Exponential Error Rates of SDP for Block Models: Beyond Grothendieck's Inequality

Yingjie Fei and Yudong Chen

***Abstract*— In this paper, we consider the cluster estimation problem under the stochastic block model. We show that the semidefinite programming (SDP) formulation for this problem achieves an error rate that decays exponentially in the signal-to-noise ratio. The error bound implies weak recovery in the sparse graph regime with bounded expected degrees as well as exact recovery in the dense regime. An immediate corollary of our results yields error bounds under the censored block model. Moreover, these error bounds are robust, continuing to hold under heterogeneous edge probabilities and a form of the so-called monotone attack. Significantly, this error rate is achieved by the SDP solution itself without any further pre- or post-processing and improves upon existing polynomially decaying error bounds proved using the Grothendieck's inequality. Our analysis builds on two key ingredients: 1) showing that the graph has a well-behaved spectrum, even in the sparse regime, after discounting an exponentially small number of edges and 2) an order-statistics argument that governs the final error rate. Both arguments highlight the implicit regularization effect of the SDP formulation.**

***Index Terms*— Stochastic block models, semidefinite programming, convex relaxation, exponential rates.**

I. Introduction

In this paper, we consider the cluster/community[1] estimation problem under the Stochastic Block Model (SBM) [1] with a growing number of clusters. In this model, a set of n nodes are partitioned into k unknown clusters of equal size; a random graph is generated by independently connecting each pair of nodes with probability p if they are in the same cluster, and with probability q otherwise. Given one realization of the graph represented by its adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$, the goal is to estimate the underlying clusters.

Much recent progress has been made on this problem, particularly in identifying the precise conditions for exact/weak recovery when there are a few communities of size linear in n. Moving beyond this regime, however, the understanding of the problem is far more limited, especially in terms of characterizing its behaviors with a growing (in n) number of clusters of sublinear sizes, and how the estimation errors depend on the model parameters in between the exact and weak recovery regimes [2], [3]. We focus on precisely these questions.

Let the ground-truth clusters be encoded by a *cluster matrix* $\mathbf{Y}^* \in \{0,1\}^{n\times n}$ defined as

$$Y^*_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are in the same community,} \\ 0 & \text{if nodes } i \text{ and } j \text{ are in different communities,} \end{cases}$$

and we use the convention that $Y^*_{ii} = 1$ for all $i \in [n]$. We consider a (now standard) semidefinite programming (SDP) approach for estimating the ground-truth $\mathbf{Y}^*$:

$$\begin{aligned} \widehat{\mathbf{Y}} = \underset{\mathbf{Y}\in\mathbb{R}^{n\times n}}{\arg\max} & \left\langle \mathbf{Y}, \mathbf{A} - \frac{p+q}{2}\mathbf{J} \right\rangle \\ \text{s.t. } & \mathbf{Y} \succeq 0,\ 0 \le \mathbf{Y} \le \mathbf{J}, \\ & Y_{ii} = 1, \forall i \in [n], \end{aligned} \tag{1}$$

where $\mathbf{J}$ is the $n \times n$ all-one matrix and $\langle \cdot, \cdot \rangle$ denotes the trace inner product. We seek to characterize the accuracy of the SDP solution $\widehat{\mathbf{Y}}$ as an estimator of the true clustering. Our main focus is the ℓ_1 error $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$, where $\|\mathbf{M}\|_1 := \sum_{i,j} |M_{ij}|$ denotes the entry-wise ℓ_1 norm. This ℓ_1 error is a natural metric that measures a form of pairwise cluster/link errors. In particular, note that the matrix $\mathbf{Y}^*$ encodes the cluster relationship between each pair of nodes; an estimator of such is given by the matrix $\widehat{\mathbf{Y}}^{\mathrm{R}} \in \{0,1\}^{n\times n}$ obtained from rounding $\widehat{\mathbf{Y}}$ element-wise. The above ℓ_1 error satisfies $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \ge \left|\{(i,j) : \widehat{Y}^{\mathrm{R}}_{ij} \ne Y^*_{ij}\}\right|/2$, and therefore upper bounds the number of pairs whose relationships are incorrectly estimated by the SDP.

In a seminal paper, Guédon and Vershynin [4] exhibited a remarkable use of the Grothendieck's inequality, and obtained the following high-probability error bound for the SDP solution $\widehat{\mathbf{Y}}$:

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1} \lesssim \sqrt{\frac{k^2}{\mathrm{SNR}\cdot n}}. \tag{2}$$

Here $\mathrm{SNR} = (p-q)^2/\left(\frac{1}{k}p + (1-\frac{1}{k})q\right) \approx (p-q)^2/p$ is a measure of the signal-to-noise ratio. This bound holds even in the sparse graph regime with constant expected degrees, namely $p, q = \Theta(1/n)$, which is a manifest of the power of the Grothendieck's inequality.

