Proceedings of Machine Learning Research vol 99:1–35, 2019 32nd Annual Conference on Learning Theory

Achieving the Bayes Error Rate in Stochastic Block Model by SDP, Robustly

Yingjie Fei YF275@CORNELL.EDU
Yudong Chen YUDONG.CHEN@CORNELL.EDU
Cornell University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the statistical performance of the semidefinite programming (SDP) relaxation approach for clustering under the binary symmetric Stochastic Block Model (SBM). We show that the SDP achieves an error rate of the form

$$\exp\left[-(1-o(1))\frac{nI}{2}\right],$$

where I is an appropriate information-theoretic measure of the signal-to-noise ratio. This bound matches the minimax lower bound on the optimal Bayes error rate for this problem, and improves upon existing results that are sub-optimal by a multiplicative constant in the exponent. As a corollary, our result implies that SDP achieves the optimal exact recovery threshold with the correct leading constant. We further show that this error rate of SDP is robust; that is, it remains unchanged under the so-called semirandom model where the graph is modified by a monotone adversary, as well as under the setting with heterogeneous edge probabilities. Our proof is based on a novel primal-dual analysis of the SDP.

Keywords: Stochastic Block Model, semidefinite programming, minimax rates, Bayes risk, robustness.

1. Introduction

The Stochastic Block Model (SBM) is a popular probabilistic model for studying clustering and community detection problems. In SBM, a set of n nodes is partitioned into several unknown clusters, where nodes in the same cluster are more likely to be connected than those in different clusters. Clustering under SBM entails determining the cluster membership of each node based on a single realization of the random graph of connections. In this paper, we focus on the canonical *binary symmetric SBM* with two equal-sized clusters. In this model, the ground-truth cluster membership can be represented by the cluster label vector $\boldsymbol{\sigma}^* \in \{\pm 1\}^n$ such that σ_i^* is the cluster label of node i and $\sum_i \sigma_i^* = 0$. Two nodes i and j are connected with probability p if $\sigma_i^*\sigma_j^* = 1$ (i.e., they are in the same cluster), and with probability q if $\sigma_i^*\sigma_j^* = -1$ (i.e., they are in different clusters), where $p > q$. The goal is to estimate the true cluster membership $\boldsymbol{\sigma}^*$ or its equivalent matrix representation $\mathbf{Y}^* \coloneqq \boldsymbol{\sigma}^*(\boldsymbol{\sigma}^*)^\top \in \{\pm 1\}^{n\times n}$.

Recent research has showcased that SBM possesses a rich set of properties interweaving algorithmic and information-theoretic considerations. A major challenge therein is that clustering under SBM is a discrete and hence non-convex problem. SDP relaxations have emerged as an efficient and robust approach to this problem, and recent work has witnessed the advances in establishing

rigorous performance guarantees for SDP (see Section 2 for a review of this literature). For the general problem of controlling the *estimation error* of SDP, the best and most general results to date are given in the line of work in Guédon and Vershynin (2016); Fei and Chen (2019b), which proves that the optimal SDP solution $\widehat{\mathbf{Y}}$ satisfies the bound

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*) \lesssim \frac{1}{n^2}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \lesssim \exp\left[-\frac{nI}{C}\right], \tag{1}$$

where $C > 0$ is a large constant, I is an appropriate measure of the signal-to-noise ratio, $\|\cdot\|_1$ denotes the entrywise ℓ_1 norm, and $\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*)$ denotes the fraction of nodes mis-clustered by an estimate $\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}} \in \{\pm 1\}^n$, extracted from $\widehat{\mathbf{Y}}$, of the ground-truth cluster labels $\boldsymbol{\sigma}^*$. The above result is, however, unsatisfactory due to the presence of a large multiplicative constant C in the exponent, rendering the bound fundamentally sub-optimal. In particular, the interesting regime for proving an error bound is when $nI \le 2\log n$, as otherwise SDP is already known to attain zero error. With a large C in the exponent, the result in (1) provides a rather loose, sometimes even uninformative,[1] bound in this regime. Moreover, this sub-optimality is intrinsic to the proof techniques used and cannot be avoided simply by more careful calculations.

In this paper, we establish a strictly tighter, and essentially optimal, error bound on SDP:

Theorem 1 (Informal) *Let* $I := -2\log\left[\sqrt{pq} + \sqrt{(1-p)(1-q)}\right]$. *As* $n \to \infty$, *with probability tending to one, the optimal solution* $\widehat{\mathbf{Y}}$ *of the SDP relaxation satisfies*

$$\frac{1}{n^2}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \le \exp\left[-(1-o(1))\frac{nI}{2}\right]. \tag{2}$$

Moreover, the explicit label estimate $\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}$ *computed by taking entrywise signs of the top eigenvector of* $\widehat{\mathbf{Y}}$ *satisfies*

$$\mathit{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*) \le \exp\left[-(1-o(1))\frac{nI}{2}\right]. \tag{3}$$

See Theorem 2 for the precise statement of our results as well as an explicit, non-asymptotic estimate of the $o(1)$ term. One should compare this result with the minimax lower bound in Zhang and Zhou (2016), which shows that *any* estimator $\widehat{\boldsymbol{\sigma}}$ must incur an error

$$\mathit{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \ge \exp\left[-(1+o(1))\frac{nI}{2}\right], \tag{4}$$

as the latter represents the best achievable *Bayes risk* of the problem. Therefore, the SDP relaxation achieves the optimal Bayes error, up to a vanishing second-order term.

Optimality as a surprise? The result above has come as unexpected to us, as it shows that relaxing the original problem via SDP incurs essentially no loss in terms of statistical accuracy. It is perhaps even more surprising considering the fact that the SDP is a relaxation of the *maximum likelihood estimator* (MLE) of $\boldsymbol{\sigma}^*$, and in general MLE minimizes the 0-1 loss $\mathbb{P}\{\widehat{\boldsymbol{\sigma}} \neq \boldsymbol{\sigma}^*\}$ but not $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$, the difference between $\widehat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}^*$. Our proof for Theorem 1 reveals one possible explanation of this phenomenon (see Section 1.1 for further discussion).

1. Note that $\frac{1}{n^2}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$ is trivially upper bounded by 2 since $\widehat{\mathbf{Y}}, \mathbf{Y}^* \in [-1,1]^{n\times n}$.

Exact recovery. As one illustration of the strength of Theorem 1, we note that it implies a sharp sufficient condition for SDP to recover $\boldsymbol{\sigma}^*$ *exactly*. In particular, when $nI > (2+\epsilon)\log n$ for any positive constant ϵ, the bound (3) ensures that $\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*) < \frac{1}{n}$ and hence $\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*) = 0$ (in fact, $\widehat{\mathbf{Y}} = \mathbf{Y}^*$ can be guaranteed). With the estimate $I = (1+o(1))\left(\sqrt{p}-\sqrt{q}\right)^2$, this means that SDP achieves exact recovery whenever $\left(\sqrt{p}-\sqrt{q}\right)^2 > (2+\epsilon)\frac{\log n}{n}$—a remarkable result first established in Hajek et al. (2016a); Bandeira (2018); Abbe et al. (2016) via specialized analysis—whereas it is known that exact recovery is information-theoretically impossible when $\left(\sqrt{p}-\sqrt{q}\right)^2 < 2\frac{\log n}{n}$. In fact, the non-asymptotic version of our results is strong enough to guarantee exact recovery with $\epsilon = \Theta\left(\frac{1}{\sqrt{\log n}}\right)$—a second-order refinement of existing results.

Robustness. Significantly, we show that the bound in Theorem 1 continues to hold under the so-called *monotone semirandom model* (Feige and Kilian (2001)), where an adversary is allowed to make arbitrary change to the graph in a way that apparently strengthens connections within each cluster and weakens connections between clusters. The same is true under SBM with heterogeneous edge probabilities only assumed to be $\geq p$ within clusters and $\leq q$ between clusters. While these two settings of SBM seemingly make the clustering problem easier, they in fact foil, provably, many existing algorithms, particularly those that over-exploit the specific structures of standard SBM in order to achieve tight recovery guarantees (Feige and Kilian (2001); Moitra et al. (2016)). In contrast, our results show that SDP relaxations enjoy a robustness property that is possessed by few other algorithms.

1.1. Primal-dual analysis

Key to the establishment of our results is a novel analysis that exploits both primal and dual characterizations of the SDP. To set the context, we note that the sub-optimal bound (1) in Fei and Chen (2019b); Giraud and Verzelen (2018) is established by utilizing the primal optimality of the SDP solution $\widehat{\mathbf{Y}}$. Their arguments, however, are too crude to provide a tight estimate of the multiplicative constant C in the exponent. On the other hand, work on exact recovery for SDP typically makes use of a dual analysis (Hajek et al., 2016a; Bandeira, 2018); in particular, the optimality of $\mathbf{Y}^*$ is certified by showing the existence of a corresponding dual optimal solution, often explicitly in the form of a diagonal matrix $\mathbf{D}$ with $D_{ii} = \sigma_i^* \sum_j A_{ij}\sigma_j^*$. As this "dual certificate" $\mathbf{D}$ is tied to (and constructed using) $\mathbf{Y}^*$, such a certification approach would only succeed when the SDP indeed admits $\mathbf{Y}^*$ as an optimal solution.

Here we are concerned with the setting where the optimal solution $\widehat{\mathbf{Y}}$ is different from $\mathbf{Y}^*$, and our goal is to bound their difference. As it is a priori unknown what $\widehat{\mathbf{Y}}$ should look like, we do not know which matrix to certify and how to construct its associated dual solution, rendering the above dual certification argument inapplicable. Instead, we make use of the fact that $\widehat{\mathbf{Y}}$ is feasible to the SDP and has a better primal objective value than $\mathbf{Y}^*$, that is, $\widehat{\mathbf{Y}}$ lies in the sublevel set defined by $\mathbf{Y}^*$ and the constraints of the SDP. We then characterize the diameter of this sublevel set by using, perhaps surprisingly, the dual certificate $\mathbf{D}$ of $\mathbf{Y}^*$. Our analysis is thus fundamentally different from the dual certification analysis in existing work, which only applies when the sublevel set consists of a single element $\mathbf{Y}^*$. At the same time, we make use of $\mathbf{D}$ in a crucial way to achieve an exponential improvement over previous primal analysis.

Interestingly, our analysis, and hence our error bounds as well, actually apply to *every element* of this sublevel set, not just the optimal solution $\widehat{\mathbf{Y}}$. As can be seen in our proof, this flexibility

plays an important role in establishing the aforementioned robustness results under semirandom and heterogeneous SBMs. On the other hand, however, with this level of generality we probably should not expect the second-order $o(1)$ term in our bounds to be optimal.

It is worth noting that a subsequent paper, Fei and Chen (2019a), shows similar results for $\mathbb{Z}_2$ Synchronization, Censored Block Model and SBM in a unified framework. It also provides insights on a joint majority voting mechanism of SDP.

1.2. Paper organization

In Section 2, we review related work on SBM and the SDP approach. In Section 3, we formally introduce the SBM and its variants, as well as the SDP approach used to solve the clustering problem. We present our main results in Section 4, and outline the main steps of the proof in Section 5. The paper is concluded in Section 6 with a discussion on future directions.

2. Related work

Without enumerating the large array of recent research efforts on the SBM, we restrict attention to those that study sharp performance bounds for SBM, with a particular focus on work about the SDP relaxation approach. We refer readers to the surveys Abbe (2018); Moore (2017); Li et al. (2018) for a more comprehensive review.

2.1. Optimal error rates

Most related to us is a line of work that seeks to characterize minimax optimal rates for the error $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$ under SBM and its variants. For any estimator $\widehat{\boldsymbol{\sigma}}$ under the binary symmetric SBM, Zhang and Zhou (2016) identify the aforementioned minimax lower bound (4). They also provide an exponential-time algorithm that achieves a matching upper bound (up to an $o(1)$ factor in the exponent). Much research effort focuses on revealing the minimax rates under more general settings and developing computationally feasible algorithms (Gao et al., 2017, 2018; Xu et al., 2017; Yun and Proutiere, 2014, 2016; Zhang and Zhou, 2017). The monograph Gao and Ma (2018) provides a review on recent work on this front. Concurrently to our work, Zhou and Li (2018) prove a refined non-asymptotic minimax lower bound and provide a polynomial-time algorithm that attains a matching upper bound. We note that this line of work does not consider the SDP relaxation approach nor deliver robustness guarantees as we do. Nevertheless, we will compare our results with theirs after stating our main theorems.

2.2. Exact recovery

The special case of achieving exact recovery, namely $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = 0$, is the focus of a large volume of recent work. In the case of the binary symmetric SBM with $p, q \asymp \frac{\log n}{n}$, Abbe et al. (2016) and Mossel et al. (2016) prove that exact recovery is information-theoretically possible if and only if $(\sqrt{p} - \sqrt{q})^2 > \frac{2 \log n}{n}$. Much research endeavor has since been devoted to identifying similar thresholds for more general SBMs and designing algorithms that succeed in exact recovery above optimal thresholds; see, e.g., Abbe et al. (2014); Abbe and Sandon (2015a,c); Abbe et al. (2017); Agarwal et al. (2017); Jog and Loh (2015); Perry and Wein (2015). As mentioned, our results imply sharp bounds for exact recovery as a corollary.

2.3. Weak recovery

Complementary to exact recovery is the notion of weak (non-trivial) recovery, that is, achieving an error $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$ that is better than random guess. For the binary symmetric SBM in the sparse regime $p, q \asymp \frac{1}{n}$, work of Lelarge et al. (2015); Massoulié (2014); Mossel et al. (2015) establishes that the necessary and sufficient condition of weak recovery is $\frac{n(p-q)^2}{p+q} > 2$. Subsequent work proves similar phase transitions and show that various algorithms achieve weak recovery above the optimal threshold for the SBM with $k \geq 2$ and possibly unbalanced clusters; see, e.g., Abbe and Sandon (2015b); Abbe et al. (2018); Banerjee (2018); Bordenave et al. (2018); Caltagirone et al. (2018); Coja-Oghlan et al. (2018); Montanari and Sen (2016); Mossel et al. (2018); Stephan and Massoulié (2018). As discussed later, our results also imply weak recovery guarantees with a sub-optimal constant.

