldsymbol{A}$, we can conclude that it holds with probability at least $1 - 2n^{-5/4}$ that the number of non-zero entries in $\boldsymbol{A}$ is less than $2n^2(p+q) = 2(\alpha+\beta)n\log n$. □