Manuscript received July 11, 2017; accepted April 9, 2018. Date of publication May 22, 2018; date of current version December 19, 2018. This work was supported in part by the National Science Foundation under Grants 1657420 and CCF-1704828 and in part by the School of Operations Research and Information Engineering, Cornell University.

The authors are with the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850 USA (e-mail: yf275@cornell.edu; yudong.chen@cornell.edu).

Communicated by E. Abbe, Associate Editor for Machine Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2839677

[1]The words *cluster* and *community* are used interchangeably in this paper.

In this paper, we go beyond the above results, and show that $\widehat{\mathbf{Y}}$ in fact satisfies (with high probability) the following exponentially-decaying error bound

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1} \lesssim \exp\left[-\Omega\left(\mathrm{SNR} \cdot \frac{n}{k}\right)\right] \tag{3}$$

as long as $\mathrm{SNR} \gtrsim \frac{k^2}{n}$ (Theorem 1). The bound is valid in both the sparse and dense regimes. Significantly, this error rate is achieved by the SDP (1) itself, without the need of a multi-step procedure, even though we are estimating a discrete structure by solving a continuous optimization problem. In particular, the SDP approach does not require pre-processing of graph (such as trimming and splitting) or an initial estimate of the clusters, nor any non-trivial post-processing of $\widehat{\mathbf{Y}}$ (such as local cluster refinement or randomized rounding).

If an explicit clustering of the nodes is concerned, the result above also yields an error bound for estimating $\boldsymbol{\sigma}^*$, the true cluster labels. In particular, an explicit cluster labeling $\widehat{\boldsymbol{\sigma}}$ can be obtained efficiently from $\widehat{\mathbf{Y}}$. Let $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$ denote the fraction of nodes that are labeled differently by $\widehat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}^*$ (after accounting for permutation of the labels). This mis-classification error can be shown to be bounded from above by the ℓ_1 error $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1/\|\mathbf{Y}^*\|_1$, and therefore satisfies the same exponential bound (Theorem 2):

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \lesssim \exp\left[-\Omega\left(\mathrm{SNR} \cdot \frac{n}{k}\right)\right]. \tag{4}$$

When specialized to different values of the errors, this single error bound (3) implies sufficient conditions for achieving exact recovery (strong consistency), almost exact recovery (weak consistency) and weak recovery; see Section I-B for the definitions of these recovery types. More generally, the above bound yields SNR conditions sufficient for achieving any δ error. As to be discussed in details in Section III-A.1, these conditions are (at least order-wise) optimal, and improve upon existing results especially when the number of clusters k is allowed to scale with n. In addition, we prove that the above guarantees for SDP are *robust* against deviations from the standard SBM: the same exponential bounds continue to hold in the presence of heterogeneous edge probabilities as well as a form of monotone attack where an adversary can modify the graph (Theorem 3). Furthermore, our results readily extend to the *Censored Block Model*, in which only partially observed data are available (Corollary 1).

In addition to providing improved error bounds, our results also involve the development of several new analytical techniques, as are discussed below. We expect these techniques to be broadly useful in the analysis of SDP and other algorithms for SBM and related statistical problems.

A. Technical Highlights

Our analysis of the SDP formulation builds on two key ingredients. The first ingredient involves showing that the graph can be partitioned into two components: one with a well-behaved spectrum, and a sparse residual with an *exponentially* small number of edges; cf. Proposition 2. Note that this partitioning is done *in the analysis*, rather than in the algorithm. This argument ensures that the SDP produces a useful solution all the way down to the sparse regime with $p, q = \Theta(\frac{1}{n})$. The second ingredient is an order-statistics argument that characterizes the interplay between the error matrix and the randomness in the graph; cf. Proposition 1. This argument establishes a connection between the estimation error and the sum of the top order statistics of certain appropriately constructed random variables; upper bounds on this sum are what ultimately dictate the exponential decay of the error. In both arguments, we make crucial use of the entry-wise boundedness of the SDP solution $\widehat{\mathbf{Y}}$, which is a manifest of the implicit regularization effect of the SDP formulation.