2.4. Optimality and robustness of SDP

SDP has been proven to succeed in exact and weak recovery above corresponding optimal thresholds (sometimes under additional assumptions). In particular, see Agarwal et al. (2017); Bandeira (2018); Hajek et al. (2015, 2016a,b) for exact recovery, and Montanari and Sen (2016) for weak recovery. Prior to our work, however, SDP was not known to achieve the optimal error rate between the exact and weak recovery regimes. Sub-optimal polynomial rates are first proved in Guédon and Vershynin (2016), later improved to exponential in Fei and Chen (2019b) and further generalized in Fei and Chen (2018); Giraud and Verzelen (2018).

Robustness has been recognized as a distinct feature of the SDP approach as compared to other more specialized algorithms for SBMs. Work in this direction has established robustness of SDP against random erasures Hajek et al. (2015, 2016b), atypical node degrees Guédon and Vershynin (2016) and adversarial corruptions Hajek et al. (2016a); Montanari and Sen (2016); Cai and Li (2015); Makarychev et al. (2016). Moitra et al. (2016) investigate the relationship between statistical optimality and robustness under monotone semirandom models; we revisit this work in more details later.

3. Problem Set-up

In this section, we formally define the binary symmetric SBM and introduce the SDP approach.

3.1. Notations

Vectors and matrices are denoted by bold letters. For a vector $\mathbf{u}$, we let u_i be its i-th entry. For a matrix $\mathbf{M}$, we let M_{ij} denote its (i,j)-th entry, $\|\mathbf{M}\|_1 \coloneqq \sum_{i,j} |M_{ij}|$ its entry-wise ℓ_1 norm, $\|\mathbf{M}\|_F \coloneqq \sqrt{\sum_{i,j} M_{ij}^2}$ its Frobenius norm, and $\|\mathbf{M}\|_{\mathrm{op}}$ its spectral norm (the maximum singular value). We write $\mathbf{M} \succeq 0$ if $\mathbf{M}$ is symmetric positive semidefinite. The trace inner product between two matrices is $\langle \mathbf{M}, \mathbf{G} \rangle \coloneqq \sum_{i,j} M_{ij} G_{ij} = \mathrm{Tr}(\mathbf{M}^\top \mathbf{G})$. We denote by $\mathbf{I}$ and $\mathbf{J}$ the $n \times n$ identity matrix and all-one matrix, respectively, and let $\mathbf{1}$ be the all-one column vector of length n.

$\mathrm{Bern}(\mu)$ denotes the Bernoulli distribution with mean $\mu \in [0,1]$. For a positive integer i, let $[i] \coloneqq \{1, 2, \ldots, i\}$. For a real number x, $\lceil x \rceil$ denotes its ceiling and $\lfloor x \rfloor$ denotes its floor. $\mathbb{I}\{\cdot\}$ is the indicator function. For two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$, $b_n = \Omega(a_n)$ or $a_n \lesssim b_n$ if there exists a universal constant $C > 0$ such that $a_n \leq C b_n$ for all n. We

write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. We are sometimes concerned with the asymptotic regime $n \to \infty$, in which case we write $a_n = o(b_n)$ or $b_n = \omega(a_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

3.2. The Stochastic Block Models

We begin with formally describing our basic model.

Model 1 (Binary symmetric SBM) *Suppose that the ground-truth clustering $\boldsymbol{\sigma}^* \in \{\pm 1\}^n$ satisfies $\langle \boldsymbol{\sigma}^*, \mathbf{1}\rangle = 0$. The graph adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ is symmetric with $A_{ii} = 0$ for $i \in [n]$ and its entries $\{A_{ij}, i < j\}$ generated independently by*

$$A_{ij} \sim \begin{cases} \mathrm{Bern}(p) & \text{if } \sigma_i^*\sigma_j^* = 1, \\ \mathrm{Bern}(q) & \text{if } \sigma_i^*\sigma_j^* = -1, \end{cases}$$

where $0 \le q < p \le 1$ are allowed to scale with n.

As the cluster labels $\boldsymbol{\sigma}^*$ are assumed to contain the same number of 1's and -1's, the binary symmetric SBM represents two clusters with equal size. Despite its simple form, this model has been of central importance in studying fundamental limits of clustering problems (Abbe et al., 2016; Hajek et al., 2016a; Lelarge et al., 2015; Massoulié, 2014; Montanari and Sen, 2016; Mossel et al., 2015, 2016).

For the purpose of studying the robustness properties of SDP relaxation, we consider two extensions of the above basic model. The first one is a semirandom model where a so-called monotone adversary, upon observing the random adjacency matrix $\mathbf{A}$ generated from Model 1 and the ground-truth clustering $\boldsymbol{\sigma}^*$, modifies $\mathbf{A}$ by *arbitrarily* adding edges between nodes of the same cluster and deleting edges between nodes of different clusters.

Model 2 (Semirandom SBM) *A monotone adversary observes $\mathbf{A}$ and $\boldsymbol{\sigma}^*$ from Model 1, picks an arbitrary set of pairs of nodes $\mathcal{L} \subset \{(i,j) \in [n]\times[n] : i < j\}$, and outputs a matrix $\mathbf{A}^{SR}$ such that for each $i, j \in [n]$,*

$$A_{ij}^{SR} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{L},\ \sigma_i^*\sigma_j^* = 1, \\ 0 & \text{if } (i,j) \in \mathcal{L},\ \sigma_i^*\sigma_j^* = -1, \\ A_{ij}, & \text{if } (i,j) \notin \mathcal{L}. \end{cases}$$

Note that the set $\mathcal{L}$ is allowed to depend on the realization of $\mathbf{A}$. Setting $\mathcal{L} = \emptyset$ in Model 2 recovers the basic Model 1. Semirandom models have a long history with many variants (Blum and Spencer, 1995). Model 2 above has been considered in Feige and Kilian (2001); Moitra et al. (2016) for SBM. While seemingly revealing more information about the underlying cluster structure, the semirandom model in fact destroys many local structures of the basic SBM, thus frustrating many algorithms that over-exploit such structures. In contrast, SDP is robust against the monotone adversary under Model 2, as we shall see in Section 4.2 below.

Another way to extend Model 1 is by allowing the edge probabilities to vary across node pairs:

Model 3 (Heterogeneous SBM) *Suppose that the ground-truth clustering $\boldsymbol{\sigma}^* \in \{\pm 1\}^n$ satisfies $\langle \boldsymbol{\sigma}^*, \mathbf{1}\rangle = 0$. The graph adjacency matrix $\mathbf{A}^H \in \{0,1\}^{n\times n}$ is symmetric with $A_{ii}^H = 0$ for $i \in [n]$*

and its entries $\{A^H_{ij}, i<j\}$ *generated independently by*

$$A^H_{ij} \sim \begin{cases} \mathrm{Bern}(p_{ij}) & \text{if } \sigma^*_i\sigma^*_j = 1, \\ \mathrm{Bern}(q_{ij}) & \text{if } \sigma^*_i\sigma^*_j = -1, \end{cases}$$

where $0 \le q_{ij} \le q < p \le p_{ij} \le 1$ *and all* $p, q, \{p_{ij}\}, \{q_{ij}\}$ *are allowed to scale with* n.

Clearly, Model 1 is a special case of Model 3 with all $\{p_{ij}\}$ equal to p and all $\{q_{ij}\}$ equal to q. For all three models, we define the following measure of the signal-to-noise ratio (SNR):

$$I := -2\log\left[\sqrt{pq} + \sqrt{(1-p)(1-q)}\right]. \tag{5}$$

Note that I is the Renyi divergence of order $\frac{1}{2}$ between $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$. As we shall see very soon, the quantity I dictates the minimax error rates. Let us also introduce the following measure of the distance between two vectors of cluster labels $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \{\pm 1\}^n$:

$$\mathtt{err}(\boldsymbol{\sigma}, \boldsymbol{\sigma}') := \min_{g\in\{\pm1\}} \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{g\sigma_i = \sigma'_i\}.$$

In words, $\mathtt{err}(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ is the fractions of nodes that are assigned a different label under $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$, modulo a global flipping of signs. With $\boldsymbol{\sigma}^*$ being the true labels, $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$ measures the relative error of the estimator $\widehat{\boldsymbol{\sigma}}$.

3.3. SDP relaxation

To motivate the SDP relaxation we are to consider, we first note that the MLE of $\boldsymbol{\sigma}^*$ under the binary symmetric SBM (Model 1) is given by the solution of the following discrete and non-convex optimization problem

$$\begin{aligned} \max_{\boldsymbol{\sigma}\in\{\pm1\}^n} \ & \left\langle \mathbf{A}, \boldsymbol{\sigma}\boldsymbol{\sigma}^\top \right\rangle \\ \text{s.t. } & \langle \boldsymbol{\sigma}, \mathbf{1}\rangle = 0. \end{aligned}$$

Derivation of the MLE in this form is now standard; see for example Li et al. (2018). We define the lifted variable $\mathbf{Y} = \boldsymbol{\sigma}\boldsymbol{\sigma}^\top$, and observe that $\mathbf{Y}$ satisfies $\mathbf{Y} \succeq 0$, $Y_{ii} = (\sigma_i)^2 = 1$ for $i \in [n]$ as well as $\langle \mathbf{J}, \mathbf{Y}\rangle = \left\langle \mathbf{1}\mathbf{1}^\top, \boldsymbol{\sigma}\boldsymbol{\sigma}^\top \right\rangle = \langle \boldsymbol{\sigma}, \mathbf{1}\rangle^2 = 0$. Dropping the constraints that $\mathbf{Y}$ has rank one and binary entries, we obtain the following SDP relaxation:

$$\begin{aligned} \widehat{\mathbf{Y}} = \mathop{\arg\max}_{\mathbf{Y}\in\mathbb{R}^{n\times n}} \ & \langle \mathbf{A}, \mathbf{Y}\rangle \\ \text{s.t. } & \mathbf{Y} \succeq 0, \\ & Y_{ii} = 1, \quad \forall i \in [n], \\ & \langle \mathbf{J}, \mathbf{Y}\rangle = 0. \end{aligned} \tag{6}$$

The optimization problem (6) is a standard SDP that can be solved in polynomial time. We remark that this SDP does not require any information about edge probabilities p and q. We note that the same SDP is also considered in Hajek et al. (2016a) for studying the optimal exact recovery

threshold by SDP under the binary symmetric SBM. It can be further traced back to the work of Feige and Kilian (2001), which studies SDP for MIN BISECTION.

In the case where we are given the monotonically modified adjacency matrix $\mathbf{A}^{\mathrm{SR}}$ generated by the semirandom SBM (Model 2), we solve the same SDP relaxation (6) with the input $\mathbf{A}$ replaced by $\mathbf{A}^{\mathrm{SR}}$. Similarly, under the Heterogeneous SBM (Model 3), we solve the SDP with $\mathbf{A}$ replaced by $\mathbf{A}^{\mathrm{H}}$.

We consider $\widehat{\mathbf{Y}}$ as an estimate of the ground-truth matrix $\mathbf{Y}^*$, and seek to characterize the accuracy of the SDP solution $\widehat{\mathbf{Y}}$ in terms of the ℓ_1 error $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$. Note that $\widehat{\mathbf{Y}}$ is not necessarily a rank one matrix of the form $\widehat{\mathbf{Y}} = \boldsymbol{\sigma}\boldsymbol{\sigma}^\top$. To extract from $\widehat{\mathbf{Y}}$ a vector of binary estimates of cluster labels, we take the signs of the entries of the top eigenvector of $\widehat{\mathbf{Y}}$ (where the sign of 0 is 1, an arbitrary choice). Letting $\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}} \in \{\pm 1\}^n$ be the vector obtained in this way, we study the error of $\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}$ as an estimate of the ground-truth label vector $\boldsymbol{\sigma}^*$, as measured by $\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*)$.

4. Main results

In this section, we provide our main results on the error rate of the SDP relaxation (6).

As mentioned, our results for the binary symmetric SBM (Model 1) in fact hold for any element in the following sublevel set (or superlevel set to be precise):

$$\mathcal{Y}(\mathbf{A}) \coloneqq \left\{\mathbf{Y} \in \mathbb{R}^{n\times n} : \langle \mathbf{A}, \mathbf{Y}\rangle \geq \langle \mathbf{A}, \mathbf{Y}^*\rangle, \mathbf{Y} \text{ is feasible to the program (6)}\right\}, \tag{7}$$

i.e., the set of feasible solutions to the SDP (6) that attain an objective value no worse than the ground-truth $\mathbf{Y}^*$. Clearly, the optimal solution to the SDP belongs to $\mathcal{Y}(\mathbf{A})$, since $\mathbf{Y}^*$ is feasible to the SDP. For the semirandom and heterogeneous SBMs (Models 2 and 3), we consider the sets $\mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$ and $\mathcal{Y}(\mathbf{A}^{\mathrm{H}})$, respectively, defined in a similar way. With a slight abuse of notation, in the sequel we shall use $\widehat{\mathbf{Y}}$ to denote an arbitrary matrix in the set $\mathcal{Y}(\mathbf{A})$, $\mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$ or $\mathcal{Y}(\mathbf{A}^{\mathrm{H}})$; accordingly, we let $\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}$ denote the corresponding vector of labels extracted from this $\widehat{\mathbf{Y}}$.

4.1. Error rates under Binary Symmetric SBM

Our first theorem is a non-asymptotic bound on the error rates of the SDP relaxation (6) under the binary symmetric SBM.

Theorem 2 (Binary Symmetric SBM) *Under Model 1 and assuming $0 < c_0 p \leq q < p \leq 1 - c_1$ for some constants $c_0, c_1 \in (0,1)$, there exist some constants $C_I, C_e, C > 0$ such that the following holds: If $nI \geq C_I$, then there hold the bounds*

$$\begin{aligned} &\frac{1}{n}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \leq \left\lfloor n \exp\left[-\left(1 - C_e\sqrt{\frac{1}{nI}}\right)\frac{nI}{2}\right]\right\rfloor, \\ &\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*) \leq \exp\left[-\left(1 - C_e\sqrt{\frac{1}{nI}}\right)\frac{nI}{2}\right], \end{aligned} \qquad \forall \widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A}),$$

with probability at least $1 - 2(e/2)^{-2n} - 3\exp\left(-\sqrt{\log n}\right) - \frac{1}{n^2} - \frac{2}{\sqrt{n}}$.

Remark 1 *The assumption $c_0 p \leq q$ arises from a technical step in our proof. We note that this assumption is common in the literature on minimax rates (Gao et al., 2017, 2018; Zhang and Zhou, 2017; Zhou and Li, 2018). In Section 5.1 in Gao et al. (2017), a weaker minimax upper bound is obtained with this assumption dropped.*

In Section 5 we outline the main steps of the proof of the above theorem, deferring the details to Appendices B and D. Note the floor operation in the first inequality above, which implies that $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 = 0$ whenever the exponent is strictly less than $-\log n$; later we explore its implication for exact recovery.

Letting $n \to \infty$ in Theorem 2, we immediately obtain the following asymptotic result:

Corollary 1 (Binary Symmetric SBM, Asymptotic) *Under Model 1 with $n \to \infty$ and assuming $0 < c_0 p \le q < p \le 1 - c_1$ for some constants $c_0, c_1 \in (0,1)$, if $nI \to \infty$, then there hold the bounds*

$$\frac{1}{n}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \le \left\lfloor n \exp\left[-(1-o(1))\frac{nI}{2}\right]\right\rfloor, \quad \forall \widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A}),$$
$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\mathrm{sdp}}, \boldsymbol{\sigma}^*) \le \exp\left[-(1-o(1))\frac{nI}{2}\right],$$

with probability $1 - o(1)$.

The results in Theorem 2 and Corollary 1 can be compared with an existing minimax lower bound on the error rate for SBM. Denote by $\widehat{\boldsymbol{\sigma}}$ an arbitrary estimator (that is, a measurable function of $\mathbf{A}$) of ground-truth labels $\boldsymbol{\sigma}^*$. Theorem 1.1 in Zhang and Zhou (2016) implies that under Model 1 with $nI \to \infty$, any $\widehat{\boldsymbol{\sigma}}$ must satisfy the lower bound

$$\mathbb{E}\,\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \ge \exp\left[-(1+o(1))\frac{nI}{2}\right].$$

In view of this lower bound and the upper bound in Corollary 1, we see that the SDP achieves the optimal error rate, up to a vanishing second-order term in the exponent.[2]

Our Theorem 2 in fact provides an explicit, non-asymptotic upper bound for the $o(1)$ term in the exponent. This bound, taking the form of $O(1/\sqrt{nI})$, yields second-order characterization of various recovery thresholds and is strong enough to provide non-trivial guarantees in the sparse graph regime. We discuss these points in Section 4.3 to follow. We do not expect this bound to be information-theoretic optimal though, for reasons discussed in the Introduction section.

4.2. Robustness under Semirandom and Heterogeneous SBMs

The following theorem shows that the error rate of the SDP is unaffected by passing to the semirandom model (Model 2). Recall that $\mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$ is the sublevel set of the SDP (6) with $\mathbf{A}^{\mathrm{SR}}$ as the input.

Theorem 3 (Semirandom SBM) *Suppose that $\mathbf{A}^{SR}$ is generated according to Model 2. The conclusions of Theorem 2 and Corollary 1 continue to hold with $\mathcal{Y}(\mathbf{A})$ replaced by $\mathcal{Y}(\mathbf{A}^{SR})$.*

A similar result holds for the heterogeneous model (Model 3). In fact, Model 3 can be reduced to Model 2 as shown in Fei and Chen (2019b, Appendix V). Recall that $\mathcal{Y}(\mathbf{A}^{\mathrm{H}})$ is the sublevel set of the SDP (6) with $\mathbf{A}^{\mathrm{H}}$ as the input.

2. We note that Theorem 1.1 in Zhang and Zhou (2016) bounds the error in expectation and holds for a parameter space containing $\boldsymbol{\sigma}^*$ with slightly unequal-sized clusters. Our results provide a high-probability bound. Extending our results to the setting of slightly unequal-sized clusters is possible, albeit tedious; we leave this to future work.

Theorem 4 (Heterogeneous SBM) *Suppose that $\mathbf{A}^H$ is generated according to Model 3. The conclusions of Theorem 2 and Corollary 1 continue to hold with $\mathcal{Y}(\mathbf{A})$ replaced by $\mathcal{Y}(\mathbf{A}^H)$.*

The proofs of the two theorems above are both given in Appendix E.

The results above show that SDP is insensitive to monotone modification and heterogeneous probabilities. We emphasize that such robustness is by no means automatic. With non-uniformity in the probabilities, the likelihood function no longer has a known, rigid form, a property heavily utilized in many algorithms. The monotone adversary can similarly alter the graph structure by creating hotspots and short cycles. Even worse, the adversary is allowed to make changes *after* observing the realized graph,[3] thus producing unspecified dependency among all edges in the observed data and leading to major obstacles for existing analysis of iterative algorithms.

We would like to mention that the work in Moitra et al. (2016) shows that the semirandom model makes weak recovery strictly harder. While not contradicting their results technically, the fact that our error bounds remain unaffected under this model does demand a closer look. We note that our bounds are optimal only up to a second order term in the exponent and consequently do not attain the optimal weak recovery limit. Also, our robustness results on error rates are tied to a specific form of SDP analysis (using the sublevel set $\mathcal{Y}(\mathbf{A})$). In comparison, for exact recovery SDP is robust *by design* to the semirandom model, as is well recognized in past work (Feige and Kilian, 2001; Chen et al., 2014; Hajek et al., 2016a).

In the following, we discuss some consequences of the results presented above, and compare them with results in other work that derives minimax rate.

4.3. Consequences and Optimality

Theorem 2 and Corollary 1 imply sharp sufficient conditions for several types of recovery.

Exact recovery: Noting the equivalence $I = (1+o(1))(\sqrt{p}-\sqrt{q})^2$ valid for $0 < q \asymp p = o(1)$, we see that whenever $n(\sqrt{p}-\sqrt{q})^2 \geq (2+\epsilon)\log n$ for any constant $\epsilon > 0$, we have $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 = 0$ by Corollary 1 (note the floor operation therein) and hence SDP achieves exactly recovery by itself. We thus recover, as a corollary, the sharp exact recovery threshold for SDP established in Hajek et al. (2016a); Bandeira (2018).

Second-order refinement: Using the non-asymptotic Theorem 2, we can obtain the following refinement of the above result: exact recovery provided $\frac{nI}{\log n} \geq 2 + \frac{C_1}{\sqrt{\log n}} + \frac{C_2}{\log n}$ for some constants $C_1, C_2 > 0$. This result is comparable to the sufficient condition $\frac{n(\sqrt{p}-\sqrt{q})^2}{\log n} \geq 2 + \frac{C}{\sqrt{\log n}} + \omega\left(\frac{1}{\log n}\right)$ for SDP established in Hajek et al. (2016a), whereas the *necessary and sufficient* condition for optimal estimator (MLE) is $\frac{n(\sqrt{p}-\sqrt{q})^2}{\log n} \geq 2 - \frac{\log\log n}{\log n} + \omega\left(\frac{1}{\log n}\right)$ (Abbe et al., 2016; Mossel et al., 2016).

Almost exact recovery: Theorem 2 ensures that $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = o(1)$ under the condition $nI \to \infty$. This condition is optimal, as proved in Mossel et al. (2016); Abbe and Sandon (2015a).

Weak recovery: When $nI \geq C$ for a sufficiently large constant C, Theorem 2 ensures that $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) < \frac{1}{2}$ and hence SDP achieves weak recovery. This condition matches, up to constants, the so-called Kesten-Stigum (KS) threshold $\frac{n(p-q)^2}{p+q} > 2$ (in view of $I \asymp \frac{(p-q)^2}{p}$; cf. Fact 5), which is optimal (Massoulié, 2014; Abbe and Sandon, 2015b; Mossel et al., 2013, 2015).

3. In this sense we have strengthened the robustness results in the previous work Fei and Chen (2019b), which does not allow such adaptivity.

Sparse regime: Our result guarantees an arbitrarily small constant error when nI is a sufficiently large but finite constant. This result is applicable even in the sparse graph regime with constant expected degrees, namely $np, nq, nI = \Theta(1)$. In comparison, many results on minimax rates require nI, and hence the degrees, to diverge (e.g., Gao et al. (2017); Zhang and Zhou (2017)).

4.4. Comparison with existing results

We compare with the existing work that derives sharp error rate bounds achievable by polynomial-time algorithms. Here we focus on the binary symmetric SBM (Model 1). To be clear, the algorithms considered in this line of work are very different from ours. In particular, most existing results and particularly those discussed below, require a good enough initial estimate of the true clusters. Obtaining such an initial solution (typically done using spectral clustering) is itself a non-trivial task.

Using neighbor voting and variational inference algorithms, the work in Gao et al. (2017) and Zhang and Zhou (2017) obtains an error bound of the same form as our Corollary 1, under the assumption that $0 < q < p < 1$, $p \asymp q$ and $nI \to \infty$ (Zhang and Zhou (2017) assume additionally $p = o(1)$). Yun and Proutiere (2014) consider a spectral algorithm and prove the error bound

$$\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \le \exp\left[-(1-\epsilon)\frac{n(\sqrt{p}-\sqrt{q})^2}{2}\right]$$

for any constant $\epsilon > 0$, under the conditions $np \to \infty$ and $(1+\varepsilon)q \le p = o\left(1/\log^2 n\right)$. Recalling $I = (1+o(1))(\sqrt{p}-\sqrt{q})^2$, we find that our error rate in Corollary 1 is better as we allow the ϵ term to vanish. A strength of our results is that we provide an explicit bound for the second-order term in the exponent; we know of few error rate results (with the exception discussed below) that offer this level of accuracy.

Concurrently to our work, Zhou and Li (2018) show that an EM-type algorithm achieves a non-asymptotic error rate of the form

$$\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \le \exp\left[-\left(1+\frac{2}{nI}\log\frac{\sqrt{np}}{C}\right)\frac{nI}{2}\right]$$

for some constant $C > 0$, under the conditions $1 \lesssim \sqrt{np} \lesssim nI$ and $q \asymp p$. This bound has a better second-order term in the exponent compared to our Theorem 2. We do note that their algorithm is fairly technical, involving data partition and the leave-one-out tricks to ensure independence, degree truncation to regularize spectral clustering, and blackbox solvers for K-means and matching problems. In comparison, the SDP approach is much simpler conceptually.

Finally, we emphasize that we also provide robustness guarantees under the monotone semirandom model and non-uniform edge probabilities. In comparison, it is unclear if comparable robustness results can be established for the work above, as their algorithms and analysis make substantial use of the properties of the standard SBM, particularly the complete independence among edges and the specific form of the likelihood function.

5. Proof Outline of Theorem 2

To prove Theorem 2, we proceed in three steps:

Step 1: As mentioned in Section 1, we construct a diagonal matrix $\mathbf{D} := \mathrm{diag}\,[\boldsymbol{\sigma}^* \circ ((\mathbf{A} - \lambda^*(\mathbf{J} - \mathbf{I}))\boldsymbol{\sigma}^*)]$, which takes the same form as the "dual certificate" used in previous work (the parameter λ^* is introduced only for technical reasons later and plays no role in this step). The construction of $\mathbf{D}$ allows us to establish the *basic inequality*:

$$0 \le \left\langle -\mathbf{D}, \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}) \right\rangle + \left\langle \mathbf{A} - \mathbb{E}\mathbf{A}, \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}) \right\rangle, \quad \text{for any } \widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A});$$

see the proof of Lemma 1 for the details of this critical step. Here $\mathcal{P}_{T^\perp}$ is an appropriate projection operator that satisfies $\mathrm{Tr}\left[\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right] = \frac{1}{n}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$, thus exposing the ℓ_1 error of $\widehat{\mathbf{Y}}$ that we seek to control.

Step 2: We then show that the second term $\langle \mathbf{A} - \mathbb{E}\mathbf{A}, \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\rangle$ in the basic inequality is negligible compared to the first term $\langle -\mathbf{D}, \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\rangle$. Consequently, the problem boils down to studying the inequality

$$0 \le \langle -\mathbf{D}, \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\rangle = \sum_{i\in[n]} (-D_{ii}) b_i,$$

where $b_i := \left(\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right)_{ii}$ satisfies $\sum_{i\in[n]} b_i = \frac{1}{n}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$.

Step 3: The b_i's take fractional values in general. Crucially, we argue that the bounding the fractional sum $\sum_i (-D_{ii}) b_i$ can be reduced to controlling the discrete sum $\sum_{i\in M}(-D_{ii})$ for appropriate subsets M of $[n]$ with $|M| \approx \frac{1}{n}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$. We then observe that each D_{ii} is roughly the difference between $\frac{n}{2}\,\mathrm{Bern}(q)$ random variables and $\frac{n}{2}\,\mathrm{Bern}(p)$ random variables. Applying a tight Chernoff inequality, in which I arises as the rate function, we show that

$$\sum_{i\in[n]} (-D_{ii}) b_i \lesssim \log\left[n \Big/ \Big(\frac{1}{n}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1\Big)\right] - \big(1 - o(1)\big)\frac{nI}{2}.$$

Combining the above two inequalities and inverting the logarithm yield the exponential error bound in the first inequality in Theorem 2.

Once we establish the first inequality in Theorem 2, the second inequality follows from a straightforward application of Davis-Kahan theorem. The details of the proof of these inequalities are given in Appendices B and D.

6. Discussion

In this paper, we have analyzed the error rates of the SDP relaxation approach for clustering under the binary symmetric SBM. We have shown that SDP achieves an exponentially-decaying error with a sharp constant, matching the minimax lower bound. As an immediate consequence, SDP achieves exact recovery above the optimal threshold. We have also shown that these results continue to hold under monotone semirandom models and non-uniform edge probabilities.

Interesting future directions include extensions to SBM with multiple and unbalanced clusters, as well as to closely related models such as the weighted and degree-corrected SBMs. It is also of interest to see if better estimates of the second order term can be obtained, and if there is a fundamental tradeoff between statistical optimality and robustness.

Acknowledgments

Y. Fei and Y. Chen were partially supported by the National Science Foundation CRII award 1657420 and grant 1704828.

References

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 670–688. IEEE, 2015a.

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

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

Emmanuel Abbe, Afonso S. Bandeira, Annina Bracher, and Amit Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

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

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Emmanuel Abbe, Enric Boix, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *arXiv preprint arXiv:1809.04818*, 2018.