Our results are non-asymptotic in nature, valid for finite values of n; letting $n \to \infty$ gives asymptotic results. All the parameters p, q and k are allowed to scale arbitrarily with n. In particular, the number of clusters k may grow with n, the clusters may have size sublinear in n, and the edge probabilities p and q may range from the sparse case $\Theta(\frac{1}{n})$ to the dense case $\Theta(1)$. Our results therefore provide a general characterization of the relationship between the SNR, the cluster sizes and the recovery errors. This point is particularly important in the regime of sublinear cluster sizes, in which case all values of p and q are of interest. The price of such generality is that we do not seek to obtain optimal values of the multiplicative constants in the error bounds — doing so typically requires asymptotic analysis with restrictions on the scaling of the parameters. In this sense, our results complement the recent work on the fundamental limits and sharp recovery thresholds of SBM [2].

B. Related Work

The Stochastic Block Model [1], [5], also known as the Planted Partition Model in the computer science literature, is a standard model for studying graph clustering and community detection in networks. There is a large body of work on the theoretical and algorithmic aspects of this model; see for example [2], [6]–[8], and the references therein. Here we briefly discuss the most relevant work, and defer to Section III for a more detailed comparison after stating our main theorems.

Existing work distinguishes between several types of recovery [2], [9], including: (a) weak recovery, where the fraction of mis-clustered nodes satisfies $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) < 1 - \frac{1}{k}$ and is hence better than random guess; (b) almost exact recovery (weak consistency), where $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = o(1)$; (c) exact recovery (strong consistency), where $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = 0$. Here a fundamental question is identifying the thresholds for the SNR above which different types of recovery can be achieved with high probability. The SDP relaxation approach to SBM has been studied in [7] and [10]–[16], which mostly focus on exact recovery in the logarithmic-degree regime $p = \Omega\left(\frac{\log n}{n}\right)$. Using the Grothendieck's inequality, the work in [4] proves for the first time that SDP achieves a non-trivial error bound in the sparse regime $p = \Theta(\frac{1}{n})$ with bounded expected degrees. In the two-cluster case, it is further shown in [17] that SDP in fact achieves the optimal weak recovery threshold as long as the expected degree is large (but still bounded). Our error

bound implies exact and weak recovery in the logarithmic and bounded degree regimes, respectively. Our result in fact goes beyond these existing ones and applies to every setting in between the two extreme regimes, capturing the exponential decay of error rates from $O(1)$ to zero.

A very recent line of research aims to precisely characterize the fundamental limits and phase transition behaviors of SBM — in particular, what are the sharp SNR thresholds, including the leading constants, for achieving the recovery types discussed in the last paragraph. When the number k of clusters is bounded, many of these questions now have satisfactory answers. Without attempting to exhaust this still growing line of remarkable work, we refer to [8] and [18]–[21] for weak recovery, [10] and [22]–[25] for almost exact recovery, and [22], [23], and [26] for exact recovery. SDP has in fact been shown to achieve the optimal exact recovery threshold [27]–[30]. Our results imply sufficient conditions for SDP achieving these various types of recovery, and moreover interpolate between them. As mentioned, we are mostly concerned with the non-asymptotic setting with a growing number of clusters, and do not attempt to optimize the values of the leading constants. Therefore, we have focused on somewhat different regimes than the work above.

Particularly relevant to us is the work in [4], [23]–[25], and [31]–[33], which provides explicit bounds on the error rates of alternative algorithms for estimating the ground-truth clustering in SBM. The Censored Block Model is studied in [31] and [34]–[38]. Robustness issues in SBM are considered in the work in [14], [28], [33], and [39]–[44]. We discuss these results in more details in Section III.

C. Notations

Column vectors are denoted by lower-case bold letters such as $\mathbf{u}$, where u_i is its i-th entry. Matrices are denoted by bold capital letters such as $\mathbf{M}$, with $\mathbf{M}^\top$ denoting the transpose of $\mathbf{M}$, $\mathrm{Tr}(\mathbf{M})$ its trace, M_{ij} its (i,j)-th entry, and $\mathrm{diag}(\mathbf{M})$ the vector of its diagonal entries. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|_1 := \sum_{i,j} |M_{ij}|$ is its entry-wise ℓ_1 norm, $\|\mathbf{M}\|_\infty := \max_{i,j} |M_{ij}|$ the entry-wise ℓ_∞ norm, and $\|\mathbf{M}\|_{\mathrm{op}}$ the spectral norm (the maximum singular value). Denote by $\mathbf{M}_{i\bullet}$ the i-th row of the matrix $\mathbf{M}$ and $\mathbf{M}_{\bullet j}$ its j-th column. We write $\mathbf{M} \succeq 0$ if $\mathbf{M}$ is symmetric and positive semidefinite. For two matrices $\mathbf{M}$ and $\mathbf{G}$ of the same dimension, we let $\langle \mathbf{M}, \mathbf{G} \rangle := \mathrm{Tr}(\mathbf{M}^\top \mathbf{G})$ denote their trace inner product, and use $\mathbf{M} \geq \mathbf{G}$ to mean that $M_{ij} \geq G_{ij}$ for all $i, j \in [n]$. Let $\mathbf{I}$ and $\mathbf{J}$ be the $n \times n$ identity matrix and all-one matrix, respectively, and $\mathbf{1}$ the all-one column vector of length n.