Naman Agarwal, Afonso S. Bandeira, Konstantinos Koiliaris, and Alexandra Kolla. Multisection in the stochastic block model using semidefinite programming. In *Compressed Sensing and its Applications*, pages 125–162. Springer, 2017.

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

Debapratim Banerjee. Contiguity and non-reconstruction results for planted partition models: the dense case. *Electronic Journal of Probability*, 23, 2018.

Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Nonbacktracking spectrum of random graphs: Community detection and nonregular ramanujan graphs. *Annals of Probability*, 46(1): 1–71, 2018.

T. Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Annals of Statistics*, 43(3):1027–1059, 2015. URL http://arxiv.org/abs/1404.6000.

Francesco Caltagirone, Marc Lelarge, and Léo Miolane. Recovering asymmetric communities in the stochastic block model. *IEEE Transactions on Network Science and Engineering*, 5(3):237–246, 2018.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014.

Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborova. Information-theoretic thresholds from the cavity method. *Advances in Mathematics*, 333:694–795, 2018.

Yingjie Fei and Yudong Chen. Hidden integrality of SDP relaxation for sub-gaussian mixture models. *arXiv preprint arXiv:1803.06510*, 2018.

Yingjie Fei and Yudong Chen. Achieving the bayes error rate in synchronization and block models by SDP, robustly. *arXiv preprint arXiv:1904.09635*, 2019a.

Yingjie Fei and Yudong Chen. Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Transactions on Information Theory*, 65(1):551–571, 2019b.

Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

Chao Gao and Zongming Ma. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *arXiv preprint arXiv:1811.06055*, 2018.

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

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k means. *arXiv preprint arXiv:1807.07547*, 2018.

Alexander Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Resenhas do Instituto de Matemática e Estatistica da Universidade de São Paulo*, 2(4):401–481, 1953.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Exact recovery threshold in the binary censored block model. In *IEEE Information Theory Workshop (ITW)*, pages 99–103, 2015.

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

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, 2016b.

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

Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labelled stochastic block model. *IEEE Transactions on Network Science and Engineering*, 2(4):152–163, 2015.

Xiaodong Li, Yudong Chen, and Jiaming Xu. Convex relaxation methods for community detection. *arXiv preprint arXiv:1810.00315*, 2018.

Joram Lindenstrauss and Aleksander Pełczyński. Absolutely summing operators in L_p-spaces and their applications. *Studia Mathematica*, 3(29):275–326, 1968.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Learning communities in the presence of errors. In *29th Annual Conference on Learning Theory*, pages 1258–1291, 2016.

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

Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 828–841. ACM, 2016.

Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 814–827, 2016.

Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *Bulletin of European Association for Theoretical Computer Science (EATCS)*, 1(121), Februrary 2017.

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

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. *Electronic Journal of Probability*, 21(21):1–24, 2016. doi: 10.1214/16-EJP4185. URL http://dx.doi.org/10.1214/16-EJP4185.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

William Perry and Alexander S. Wein. A semidefinite program for unbalanced multisection in the stochastic block model. *arXiv preprint arXiv:1507.05605*, 2015.

Elizaveta Rebrova and Roman Vershynin. Norms of random matrices: local and global problems. *arXiv preprint arXiv:1608.06953*, 2016.

Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. *arXiv preprint arXiv:1811.05808*, 2018.

Van Vu. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4): 526–538, 2011.

Min Xu, Varun Jog, and Po-Ling Loh. Optimal rates for community estimation in the weighted stochastic block model. *arXiv preprint arXiv:1706.01175*, 2017.

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

Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

Anderson Y. Zhang and Harrison H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*, 2017.

Zhixin Zhou and Ping Li. Non-asymptotic chernoff lower bound and its application to community detection in stochastic block model. *arXiv preprint arXiv:1812.11269*, 2018.

Appendix A. Additional notations

We introduce some additional notations for subsequent proofs. For a matrix $\mathbf{M}$, we let $\mathbf{M}^\top$ denote the transpose of $\mathbf{M}$, $\mathrm{Tr}(\mathbf{M})$ its trace, $\|\mathbf{M}\|_\infty \coloneqq \max_{i,j} |M_{ij}|$ its entrywise ℓ_∞ norm, and $\mathrm{diag}\,(\mathbf{M})$ the vector of its diagonal entries. With another matrix $\mathbf{G}$ of the same shape as $\mathbf{M}$, we use $\mathbf{M} \geq \mathbf{G}$ to mean that $M_{ij} \geq G_{ij}$ for all i, j.

Appendix B. Proof of the first inequality in Theorem 2

Here we prove the first inequality in Theorem 2. The proof of the second inequality is given in Section D

Consider any $\widehat{\mathbf{Y}}$ from the set $\mathcal{Y}(\mathbf{A})$. Let us define $\gamma \coloneqq \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$, the key quantity of our interest. Let t^* be the minimizer of $t \mapsto \mathbb{E}e^{t(X-Y)}$ where $X \sim \mathrm{Bern}(q)$, $Y \sim \mathrm{Bern}(p)$ and X and Y are independent. It can be shown that

$$t^* = \frac{1}{2} \log \frac{p(1-q)}{q(1-p)}. \tag{8}$$

Since we assume $p > q$, we have $t^* > 0$. Let

$$\lambda^* \coloneqq \frac{1}{2t^*} \log \frac{1-q}{1-p}. \tag{9}$$

Let $\mathbf{A}^0 \coloneqq \mathbf{A} - \lambda^*(\mathbf{J} - \mathbf{I})$ and $\mathbf{W} \coloneqq \mathbf{A} - \mathbb{E}\mathbf{A}$. Define an $n \times n$ diagonal matrix

$$\mathbf{D} \coloneqq \begin{bmatrix} D_{11} & & \\ & \ddots & \\ & & D_{nn} \end{bmatrix}$$

such that for each $i \in [n]$, $D_{ii} \coloneqq \sum_{j\in[n]} A^0_{ij} Y^*_{ij} = \sigma^*_i \sum_{j\in[n]} A^0_{ij}\sigma^*_j$. Let $\mathbf{d} \coloneqq [D_{11}, \ldots, D_{nn}]^\top$. Let $\mathbf{U} \in \mathbb{R}^n$ be a vector of the left singular vector of $\mathbf{Y}^*$ and it can be seen that $\mathbf{U} = \frac{1}{\sqrt{n}}\boldsymbol{\sigma}^*$. Define the projection $\mathcal{P}_T(\mathbf{M}) \coloneqq \mathbf{U}\mathbf{U}^\top\mathbf{M} + \mathbf{M}\mathbf{U}\mathbf{U}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{M}\mathbf{U}\mathbf{U}^\top$ and $\mathcal{P}_{T^\perp}(\mathbf{M}) = \mathbf{M} - \mathcal{P}_T(\mathbf{M})$ for any $\mathbf{M} \in \mathbb{R}^{n\times n}$.

B.1. Establishing basic inequality

Let us record some facts about $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A})$ and parameters t^*, λ^*, I that will be useful for the subsequent proof.

Fact 1 *We have* $\left\langle \mathbb{E}\mathbf{A} - \lambda^*(\mathbf{J} - \mathbf{I}), \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle = -(p - \lambda^*)\,\mathrm{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right)$.

The proof is given in Section C.1.

Fact 2 $\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \succeq 0$ *and* $\mathrm{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) = \frac{\gamma}{n}$.

The proof is given in Section C.2.

Fact 3 *We have* $\|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_\infty \le 4$.

The proof is given in Section C.3.

Fact 4 *If* $0 < c_0 p \le q < p \le 1 - c_1$ *for some constants* $c_0, c_1 \in (0,1)$, *then there exists some constant* $C > 0$ *such that* $t^* \le C\frac{p-q}{p}$.

The proof is given in Section C.4.

Fact 5 *If* $0 < q < p \le 1 - c$ *for some constant* $c \in (0,1)$, *then* $I \asymp \frac{(p-q)^2}{p}$.

This is a partial result of Lemma B.1 of Zhang and Zhou (2016).

Fact 6 *If* $0 < q < p < 1$, *we have* $\lambda^* \in (q, p)$.

The proof is given in Section C.5.

With the facts above, we establish the following critical basic inequality.

Lemma 1 *Any* $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A})$ *satisfies the inequality*

$$0 \le \left\langle -\mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle + \left\langle \mathbf{W}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle.$$

The proof is given in Section C.6.

We define the shorthands $S_1 \coloneqq \left\langle -\mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle$ and $S_2 \coloneqq \left\langle \mathbf{W}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle$. In the following, we first control S_2 and then derive the exponential error rate from $S_1 + S_2$.

B.2. Controlling S_2

The proposition below provides a bound on S_2.

Proposition 1 *Under the conditions of Theorem 2, with probability at least* $1 - \frac{1}{n^2} - \frac{3}{\sqrt{n}}$ *at least one of the following inequalities holds:*

$$\frac{\gamma}{n} \le \left\lfloor n \exp\left[-\left(1 - C_e\sqrt{\frac{1}{nI}}\right)\frac{nI}{2}\right]\right\rfloor,$$
$$S_2 \le C\frac{1}{t^*}\gamma\sqrt{\frac{I}{n}}, \tag{10}$$

for some constants $C_e, C > 0$.

The proof is given in Section C.7. WLOG we assume the second equation in Proposition 1 holds. By Lemma 1, we have

$$0 \le \left\langle -\mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle + C\frac{1}{t^*}\gamma I\sqrt{\frac{1}{nI}}. \tag{11}$$

B.3. Analyzing $S_1 + S_2$

If $\gamma = 0$ then we are done. Therefore, we assume $\gamma > 0$ in the following. By Fact 2, $\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)$ is psd and all its diagonal entries are non-negative, with

$$\mathrm{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) = \frac{\gamma}{n}$$

Let $b_i := \left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right)_{ii}$, $b_{\max} = 4$, $\beta := \frac{1}{b_{\max}}\sum_{i\in[n]} b_i = \frac{\gamma}{b_{\max} n}$ and $X_i := -d_i$. By Fact 3, $\frac{b_i}{b_{\max}} \in [0,1]$. Note that each X_i is the sum of $n-1$ independent random variables, but X_i is not independent with X_j for $i \neq j$. To proceed, we show that the quantity $\left\langle -\mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle$ is bounded by the sum of a certain subset of $\{X_i\}$ with size roughly β, which can be further controlled by the following lemma.

Lemma 2 *Let* $\eta := C\sqrt{\frac{1}{nI}}$ *for some large constant* $C > 0$ *and* M *be any positive number satisfying* $1 \le M \le C\sqrt{\frac{n}{I}}$. *There exists some constant* $C_I > 0$ *that only depends on* C *such that the following holds. If* $nI \ge C_I$ *then we have*

$$\max_{\mathcal{M}\subset[n],\, |\mathcal{M}|=m}\left[\sum_{i\in\mathcal{M}} X_i\right] \le \frac{1}{t^*}\left((1+\eta)m\log\left(\frac{ne}{m}\right) - (1-2\eta)\frac{mn}{2}I\right), \quad m = 1, 2, \ldots, \lfloor M\rfloor$$

with probability at least $1 - 3\exp\left(-\sqrt{\log n}\right)$.

The proof is given in Section C.8. To see that the range of M is valid, one can verify that $1 \le C'\sqrt{\frac{n}{I}}$ for some constant $C' > 0$. Indeed, $p \le 1$ implies $1 \le \sqrt{\frac{n}{p}}$, and $p \ge \frac{(p-q)^2}{p}$ together with Fact 5 implies $\sqrt{\frac{n}{p}} \le \sqrt{\frac{np}{(p-q)^2}} \le C'\sqrt{\frac{n}{I}}$.

We need a simple pilot bound, which ensures that the SDP solution satisfies a non-trivial error bound.

Lemma 3 *Under Model 1, if $p \geq \frac{1}{n}$ then there exists some constant $C > 0$ such that*

$$\gamma \leq C\sqrt{\frac{n^3}{I}}$$

with probability at least $1 - 2(e/2)^{-2n}$.

The proof is given in Section C.9.

To proceed, we employ a technique involving order statistics similar to those in Fei and Chen (2018, 2019b). Let $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ be the order statistics of $\{X_i\}$. We can write

$$\begin{aligned} S_1 &= \sum_{i\in[n]} X_i b_i \\ &= b_{\max} \sum_{i\in[n]} X_i \left(\frac{b_i}{b_{\max}}\right). \end{aligned}$$

Let η be as defined in Lemma 2.

If $0 < \beta \leq 1$, we have

$$\begin{aligned} S_1 &\leq b_{\max}\beta X_{(1)} \\ &\overset{(a)}{\leq} b_{\max}\beta \left[\frac{1}{t^*}\left((1+\eta)\log\left(\frac{ne}{1}\right) - (1-2\eta)\frac{n}{2}I\right)\right] \\ &= b_{\max}\beta \left[\frac{1}{t^*}\left((1+\eta)\log\left(\frac{ne}{\lceil\beta\rceil}\right) - (1-2\eta)\frac{n}{2}I\right)\right] \end{aligned}$$

where step (a) follows from Lemma 2 with $M = 1$ and the last step holds since $\lceil\beta\rceil = 1$. Then Equation (11) yields

$$\begin{aligned} 0 &\leq C\frac{1}{t^*}\sqrt{\frac{1}{nI}}\gamma I + b_{\max}\beta \left[\frac{1}{t^*}\left((1+\eta)\log\left(\frac{ne}{\lceil\beta\rceil}\right) - (1-2\eta)\frac{n}{2}I\right)\right] \\ &= b_{\max}\beta \left[\frac{1}{t^*}\left((1+\eta)\log\left(\frac{ne}{\lceil\beta\rceil}\right) - (1-C_e''\eta)\frac{n}{2}I\right)\right] \end{aligned}$$

for some constant $C_e'' > 0$. Since $nI \geq C_I$ implies $\eta \leq C'\sqrt{\frac{1}{C_I}}$, rearranging the above equation yields

$$\lceil\beta\rceil \leq en\exp\left[-(1-C_e'\eta)\frac{nI}{2}\right]$$

for some constant $C_e' > 0$. Since $\lceil x\rceil \leq y$ implies $x \leq \lfloor y\rfloor$ for any $x, y \geq 0$, the above inequality yields

$$\frac{\gamma}{b_{\max}n} = \beta \leq \left\lfloor en\exp\left[-(1-C_e'\eta)\frac{nI}{2}\right]\right\rfloor.$$

To proceed, we record a fact related to the floor function.

Fact 7 *For any $x \geq 0$ and positive integer c, we have $c\lfloor x\rfloor \leq \lfloor cx\rfloor$.*

Proof We have $\lfloor cx \rfloor = \lfloor c \lfloor x \rfloor + (cx - c \lfloor x \rfloor) \rfloor \geq \lfloor c \lfloor x \rfloor \rfloor = c \lfloor x \rfloor$ by noting that $cx - c \lfloor x \rfloor \geq 0$ and $c \lfloor x \rfloor$ is an integer. ■

By Fact 7 and the definition $b_{\max} = 4$, we have

$$\begin{aligned}\frac{\gamma}{n} &\leq b_{\max} \left\lfloor en \exp\left[-(1 - C'_e\eta)\frac{nI}{2}\right]\right\rfloor \\ &= \left\lfloor b_{\max} en \exp\left[-(1 - C'_e\eta)\frac{nI}{2}\right]\right\rfloor \\ &= \left\lfloor n \exp\left[-(1 - C'_e\eta)\frac{nI}{2} + \log\left(b_{\max} e\right)\right]\right\rfloor\end{aligned}$$

As long as $nI \geq 1$, we have $\sqrt{\frac{1}{nI}} \geq \frac{1}{nI}$ and

$$\frac{\gamma}{n} \leq \left\lfloor n \exp\left[-(1 - C_e\eta)\frac{nI}{2}\right]\right\rfloor$$

for some constant $C_e \geq C'_e$. We have arrived at the desired result for $0 < \beta \leq 1$.

Now if $\beta > 1$, then

$$S_1 \leq b_{\max}\left[\sum_{i \in [\lfloor \beta \rfloor]} X_{(i)} + (\beta - \lfloor \beta \rfloor) X_{(\lceil \beta \rceil)}\right]$$

Continuing from Equation (11), we have

$$\begin{aligned}0 &\leq S_1 + C b_{\max} \beta \frac{1}{t^*}\sqrt{\frac{1}{nI}} nI \\ &\leq b_{\max}\left[\sum_{i \in [\lfloor \beta \rfloor]} \left(X_{(i)} + C\frac{1}{t^*}\sqrt{\frac{1}{nI}} nI\right) + (\beta - \lfloor \beta \rfloor)\left(X_{(\lceil \beta \rceil)} + C\frac{1}{t^*}\sqrt{\frac{1}{nI}} nI\right)\right].\end{aligned}$$

It would be challenging to handle the residual term $(\beta - \lfloor \beta \rfloor) X_{(\lceil \beta \rceil)}$ when β is not an integer. Fortunately, taking the integer part of β enables us to control the SDP error γ more easily while not loosening the bound by too much, with help of the next lemma.

Lemma 4 *Let $\{\phi_i\}_{i \in [n]}$ be such that $\phi_1 \geq \phi_2 \geq \ldots$, and $u' \in [1, n]$. Define $V(u) \coloneqq \sum_{i \in [\lfloor u \rfloor]} \phi_i + (u - \lfloor u \rfloor)\phi_{\lceil u \rceil}$. If $0 \leq V(u')$, then we have $0 \leq V(u_0)$ for any $u_0 \in [1, u']$.*

The proof can be found in Section C.10.

Let $\beta_0 \coloneqq \lfloor \beta \rfloor$. By Lemma 4 with $u' = \beta$ and $\phi_{(i)} = X_{(i)} + C\frac{1}{t^*}\sqrt{\frac{1}{nI}} nI$, last displayed equation implies that

$$0 \leq b_{\max} \sum_{i \in [\beta_0]} \left(X_{(i)} + C\frac{1}{t^*}\sqrt{\frac{1}{nI}} nI\right)$$

$$= b_{\max} \sum_{i \in [\beta_0]} X_{(i)} + b_{\max} \beta_0 C \frac{1}{t^*} \sqrt{\frac{1}{nI}} nI.$$

Lemma 3 implies that $\beta_0 \le \beta \le C' \sqrt{\frac{n}{I}}$ for some constant $C' > 0$ with high probability, and therefore by Lemma 2 with $M = C' \sqrt{\frac{n}{I}}$, with high probability we have

$$\begin{aligned} 0 &\le \frac{1}{t^*} \left((1+\eta) \beta_0 \log \left(\frac{ne}{\beta_0} \right) - (1-2\eta) \frac{\beta_0 n}{2} I \right) + C \frac{1}{t^*} \beta_0 nI \sqrt{\frac{1}{nI}} \\ &= \beta_0 \frac{1}{t^*} \left((1+\eta) \log \left(\frac{ne}{\beta_0} \right) - (1 - C_e'' \eta) \frac{n}{2} I \right) \end{aligned}$$

for some constant $C_e'' > 0$. Since $nI \ge C_I$ implies $\eta \le C' \sqrt{\frac{1}{C_I}}$, rearranging the above equation yields

$$\beta_0 \le en \exp \left[-(1 - C_e' \eta) \frac{nI}{2} \right]$$

for some constant $C_e' > 0$. Since β_0 is an integer by definition, we must also have

$$\beta_0 \le \left\lfloor en \exp \left[-(1 - C_e' \eta) \frac{nI}{2} \right] \right\rfloor.$$

Given $\beta > 1$ and the definition of β, we have $\beta \le 2\beta_0$ and $\frac{\gamma}{n} = b_{\max} \beta$. Then Fact 7 and the definition $b_{\max} = 4$ implies

$$\begin{aligned} \frac{\gamma}{n} &\le 2 b_{\max} \left\lfloor en \exp \left[-(1 - C_e' \eta) \frac{nI}{2} \right] \right\rfloor \\ &\le \left\lfloor 2 b_{\max} en \exp \left[-(1 - C_e' \eta) \frac{nI}{2} \right] \right\rfloor. \end{aligned}$$

As long as $nI \ge 1$, we have $\sqrt{\frac{1}{nI}} \ge \frac{1}{nI}$ and

$$\begin{aligned} \frac{\gamma}{n} &\le \left\lfloor n \exp \left[-(1 - C_e' \eta) \frac{nI}{2} + \log(2 b_{\max} e) \right] \right\rfloor \\ &= \left\lfloor n \exp \left[-(1 - C_e \eta) \frac{nI}{2} \right] \right\rfloor \end{aligned}$$

for some constant $C_e \ge C_e'$. The proof is completed.

Appendix C. Proofs of Technical Lemmas in Section B

C.1. Proof of Fact 1

The conclusion follows that

$$\left\langle \mathbb{E}\mathbf{A} - \lambda^* (\mathbf{J} - \mathbf{I}), \mathcal{P}_{T^\perp} \left(\widehat{\mathbf{Y}} \right) \right\rangle = \left\langle (\mathbb{E}\mathbf{A} + p\mathbf{I}) - \lambda^* \mathbf{J} - (p - \lambda^*) \mathbf{I}, \mathcal{P}_{T^\perp} \left(\widehat{\mathbf{Y}} \right) \right\rangle$$

$$
\begin{aligned}
&= \left\langle (\mathbb{E}\mathbf{A} + p\mathbf{I}) - \lambda^*\mathbf{J}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\rangle - (p-\lambda^*)\operatorname{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) \\
&= \left\langle \mathcal{P}_{T^\perp}((\mathbb{E}\mathbf{A} + p\mathbf{I}) - \lambda^*\mathbf{J}), \widehat{\mathbf{Y}}\right\rangle - (p-\lambda^*)\operatorname{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) \\
&\overset{(a)}{=} \left\langle \mathcal{P}_{T^\perp}\left(\frac{p-q}{2}\mathbf{Y}^* + \frac{p+q}{2}\mathbf{J} - \lambda^*\mathbf{J}\right), \widehat{\mathbf{Y}}\right\rangle - (p-\lambda^*)\operatorname{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) \\
&\overset{(b)}{=} \left\langle \left(\frac{p+q}{2} - \lambda^*\right)\mathbf{J}, \widehat{\mathbf{Y}}\right\rangle - (p-\lambda^*)\operatorname{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) \\
&= -(p-\lambda^*)\operatorname{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right)
\end{aligned}
$$

where step (a) holds since $(\mathbb{E}\mathbf{A}+p\mathbf{I})-\lambda^*\mathbf{J} = \frac{p-q}{2}\mathbf{Y}^* + \frac{p+q}{2}\mathbf{J} - \lambda^*\mathbf{J}$, step (b) holds since $\mathcal{P}_{T^\perp}(\mathbf{Y}^*)$ is equal to the all-zero matrix and $\mathcal{P}_{T^\perp}(\mathbf{J}) = \mathbf{J}$ (since $\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\mathbf{J} = \left(\mathbf{I} - \frac{1}{n}\mathbf{Y}^*\right)\mathbf{J} = \mathbf{J}$), and the last step holds since $\left\langle \mathbf{J}, \widehat{\mathbf{Y}}\right\rangle = 0$.

C.2. Proof of Fact 2

Fix any $\mathbf{x} \in \mathbb{R}^n$ and let $\mathbf{v} \coloneqq \mathbf{x}^\top\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)$. Then

$$
\begin{aligned}
\mathbf{x}^\top\left[\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right]\mathbf{x} &= \mathbf{x}^\top\left[\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\widehat{\mathbf{Y}}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\right]\mathbf{x} \\
&= \mathbf{v}^\top\widehat{\mathbf{Y}}\mathbf{v} \\
&\geq 0
\end{aligned}
$$

where the last step holds since $\widehat{\mathbf{Y}} \succeq 0$ by feasibility to program (6). Therefore, $\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \succeq 0$. We also have

$$
\begin{aligned}
\operatorname{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) &= \operatorname{Tr}\left(\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\right) \\
&\overset{(a)}{=} \operatorname{Tr}\left(\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\right) \\
&\overset{(b)}{=} \operatorname{Tr}\left(\left(-\mathbf{U}\mathbf{U}^\top\right)\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\right) \\
&= \frac{1}{n}\operatorname{Tr}\left((-\mathbf{Y}^*)\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\right) \\
&= \frac{\gamma}{n},
\end{aligned}
$$

where step (a) holds since trace is invariant under cyclic permutations and the matrix $\mathbf{I} - \mathbf{U}\mathbf{U}^\top$ is idempotent, and step (b) holds since $Y^*_{ii} - \widehat{Y}_{ii} = 0$ for $i \in [n]$.

C.3. Proof of Fact 3

The result follows from the definition of $\mathcal{P}_{T^\perp}(\cdot)$ and direct calculation

$$
\begin{aligned}
\|\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\|_\infty &\leq \|\widehat{\mathbf{Y}}\|_\infty + \|\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\|_\infty + \|\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\|_\infty \\
&\quad + \|\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\|_\infty \\
&\leq 4.
\end{aligned}
$$

C.4. Proof of Fact 4

By definition of t^* in Equation (8), we have

$$t^* = \frac{1}{2}\left[\log\frac{p}{q} + \log\frac{1-q}{1-p}\right].$$

We discuss two cases based on whether $\frac{p}{q}$ is larger than $\frac{1-q}{1-p}$.

If $\frac{p}{q} \geq \frac{1-q}{1-p}$, we have

$$\begin{aligned} t^* &\leq \log\frac{p}{q} \\ &= \log\left(1 + \frac{p-q}{q}\right) \\ &\overset{(a)}{\leq} \frac{p-q}{q} \\ &\leq \frac{p-q}{c_0 p} \end{aligned}$$

where step (a) holds since the fact that $1 + x \leq e^x$ for $x \in \mathbb{R}$ implies $\log(1+x) \leq x$ for $x \geq 0$, and the last step holds by the assumption that $c_0 p \leq q$.

If $\frac{p}{q} \leq \frac{1-q}{1-p}$, then

$$\begin{aligned} t^* &\leq \log\frac{1-q}{1-p} \\ &\overset{(a)}{\leq} \frac{1-q}{1-p} - 1 \\ &\overset{(b)}{\leq} \frac{p-q}{c_1} \\ &\leq \frac{p-q}{c_1 p}, \end{aligned}$$

where step (a) holds since $\log x \leq x - 1$ for $x > 0$, step (b) holds by our assumption $p \leq 1 - c_1$, and the last step holds since $p \leq 1$.

Hence, taking $C = \max\left\{\frac{1}{c_0}, \frac{1}{c_1}\right\}$ finishes the proof.

C.5. Proof of Fact 6

We want to show

$$0 < p - \lambda^* = \left[\log\frac{p(1-q)}{q(1-p)}\right]^{-1}\left[p\log\frac{p(1-q)}{q(1-p)} - \log\frac{1-q}{1-p}\right]$$

by definition of λ^*. Since $0 < q < p < 1$, we have $\log\frac{p(1-q)}{q(1-p)} > 0$. We also have

$$p\log\frac{p(1-q)}{q(1-p)} - \log\frac{1-q}{1-p} = p\log\frac{p}{q} + p\log\frac{1-q}{1-p} - \log\frac{1-q}{1-p}$$

$$= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q},$$

which is the KL divergence between $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$. It is positive since $p \neq q$, whence $p - \lambda^* > 0$ as claimed.

Similarly, we have

$$\lambda^* - q = \left[\log \frac{p(1-q)}{q(1-p)}\right]^{-1} \left[q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}\right].$$

With $0 < q < p < 1$, the quantity inside the first bracket on the RHS is positive, and the quantity inside the second bracket is the KL divergence and thus is also positive. We therefore conclude that $\lambda^* - q > 0$.

C.6. Proof of Lemma 1

Recall the definitions of $\mathbf{A}^0, \mathbf{W}, \mathbf{D}, \mathbf{d}$ defined in the beginning of Section B. Note that

$$\begin{aligned}
\mathbf{D}\boldsymbol{\sigma}^* &= \boldsymbol{\sigma}^* \circ \mathbf{d} \\
&= \boldsymbol{\sigma}^* \circ \boldsymbol{\sigma}^* \circ \left(\mathbf{A}^0 \boldsymbol{\sigma}^*\right) \\
&= \mathbf{A}^0 \boldsymbol{\sigma}^*
\end{aligned}$$

and therefore

$$\mathbf{D}\mathbf{Y}^* = \mathbf{D}\boldsymbol{\sigma}^* \left(\boldsymbol{\sigma}^*\right)^\top = \mathbf{A}^0 \boldsymbol{\sigma}^* \left(\boldsymbol{\sigma}^*\right)^\top = \mathbf{A}^0 \mathbf{Y}^*.$$

Since $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A})$, we have

$$\begin{aligned}
0 &\leq \left\langle \mathbf{A}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\overset{(a)}{=} \left\langle \mathbf{A}^0 - \mathbf{D}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\overset{(b)}{=} \left\langle \mathbf{A}^0 - \mathbf{D}, \widehat{\mathbf{Y}} \right\rangle \\
&= \left\langle \mathbf{A}^0 - \mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \right\rangle + \left\langle \mathbf{A}^0 - \mathbf{D}, \mathcal{P}_{T}\left(\widehat{\mathbf{Y}}\right) \right\rangle \\
&\overset{(c)}{=} \left\langle \mathbf{A}^0 - \mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \right\rangle \\
&= \left\langle \mathbf{W} - \mathbf{D} + (\mathbb{E}\mathbf{A} - \lambda^*(\mathbf{J} - \mathbf{I})), \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \right\rangle \\
&\overset{(d)}{\leq} \left\langle -\mathbf{D}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \right\rangle + \left\langle \mathbf{W}, \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right) \right\rangle
\end{aligned}$$

where

- step (a) holds since $\left\langle \mathbf{J}, \widehat{\mathbf{Y}} \right\rangle = \langle \mathbf{J}, \mathbf{Y}^* \rangle = 0$, $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ has zero diagonal and $\mathbf{D}$ is a diagonal matrix;
- step (b) holds since $\mathbf{D}\mathbf{Y}^* = \mathbf{A}^0 \mathbf{Y}^*$;

- step (c) holds since

$$\left\langle \mathbf{A}^0 - \mathbf{D}, \mathcal{P}_T\left(\widehat{\mathbf{Y}}\right)\right\rangle = \left\langle \mathbf{A}^0 - \mathbf{D}, \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\right\rangle + \left\langle \mathbf{A}^0 - \mathbf{D}, \widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\right\rangle - \left\langle \mathbf{A}^0 - \mathbf{D}, \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\right\rangle$$

  and $\mathbf{U}\mathbf{U}^\top = n^{-1}\mathbf{Y}^* = n^{-1}\boldsymbol{\sigma}^*\left(\boldsymbol{\sigma}^*\right)^\top$, there exist some vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ such that

$$\left\langle \mathbf{A}^0 - \mathbf{D}, \mathcal{P}_T\left(\widehat{\mathbf{Y}}\right)\right\rangle = \left\langle \mathbf{A}^0 - \mathbf{D}, \boldsymbol{\sigma}^*\mathbf{u}^\top + \mathbf{v}\left(\boldsymbol{\sigma}^*\right)^\top\right\rangle = 0$$

  where the last step follows $\mathbf{D}\boldsymbol{\sigma}^* = \mathbf{A}^0\boldsymbol{\sigma}^*$;

- step (d) follows from Fact 1, and $\mathrm{Tr}\left(\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right) \geq 0$ by Fact 2, and $p - \lambda^* \geq 0$ by Fact 6.

C.7. Proof of Proposition 1

Recall that

$$S_2 = \left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{W}\right\rangle. \tag{12}$$

We control the right hand side of Equation (12) by splitting the term into two parts, one involving a trimmed version of $\mathbf{W}$ and the other the residual. This technique is similar to those in Fei and Chen (2019b); Zhang and Zhou (2017), but here we provide somewhat tighter bounds.

C.7.1. TRIMMING

We need two technical lemmas concerning properties of a trimmed Bernoulli matrix and its residual.

Lemma 5 *Suppose $\mathbf{M} \in \mathbb{R}^{n\times n}$ is a random matrix with zero on the diagonal and independent entries $\{M_{ij}\}$ with the following distribution*

$$M_{ij} = \begin{cases} 1 - p_{ij}, & \text{w.p. } p_{ij}, \\ -p_{ij}, & \text{w.p. } 1 - p_{ij}. \end{cases}$$

Let $p' \coloneqq \max_{ij} p_{ij}$ and $\widetilde{\mathbf{M}}$ be the matrix obtained from $\mathbf{M}$ by zeroing out all the rows and columns having more than $40np'$ positive entries. Then there exists some constant $C > 0$ such that

$$\|\widetilde{\mathbf{M}}\|_{\mathrm{op}} \leq C\sqrt{np'}$$

with probability at least $1 - \frac{1}{n^2}$.

Proof The claim follows from Lemma 9 in Fei and Chen (2019b) with σ^2 therein set to p'. ■

Lemma 6 *Let $\mathbf{M} \in \{0,1\}^{n\times n}$ be a binary matrix with $M_{ii} = 0$ for all $i \in [n]$, and $\{M_{ij}\}_{i\in[n],j\in[n]}$ being independent Bernoulli random variables. Let $p' \coloneqq \max_{ij}\mathbb{E}M_{ij}$. Define $\mathcal{T} \coloneqq \{i \in [n] : \sum_j M_{ij} \geq 40np'\}$ and $Z_i \coloneqq \sum_j |M_{ij} - \mathbb{E}M_{ij}|\,\mathbb{I}\{i \in \mathcal{T}\}$. If $p' \geq \frac{C}{n}$ for a sufficiently large positive constant C, then we have*

$$\sum_i Z_i \leq 40n^2p'e^{-5np'}$$

with probability at least $1 - \frac{1}{\sqrt{n}}$.

Proof Define the event $B := \left\{\sum_i Z_i > 40n^2p'e^{-5np'}\right\}$. First consider the case where $np' \geq \frac{1}{10}\log n$. Applying Lemma C.5 in Zhang and Zhou (2017) with p therein set to $2p'$,[4] we obtain that $\mathbb{P}\{B\} \leq e^{-10np'}$. Under the case $np' \geq \frac{1}{10}\log n$, this probability is at most $e^{-\log n} \leq \frac{1}{\sqrt{n}}$ as claimed. Next consider the case where $C \leq np' < \frac{1}{10}\log n$. Since $\sum_j M_{ij}\mathbb{I}\{i \in \mathcal{T}\} \geq 40np'\mathbb{I}\{i \in \mathcal{T}\}$ by definition of $\mathcal{T}$, we have

$$\begin{aligned}\sum_i Z_i &\leq \sum_i \sum_j M_{ij}\mathbb{I}\{i \in \mathcal{T}\} + np' \sum_i \mathbb{I}\{i \in \mathcal{T}\} \\ &\leq 2\sum_i \sum_j M_{ij}\mathbb{I}\{i \in \mathcal{T}\}.\end{aligned}$$

Set $\varepsilon = 20np'e^{-5np'}$; note that $\varepsilon \in (0, 1/2]$ and $21np' + 2\log\varepsilon^{-1} \leq 40np'$ since $p' \geq \frac{C}{n}$. Applying Lemma 8.1 in Rebrova and Vershynin (2016) with the above ε,[5] we obtain that

$$\begin{aligned}\mathbb{P}\left\{\sum_i \sum_j M_{ij}\mathbb{I}\{i \in \mathcal{T}\} > 20n^2p'e^{-5np'}\right\} &\leq \exp\left(-10n^2p'e^{-5np'}\right) \\ &\leq \exp\left(-10Cne^{-\frac{1}{2}\log n}\right) \\ &= \exp\left(-10C\sqrt{n}\right) \leq \frac{1}{\sqrt{n}}.\end{aligned}$$

Combining the last two display equations proves that $\mathbb{P}\{B\} \leq \frac{1}{\sqrt{n}}$ as claimed. ■

Let $\mathbf{W}^{\mathrm{up}}$ be obtained from $\mathbf{W}$ by zeroing out its lower triangular entries. Turning to S_2, we observe that

$$\begin{aligned}S_2 &= 2\left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{W}^{\mathrm{up}}\right\rangle \\ &= 2\left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \widetilde{\mathbf{W}^{\mathrm{up}}}\right\rangle + 2\left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{W}^{\mathrm{up}} - \widetilde{\mathbf{W}^{\mathrm{up}}}\right\rangle \\ &\overset{(a)}{\leq} 2\,\mathrm{Tr}\left[\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right] \cdot \|\widetilde{\mathbf{W}^{\mathrm{up}}}\|_{\mathrm{op}} + 2\|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_\infty \|\mathbf{W}^{\mathrm{up}} - \widetilde{\mathbf{W}^{\mathrm{up}}}\|_1 \\ &\overset{(b)}{\leq} 2\frac{\gamma}{n}\|\widetilde{\mathbf{W}^{\mathrm{up}}}\|_{\mathrm{op}} + 8\|\mathbf{W}^{\mathrm{up}} - \widetilde{\mathbf{W}^{\mathrm{up}}}\|_1,\end{aligned}$$

where step (a) holds since $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}) \succeq 0$ (by Fact 2) implies $\|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_* = \mathrm{Tr}\left\{\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\}$, and step (b) holds by Fact 2 and Fact 3. We then apply Lemma 5 to $\widetilde{\mathbf{W}^{\mathrm{up}}}$ to bound $\|\widetilde{\mathbf{W}^{\mathrm{up}}}\|_{\mathrm{op}}$, and apply Lemma 6 to $\mathbf{W}^{\mathrm{up}}$ and $(\mathbf{W}^{\mathrm{up}})^\top$ to bound $\|\mathbf{W}^{\mathrm{up}} - \widetilde{\mathbf{W}^{\mathrm{up}}}\|_1$. Note that the assumption $p' \geq \frac{C}{n}$ of Lemma 6 is satisfied by the assumption of this proposition that $nI \geq C_I$ for some large enough $C_I > 0$ (since Fact 5 implies $I \lesssim \frac{(p-q)^2}{p} \leq p$). Then with probability at least $1 - \frac{1}{n^2} - \frac{2}{\sqrt{n}}$, there holds

$$S_2 \leq C_0\frac{\gamma}{n}\sqrt{np} + C_1n^2pe^{-5np} =: C_0Q_1 + C_1Q_2$$

for some constants $C_0, C_1 > 0$.

4. Inspecting their proof, we see that their bound holds without change for matrices with independent entries.
5. Inspecting their proof, we see that their bound holds without change when the means of the Bernoulli are upper bounded by the same p'.

C.7.2. CONTROLLING Q_1

Note that Fact 5 implies $\sqrt{I} \le C' \frac{p-q}{\sqrt{p}}$ for some constant $C' > 0$ and therefore $Q_1 = \gamma \frac{p-q}{p-q}\sqrt{\frac{p}{n}} \le \frac{1}{C'}\gamma(p-q)\sqrt{\frac{1}{nI}}$.

C.7.3. CONTROLLING Q_2

We record an elementary inequality.

Lemma 7 *For any number $\alpha > 0$ with $\alpha \asymp 1$, there exists a constant $C(\alpha) \ge 1$ (that only depends on α) such that if $nI \ge 2C(\alpha)$, then*

$$pe^{-pn/(2\alpha)} \le (p-q)e^{-nI/(4\alpha)}.$$

Proof Note that $pn \ge (p-q)n \ge \frac{(p-q)^2}{p} n \ge \frac{1}{C'} nI \ge \frac{1}{C'} 2C(\alpha)$ for some constant $C' > 0$ by Fact 5. As long as $C(\alpha)$ is sufficiently large, we have $\frac{pn}{2} \le e^{pn/(4\alpha)}$. These inequalities imply that

$$\frac{p}{p-q} \le \frac{pn}{2} \le e^{pn/(4\alpha)} \le e^{(2p-I)n/(4\alpha)}.$$

Multiplying both sides by $(p-q)e^{-2pn/(4\alpha)}$ yields the claimed inequality. ■

Equipped with the above bound, we are ready to bound Q_2. Let $\xi = C_e\sqrt{\frac{1}{nI}}$ for some constant $C_e > 0$ such that $\lfloor ne^{-(1-\xi)nI/2} \rfloor > 0$. If $\frac{\gamma}{n} = 0$ or $\frac{\gamma}{n} \le \lfloor ne^{-(1-\xi)nI/2} \rfloor$, then the first inequality in Proposition 1 holds and we are done. It remains to consider the case $\frac{\gamma}{n} > \lfloor ne^{-(1-\xi)nI/2} \rfloor > 0$. We have that $\lfloor ne^{-(1-\xi)nI/2} \rfloor$ is a positive integer and $\gamma > n \lfloor ne^{-(1-\xi)nI/2} \rfloor \ge \frac{1}{2} n^2 e^{-(1-\xi)nI/2}$. Lemma 7 with $\alpha = \frac{1}{5}$ implies that

$$\begin{aligned}
Q_2 &\le pn^2 e^{-5pn/2} \\
&\le (p-q)n^2 e^{-5nI/4} \\
&\le (p-q)e^{-nI/2} \cdot n^2 e^{-nI/2} \\
&\le (p-q)e^{-nI/2} \cdot n^2 e^{-(1-\xi)nI/2} \\
&\le 2(p-q)e^{-nI/2} \cdot \gamma
\end{aligned}$$

Choosing $C_I > 0$ large enough so that $e^{-nI/2} \le \sqrt{\frac{1}{nI}}$, we have $Q_2 \le 2\gamma(p-q)\sqrt{\frac{1}{nI}}$.

C.7.4. PUTTING TOGETHER

So far, we have shown that

$$S_2 \le C_2 \gamma (p-q)\sqrt{\frac{1}{nI}}$$

for some constant $C_2 > 0$. Under the assumption $0 < c_0 p \le q < p \le 1 - c_1$, we have $p - q \le \frac{C'I}{t^*}$ for some constant $C' > 0$ by Fact 4 and Fact 5. The proof is completed.

C.8. Proof of Lemma 2

Recall that $X_i = -d_i$ and $\mathbf{d} = \operatorname{diag}(\mathbf{D}) = \boldsymbol{\sigma}^* \circ ((\mathbf{A} - \lambda^*(\mathbf{J} - \mathbf{I}))\boldsymbol{\sigma}^*)$. For clarity of exposition, we define the shorthands

$$\begin{aligned}
L_m &\coloneqq \max_{\mathcal{M}\subset[n],\, |\mathcal{M}|=m} \left[-\sum_{i\in\mathcal{M}} d_i \right], \quad m \in [\lfloor M \rfloor], \\
L_{m,\mathcal{M}} &\coloneqq -\sum_{i\in\mathcal{M}} d_i, \quad \mathcal{M} \subset [n], |\mathcal{M}| = m, \\
R_m &\coloneqq \frac{1}{t^*}\left((1+\eta) m \log\left(\frac{ne}{m}\right) - (1-2\eta)\frac{mn}{2} I \right), \quad m \in [\lfloor M \rfloor], \\
P_{m,\mathcal{M}} &\coloneqq \mathbb{P}\left(L_{m,\mathcal{M}} \geq R_m \right), \\
P_m &\coloneqq \mathbb{P}\left(\exists \mathcal{M} \subset [n], |\mathcal{M}| = m : L_{m,\mathcal{M}} \geq R_m \right), \\
P &\coloneqq \mathbb{P}\left(\exists m \in [M] : L_m \geq R_m \right).
\end{aligned}$$

Our goal is to show that $P \leq 3\exp\left(-\sqrt{\log n}\right)$. We start the proof by controlling $P_{m,\mathcal{M}}$ for a fixed $\mathcal{M} \subset [n]$ with $|\mathcal{M}| = m$.

C.8.1. A CLOSER LOOK AT $L_{m,\mathcal{M}}$

Let $\{Z_j\}, \{Z_j'\} \overset{\text{i.i.d.}}{\sim} \operatorname{Bern}(q)$,$\{Y_j\}, \{Y_j'\} \overset{\text{i.i.d.}}{\sim} \operatorname{Bern}(p)$ such that each of them is independent of the rest. By definition, each $-d_i$ is the sum of $\frac{n}{2}$ random variables $\{Z_j - \lambda^*\}$ and $\frac{n}{2} - 1$ random variables $\{-Y_j + \lambda^*\}$. For fixed m and $\mathcal{M}$, the quantity $L_{m,\mathcal{M}}$ is the sum of $mn - m$ random variables. It can be seen that $L_{m,\mathcal{M}} = \sum_{j\in[mn-m]} V_j$, where each V_j distributes as either $Z_1 - \lambda^*$ or $Y_1 - \lambda^*$.

Due to symmetry of $\mathbf{A}$, there may exist some $j \neq j' \in [mn - m]$ such that V_j and $V_{j'}$ identify the same random variable. Let us define a set to group together all such random variables. We set

$$\mathcal{J} \coloneqq \left\{ j \in [mn - m] : \exists j' \in [mn - m] \backslash \{j\} \text{ s.t. } V_j = V_{j'} \right\}.$$

Note that $|\mathcal{J}| = 0$ for $m = 1$ and $|\mathcal{J}| > 0$ for $m > 1$. We further split the set $\{V_j\}$ according to their distribution and quantify the sizes of the resulting partitions:

$$\begin{aligned}
m_p &\coloneqq \left| \{ j \notin \mathcal{J} : \mathbb{E}U_j = -p + \lambda^* \} \right|, \\
m_q &\coloneqq \left| \{ j \notin \mathcal{J} : \mathbb{E}U_j = q - \lambda^* \} \right|, \\
m_p' &\coloneqq \frac{1}{2} \left| \{ j \in \mathcal{J} : \mathbb{E}U_j = -p + \lambda^* \} \right|, \\
m_q' &\coloneqq \frac{1}{2} \left| \{ j \in \mathcal{J} : \mathbb{E}U_j = q - \lambda^* \} \right|.
\end{aligned}$$

It is not hard to see that $m_q + m_p = mn - m^2$ and $m_q' + m_p' = \frac{m(m-1)}{2}$. Now we can write

$$L_{m,\mathcal{M}} = \left[\sum_{j\in[m_q]} (Z_j - \lambda^*) - \sum_{j\in[m_p]} (Y_j - \lambda^*) \right] + 2\left[\sum_{j\in[m_q']} (Z_j' - \lambda^*) - \sum_{j\in[m_p']} (Y_j' - \lambda^*) \right].$$

Recall the definition of $t^* > 0$ in Equation (8). We have

$$P_{m,\mathcal{M}} = \mathbb{P}\left(L_{m,\mathcal{M}} \geq R_m \right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\exp\left(t^* L_{m,\mathcal{M}}\right) \geq \exp\left(t^* R_m\right)\right) \\
&\overset{(a)}{\leq} \exp\left(-t^* R_m\right) \left[\mathbb{E}\exp\left(t^* \sum_{j\in[m_q]} (Z_j - \lambda^*) - t^* \sum_{j\in[m_p]} (Y_j - \lambda^*)\right)\right] \\
&\quad \times \left[\mathbb{E}\exp\left(2t^* \sum_{j\in[m'_q]} (Z'_j - \lambda^*) - 2t^* \sum_{j\in[m'_p]} (Y'_j - \lambda^*)\right)\right] \\
&=: Q_1 Q_2 Q_3,
\end{aligned}
$$

where step (a) holds by Chernoff inequality.

C.8.2. CONTROLLING Q_1

By definition of R_m, we have

$$
Q_1 = \exp\left(-(1+\eta) m \log\left(\frac{ne}{m}\right) + (1-2\eta)\frac{mn}{2} I\right)
$$

To control Q_2 and Q_3, we need the following result about Bernoulli moment generating functions.

Fact 8 *Let $Z \sim \mathrm{Bern}(q)$ and $Y \sim \mathrm{Bern}(p)$. We have the following identities*

$$
\begin{aligned}
\mathbb{E}e^{t^* Z}\mathbb{E}e^{-t^* Y} &= e^{-I}, \\
\left(\mathbb{E}e^{t^* Z}\right)^{\frac{1}{2}} \left(\mathbb{E}e^{-t^* Y}\right)^{-\frac{1}{2}} e^{-t^*\lambda^*} &= 1, \\
\mathbb{E}e^{2t^* Z}\mathbb{E}e^{-2t^* Y} &= 1, \\
\left(\mathbb{E}e^{2t^* Z}\right)^{\frac{1}{2}} \left(\mathbb{E}e^{-2t^* Y}\right)^{-\frac{1}{2}} e^{-2t^*\lambda^*} &= 1.
\end{aligned}
$$

The proof is given in Section C.8.6.

C.8.3. CONTROLLING Q_2

We have

$$
\begin{aligned}
Q_2 &= \mathbb{E}\exp\left(t^* \sum_{j\in[m_q]} (Z_j - \lambda^*) - t^* \sum_{j\in[m_p]} (Y_j - \lambda^*)\right) \\
&= e^{-t^*\lambda^*(m_q - m_p)} \left(\mathbb{E}e^{t^* Z_1}\right)^{m_q} \left(\mathbb{E}e^{-t^* Y_1}\right)^{m_p} \\
&= \left(\mathbb{E}e^{t^* Z_1}\mathbb{E}e^{-t^* Y_1}\right)^{\frac{1}{2}m_p + \frac{1}{2}m_q} \left(\left(\frac{\mathbb{E}e^{t^* Z_1}}{\mathbb{E}e^{-t^* Y_1}}\right)^{\frac{1}{2}} e^{-t^*\lambda^*}\right)^{m_q - m_p}.
\end{aligned}
$$

By Fact 8, we can continue to write

$$
Q_2 \leq \exp\left(-\left(\frac{1}{2}m_p + \frac{1}{2}m_q\right) I\right)
$$

$$\leq \exp\left(-\frac{1}{2}(mn-m^2)I\right)$$
$$\leq \exp\left(-(1-\eta)\frac{mn}{2}I\right)$$

where the last step holds since $m \leq M \leq C\sqrt{\frac{n}{I}} = n\eta$.

C.8.4. CONTROLLING Q_3

Similar to controlling Q_2, we compute

$$\begin{aligned}
Q_3 &= \mathbb{E}\exp\left(2t^*\sum_{j\in[m_q']}(Z_j'-\lambda^*) - 2t^*\sum_{j\in[m_p']}(Y_j'-\lambda^*)\right)\\
&= e^{-2t^*\lambda^*(m_q'-m_p')}\left(\mathbb{E}e^{2t^*Z_1'}\right)^{m_q'}\left(\mathbb{E}e^{-2t^*Y_1'}\right)^{m_p'}\\
&= \left(\mathbb{E}e^{2t^*Z_1'}\mathbb{E}e^{-2t^*Y_1'}\right)^{\frac{1}{2}m_p'+\frac{1}{2}m_q'}\left(\left(\frac{\mathbb{E}e^{2t^*Z_1'}}{\mathbb{E}e^{-2t^*Y_1'}}\right)^{\frac{1}{2}}e^{-2t^*\lambda^*}\right)^{m_q'-m_p'}\\
&= 1
\end{aligned}$$

where the last step holds by by Fact 8.

C.8.5. PUTTING TOGETHER

We have

$$\begin{aligned}
P_{m,\mathcal{M}} &\leq \exp\left((1-2\eta)\frac{mn}{2}I - (1+\eta)m\log\left(\frac{ne}{m}\right)\right)\cdot\exp\left(-(1-\eta)\frac{mn}{2}I\right)\cdot 1\\
&= \exp\left(-(1+\eta)m\log\left(\frac{ne}{m}\right) - \eta\frac{mn}{2}I\right)
\end{aligned}$$

and by the union bound

$$\begin{aligned}
P_m &\leq \binom{n}{m}\exp\left(-(1+\eta)m\log\left(\frac{ne}{m}\right) - \eta\frac{mn}{2}I\right)\\
&\overset{(a)}{\leq} \exp\left(-\eta m\log\left(\frac{ne}{m}\right) - \eta\frac{mn}{2}I\right)\\
&= \left[\exp\left(-C\sqrt{\frac{1}{nI}}\log\left(\frac{ne}{m}\right) - \frac{C}{2}\sqrt{nI}\right)\right]^m
\end{aligned}$$

where step (a) holds by $\binom{n}{m} \leq \left(\frac{en}{m}\right)^m$ and the last step holds by the definition of η. Note that if $m \leq \sqrt{n}$, then

$$C\sqrt{\frac{1}{nI}}\log\left(\frac{ne}{m}\right) + \frac{C}{2}\sqrt{nI} \geq C\sqrt{\frac{1}{2}\log\left(\frac{ne}{m}\right)} \geq \sqrt{\log n}$$

where the last step holds since $C > 10$. We also have $P_m \leq \left[\exp\left(-\sqrt{\log n}\right)\right]^m < \frac{1}{2}$. If $m > \sqrt{n}$, then

$$C\sqrt{\frac{1}{nI}}\log\left(\frac{ne}{m}\right) + \frac{C}{2}\sqrt{nI} \geq \frac{C}{2}\sqrt{nI} \geq \log 10$$

by choosing C_I large enough and hence $P_m \le [\exp(-\log 10)]^m \le \frac{1}{10^m}$. It follows that

$$\begin{aligned}
P &\le \sum_{m\in[M]} P_m \\
&\le \sum_{1\le m\le\sqrt{n}} P_m + \sum_{\sqrt{n}<m\le n} P_m \\
&\le \sum_{1\le m<\infty} \left[\exp\left(-\sqrt{\log n}\right)\right]^m + n\cdot\frac{1}{10^{\sqrt{n}}} \\
&\le \frac{\exp\left(-\sqrt{\log n}\right)}{1-\exp\left(-\sqrt{\log n}\right)} + \frac{1}{n} \\
&\le 2\exp\left(-\sqrt{\log n}\right) + \exp(-\log n) \\
&\le 3\exp\left(-\sqrt{\log n}\right)
\end{aligned}$$

as desired.

C.8.6. PROOF OF FACT 8

The first equation can be shown by

$$\begin{aligned}
\mathbb{E}e^{t^*Z}\mathbb{E}e^{-t^*Y} &= \left(qe^{t^*}+1-q\right)\left(pe^{-t^*}+1-p\right) \\
&= pq+(1-p)(1-q)+q(1-p)e^{t^*}+p(1-q)pe^{-t^*} \\
&= pq+(1-p)(1-q)+2\sqrt{pq(1-p)(1-q)} \\
&= \left(\sqrt{pq}+\sqrt{(1-p)(1-q)}\right)^2 \\
&= e^{-I}
\end{aligned}$$

Note that $e^{2t^*\lambda^*} = \frac{1-q}{1-p}$. For the second equation, we compute

$$\begin{aligned}
\frac{\mathbb{E}e^{t^*Z}}{\mathbb{E}e^{-t^*Y}}e^{-2t^*\lambda^*} &= \left(\frac{qe^{t^*}+1-q}{pe^{-t^*}+1-p}\right)\left(\frac{1-p}{1-q}\right) \\
&= \left(\frac{q\sqrt{\frac{p(1-q)}{q(1-p)}}+1-q}{p\sqrt{\frac{q(1-p)}{p(1-q)}}+1-p}\right)\left(\frac{1-p}{1-q}\right) \\
&= 1
\end{aligned}$$

and take square roots on both sides. Finally, the remaining equations are combinations of the following key identities:

$$\begin{aligned}
\mathbb{E}e^{2t^*Z} &= qe^{2t^*}+1-q \\
&= q\frac{p(1-q)}{q(1-p)}+1-q \\
&= \frac{p(1-q)}{1-p}+\frac{(1-p)(1-q)}{1-p}
\end{aligned}$$

$$= \frac{1-q}{1-p}$$

and

$$\begin{aligned}\mathbb{E}e^{-2t^*Y} &= pe^{-2t^*} + 1 - p \\ &= p\frac{q(1-p)}{p(1-q)} + 1 - p \\ &= \frac{q(1-p)}{1-q} + \frac{(1-q)(1-p)}{1-q} \\ &= \frac{1-p}{1-q}\end{aligned}$$

and $e^{2t^*\lambda^*} = \frac{1-q}{1-p}$.

C.9. Proof of Lemma 3

Since $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A})$, we have

$$\begin{aligned}0 &\le \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} \right\rangle \\ &\overset{(a)}{=} \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \frac{p+q}{2}(\mathbf{J} - \mathbf{I}) \right\rangle \\ &= \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbb{E}\mathbf{A} - \frac{p+q}{2}(\mathbf{J} - \mathbf{I}) \right\rangle + \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A} \right\rangle\end{aligned}$$

where step (a) holds since $\left\langle \mathbf{J}, \widehat{\mathbf{Y}} \right\rangle = \langle \mathbf{J}, \mathbf{Y}^* \rangle$ and $\widehat{Y}_{ii} = Y^*_{ii}$. Noting that

$$\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbb{E}\mathbf{A} - \frac{p+q}{2}(\mathbf{J} - \mathbf{I}) \right\rangle = -\frac{p-q}{2}\gamma,$$

we have the bound $\gamma \le \frac{2}{p-q} \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A} \right\rangle$. To control the RHS, we compute

$$\begin{aligned}\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A} \right\rangle &= \left\langle \widehat{\mathbf{Y}}, \mathbf{A} - \mathbb{E}\mathbf{A} \right\rangle - \langle \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A} \rangle \\ &\le 2 \sup_{\mathbf{Y} \succeq 0, \mathrm{diag}(\mathbf{Y}) \le \mathbf{1}} |\langle \mathbf{Y}, \mathbf{A} - \mathbb{E}\mathbf{A} \rangle| .\end{aligned}$$

The Grothendieck's inequality (Grothendieck, 1953; Lindenstrauss and Pełczyński, 1968) guarantees that

$$\sup_{\mathbf{Y} \succeq 0, \mathrm{diag}(\mathbf{Y}) \le \mathbf{1}} |\langle \mathbf{Y}, \mathbf{A} - \mathbb{E}\mathbf{A} \rangle| \le K_G \|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\infty \to 1}$$

where K_G denotes the Grothendieck's constant ($0 < K_G \le 1.783$) and $\|\mathbf{M}\|_{\infty \to 1} \coloneqq \sup_{\mathbf{x}:\|\mathbf{x}\|_\infty \le 1} \|\mathbf{M}\mathbf{x}\|_1$ is the $\ell_\infty \to \ell_1$ operator norm for a matrix $\mathbf{M}$. Furthermore, we have the identity

$$\|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\infty \to 1} = \sup_{\mathbf{x}:\|\mathbf{x}\|_\infty \le 1} \|(\mathbf{A} - \mathbb{E}\mathbf{A})\mathbf{x}\|_1 = \sup_{\mathbf{y},\mathbf{z} \in \{\pm 1\}^n} \left| \mathbf{y}^\top (\mathbf{A} - \mathbb{E}\mathbf{A})\mathbf{z} \right| .$$

Set $v^2 := \sum_{1\le i<j\le n} \mathrm{Var}(A_{ij})$. For each pair of fixed vectors $\mathbf{y}, \mathbf{z} \in \{\pm 1\}^n$, the Bernstein inequality ensures that for each number $t \ge 0$,

$$\Pr\left\{\left|\mathbf{y}^\top(\mathbf{A}-\mathbb{E}\mathbf{A})\mathbf{z}\right| > t\right\} \le 2\exp\left\{-\frac{t^2}{2v^2+4t/3}\right\}.$$

Setting $t = \sqrt{16nv^2} + \frac{8}{3}n$ gives

$$\Pr\left\{\left|\mathbf{y}^\top(\mathbf{A}-\mathbb{E}\mathbf{A})\mathbf{z}\right| > \sqrt{16nv^2} + \frac{8}{3}n\right\} \le 2e^{-2n}.$$

Applying the union bound and using the fact that $v^2 \le p(n^2-n)/2$, we obtain that with probability at most $2^{2n}\cdot 2e^{-2n} = 2(e/2)^{-2n}$,

$$\|\mathbf{A}-\mathbb{E}\mathbf{A}\|_{\infty\to 1} > 2\sqrt{2p(n^3-n^2)} + \frac{8}{3}n.$$

Combining pieces, we conclude that with probability at least $1-2(e/2)^{-2n}$,

$$\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A}-\mathbb{E}\mathbf{A}\right\rangle \le 8\sqrt{2p(n^3-n^2)} + \frac{32}{3}n;$$

whence

$$\begin{aligned}\gamma &\le \frac{2}{p-q}\left(8\sqrt{2p(n^3-n^2)} + \frac{32}{3}n\right)\\ &\overset{(a)}{\le} \frac{45\sqrt{pn^3}}{p-q}\\ &\le \frac{45}{C'}\sqrt{\frac{n^3}{I}},\end{aligned}$$

for some constant $C' > 0$, where step (a) holds by our assumption $p \ge \frac{1}{n}$ and the last step follows from Fact 5.

C.10. Proof of Lemma 4

If $\phi_i \ge 0$ for all $i \in [\lceil u' \rceil]$, then the result follows immediately. Now we assume that at least one of $\{\phi_i\}$ is negative. Define $w := \arg\min\{i \in [\lceil u' \rceil] : \phi_i < 0\}$ be the smallest index of negative ϕ_i. Then for $u_0 \in [1, w-1]$, we have $0 \le V(u_0)$ as $\phi_i \ge 0$ for all $i \in [1, w-1]$. On the other hand, it is not hard to see that V is decreasing on $[w-1, u']$ since $\phi_i < 0$ for all $i \in [w, \lceil u' \rceil]$. In view of our assumption $0 \le V(u')$, the proof is completed.

Appendix D. Proof of the second inequality in Theorem 2

Observe that for any $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A})$, we have

$$\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 = \sum_{i,j\in[n]} (\widehat{Y}_{ij} - Y^*_{ij})^2$$

$$\overset{(a)}{\leq} \sum_{i,j\in[n]} 2\left|\widehat{Y}_{ij} - Y^*_{ij}\right|$$
$$= 2\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$$

where step (a) holds since $\widehat{\mathbf{Y}}, \mathbf{Y}^* \in [-1,1]^{n\times n}$ by feasibility to the program (6). Then combined with the first inequality of Theorem 2, this implies

$$\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 \leq 2n^2 \exp\left[-\left(1 - C'_e\sqrt{\frac{1}{nI}}\right)\frac{nI}{2}\right]$$
$$\leq n^2 \exp\left[-\left(1 - C_e\sqrt{\frac{1}{nI}}\right)\frac{nI}{2}\right]$$

for some constants $C_e > C'_e > 0$.

Let $\varepsilon := \exp\left[-\left(1 - C_e\sqrt{\frac{1}{nI}}\right)\frac{nI}{2}\right]$ and $\hat{\mathbf{v}}$ be an eigenvector of $\widehat{\mathbf{Y}}$ corresponding to the largest eigenvalue with $\|\hat{\mathbf{v}}\|_2^2 = n$. It can seen that $\boldsymbol{\sigma}^*$ is an eigenvector of $\mathbf{Y}^*$ corresponding to the largest eigenvalue. We claim that

$$\min_{g\in\{\pm 1\}} \|g\hat{\mathbf{v}} - \boldsymbol{\sigma}^*\|_2^2 \leq C^2\varepsilon n$$

for an absolute constant $C > 0$. This follows from Davis-Kahan theorem. To see this, note that the largest eigenvalue of $\mathbf{Y}^*$ is n and all the others are equal to 0, so the eigengap is n. Because $\widehat{\mathbf{Y}} = \mathbf{Y}^* + (\widehat{\mathbf{Y}} - \mathbf{Y}^*)$ and $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_F \leq \sqrt{\varepsilon}n$, David-Kahan theorem (see, for example, Corollary 3 in Vu (2011)) implies that

$$\min_{g\in\{\pm 1\}} \|g\hat{\mathbf{u}} - \mathbf{u}^*\|_2 = 2\left|\sin\left(\frac{\theta}{2}\right)\right| \leq C\sqrt{\varepsilon}$$

where $\hat{\mathbf{u}}$ and $\mathbf{u}^*$ denote the unit-norm eigenvectors associated to the largest eigenvalues of $\widehat{\mathbf{Y}}$ and $\mathbf{Y}^*$, respectively, and $\theta \in [0, \frac{\pi}{2}]$ denotes the angle between these two vectors. By definition $\hat{\mathbf{v}} = \sqrt{n}\hat{\mathbf{u}}$ and $\boldsymbol{\sigma}^* = \sqrt{n}\mathbf{u}^*$, we have the desired result.

Finally, we relate $\mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\text{sdp}}, \boldsymbol{\sigma}^*)$ to ε via the quantity $\min_{g\in\{\pm1\}} \|g\mathbf{v} - \boldsymbol{\sigma}^*\|_2^2$. Let $g^* := \arg\min_{g\in\{\pm1\}} \|g\mathbf{v} - \boldsymbol{\sigma}^*\|_2^2$ and WLOG we assume $g^* = 1$. By definition we have $\widehat{\sigma}_i^{\text{sdp}} = \text{sign}(\hat{v}_i)$. It can be seen that

$$\begin{aligned} C^2\varepsilon n &\geq \|\hat{\mathbf{v}} - \boldsymbol{\sigma}^*\|_2^2 \\ &= \sum_{i\in[n]} (\hat{v}_i - \sigma_i^*)^2 \\ &\geq \sum_{i\in[n]} (\hat{v}_i - \sigma_i^*)^2 \mathbb{I}\{\text{sign}(\hat{v}_i) \neq \sigma_i^*\} \\ &\geq \sum_{i\in[n]} \mathbb{I}\{\text{sign}(\hat{v}_i) \neq \sigma_i^*\} \\ &\geq n \cdot \mathtt{err}(\widehat{\boldsymbol{\sigma}}^{\text{sdp}}, \boldsymbol{\sigma}^*). \end{aligned}$$

We divide both sides of the above equation by n, and note that the constant C^2 can be absorbed into C_e under the assumption that $nI \geq C_I$ for C_I sufficiently large. The result follows.

Appendix E. Proof of Theorems 3 and 4

We first prove Theorem 3 for the semirandom Model 2. Recall that $\mathbf{A}^{\mathrm{SR}}$ is the observed adjacency matrix generated from Model 2, and take $\widehat{\mathbf{Y}}$ to be an arbitrary element of $\mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$. By definition of $\mathbf{A}^{\mathrm{SR}}$ and the feasibility of $\widehat{\mathbf{Y}}$, we have for all $i, j \in [n]$

$$\begin{cases} A_{ij}^{\mathrm{SR}} \geq A_{ij}, \widehat{Y}_{ij} - Y_{ij}^* \leq 0, & \text{if } Y_{ij}^* = 1, \\ A_{ij}^{\mathrm{SR}} \leq A_{ij}, \widehat{Y}_{ij} - Y_{ij}^* \geq 0, & \text{if } Y_{ij}^* = -1. \end{cases}$$

The fact that $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$, together with the above displayed equation, implies that $0 \leq \left\langle \mathbf{A}^{\mathrm{SR}}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \leq \left\langle \mathbf{A}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle$. This further implies that $\widehat{\mathbf{Y}} \in \mathcal{Y}(\mathbf{A})$. Therefore, invoking Theorem 2 gives the desired result.

We next prove Theorem 4 for the heterogeneous Model 3. To this end, we show that Model 3 can be reduced to Model 2, by applying a coupling argument similar to that in Appendix V in Fei and Chen (2019b). Here we include the proof for completeness. Recall that for each $i < j$, we have $A_{ij}^{\mathrm{H}} \sim \mathrm{Bern}(p_{ij})$ if $Y_{ij}^* = 1$, $A_{ij}^{\mathrm{H}} \sim \mathrm{Bern}(q_{ij})$ if $Y_{ij}^* = -1$ and $A_{ij}^{\mathrm{H}} = A_{ji}^{\mathrm{H}}$, where we assume $p_{ij} \geq p$ and $q_{ij} \leq q$. We claim that such a $\mathbf{A}^{\mathrm{H}}$ can also be generated by the following 3-step process:

1. We first construct a set of node pairs $\mathcal{L}$ as follows: independently for each $i < j$, if $Y_{ij}^* = 1$ we include (i, j) in $\mathcal{L}$ with probability $1 - \frac{1-p_{ij}}{1-p}$, and if $Y_{ij}^* = -1$ we include (i, j) in $\mathcal{L}$ with probability $1 - \frac{q_{ij}}{q}$;
2. Independently of above, we sample a graph adjacency matrix $\mathbf{A}$ from Model 1;
3. The final graph adjacency matrix $\mathbf{A}^{\mathrm{H}}$ is constructed as follows: for each $i < j$, if $Y_{ij}^* = 1$ we let $A_{ij}^{\mathrm{H}} = \mathbb{I}\{A_{ij} = 1 \text{ or } (i, j) \in \mathcal{L}\}$, and if $Y_{ij}^* = -1$ we let $A_{ij}^{\mathrm{H}} = \mathbb{I}\{A_{ij} = 1 \text{ and } (i, j) \notin \mathcal{L}\}$.

Note that the assumption $p_{ij} \geq p$ and $q_{ij} \leq q$ ensures that the probabilities in step 1 are in $[0, 1]$. We verify that the distribution of $\mathbf{A}^{\mathrm{H}}$ generated above is the same as that of the adjacency matrix from Model 3 as claimed. Indeed, for each $i < j$ we have

$$\begin{aligned} \mathbb{P}\left(A_{ij}^{\mathrm{H}} = 1\right) &= \begin{cases} \mathbb{P}\left(A_{ij} = 1 \text{ or } (i, j) \in \mathcal{L}\right), & \text{if } Y_{ij}^* = 1, \\ \mathbb{P}\left(A_{ij} = 1 \text{ and } (i, j) \notin \mathcal{L}\right), & \text{if } Y_{ij}^* = -1. \end{cases} \\ &= \begin{cases} 1 - (1 - p) \cdot \frac{1-p_{ij}}{1-p} = p_{ij}, & \text{if } Y_{ij}^* = 1, \\ q \cdot \frac{q_{ij}}{q} = q_{ij}, & \text{if } Y_{ij}^* = -1. \end{cases} \end{aligned}$$

On the other hand, conditioned on the set $\mathcal{L}$, the distribution of $\mathbf{A}^{\mathrm{H}}$ is identical to that of $\mathbf{A}^{\mathrm{SR}}$ from Model 2 with the same $\mathcal{L}$, since step 2 is independent of step 1 and in step 3 the matrix $\mathbf{A}$ is modified monotonically to produce $\mathbf{A}^{\mathrm{H}}$. This means that $\mathcal{Y}(\mathbf{A}^{\mathrm{H}}) = \mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$ conditioned on $\mathcal{L}$. But we have established above that the error bounds in Theorem 2 continue to hold for $\mathcal{Y}(\mathbf{A}^{\mathrm{SR}})$, and hence also for $\mathcal{Y}(\mathbf{A}^{\mathrm{H}})$ when conditioned on $\mathcal{L}$. Integrating out the randomness of the set $\mathcal{L}$, we see that the same error bounds hold for $\mathcal{Y}(\mathbf{A}^{\mathrm{H}})$ unconditionally.