We use $\mathrm{Bern}(\mu)$ to denote the Bernoulli distribution with rate $\mu \in [0,1]$. For a positive integer i, let $[i] := \{1, 2, \ldots, i\}$. For a real number x, $\lceil x \rceil$ denotes its ceiling. Throughout the paper, a universal constant C means a fixed number that is independent of the model parameters (n, k, p, q, etc.) and the graph distribution. We use the following standard notations for order comparison of 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 $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. We write $a_n = o(b_n)$ or $b_n = \omega(a_n)$ if $\lim_{n\to\infty} a_n / b_n = 0$.

II. Problem Setup

In this section, we formally set up the problem of cluster estimation under SBM and describe the SDP approach.

A. The Stochastic Block Model

Given n nodes, we assume that each node belongs to exactly one of k ground truth clusters, where the clusters have equal size n/k. Let $\boldsymbol{\sigma}^* \in [k]^n$ be the vector of ground-truth cluster labels, where σ_i^* is the index of the cluster that contains node i. (The cluster labels are unique only up to permutation; here $\boldsymbol{\sigma}^*$ is defined with respect to an arbitrary permutation.) This ground-truth can be equivalently encoded in the cluster matrix $\mathbf{Y}^* \in \{0,1\}^{n\times n}$ defined in Section I. We do not know $\boldsymbol{\sigma}^*$ or $\mathbf{Y}^*$, but we observe the adjacency matrix $\mathbf{A}$ of a graph generated from the following Stochastic Block Model (SBM).

Model 1 (Standard Stochastic Block Model): The graph adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ is symmetric with its entries $\{A_{ij}, i < j\}$ generated independently by

$$A_{ij} \sim \begin{cases} \mathrm{Bern}(p) & \text{if } Y^*_{ij} = 1, \\ \mathrm{Bern}(q) & \text{if } Y^*_{ij} = 0, \end{cases}$$

where $0 \leq q < p \leq 1$.

The values of the diagonal entries of $\mathbf{A}$ are inconsequential for the SDP formulation (1) due to the constraint $Y_{ii} = 1, \forall i$. Therefore, we may assume that $A_{ii} = 0$ for all $i \in [n]$, which simplifies the presentation of the analysis. The goal is to estimate $\mathbf{Y}^*$ given the observed graph $\mathbf{A}$.

Playing a crucial role in our results is the quantity

$$s := \frac{(p-q)^2}{\frac{1}{k}p + (1-\frac{1}{k})q}, \tag{5}$$

which is a measure of the SNR of the model. In particular, the numerator of s is the squared expected difference between the in- and cross-cluster edge probabilities, and the denominator is essentially the average variance of the entries of $\mathbf{A}$. The quantity s has been shown to capture the hardness of SBM, and defines the celebrated Kesten-Stigum threshold [21]. To avoid cluttered notation, we assume throughout the paper that $n \geq 4$, $2 \leq k < n$ and there exists a universal constant $0 < c < 1$ such that $q \geq cp$; this setting encompasses most interesting regimes of the problem, as clustering is more challenging when q is large.

B. Semidefinite Programming Relaxation

We consider the SDP formulation in (1), whose optimal solution $\widehat{\mathbf{Y}}$ serves as an estimator of the ground-truth cluster matrix $\mathbf{Y}^*$. This SDP can be interpreted as a convex relaxation for the maximum likelihood estimator, the modularity maximization problem, the optimal subgraph/cut problem, or a variant of the robust/sparse PCA problem; see [5], [7], [10], [12], [15] for such derivations. Our goal is to study the recovery error of $\widehat{\mathbf{Y}}$, and in particular how it depends on the

number of nodes n, the number of clusters k and the SNR measure s defined above.

Note that there is nothing special about the particular formulation in (1). All our results apply to, for example, the following alternative SDP formulation:

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

This formulation was previously considered in [7]. We may further replace the third constraint above with the row-wise constraints $\sum_{j=1}^{n} Y_{ij} = \sum_{j=1}^{n} Y^*_{ij}, \forall i \in [n]$, akin to the formulation in [10] motivated by weak assortative SBMs. Under the standard assumption of equal-sized clusters, the values $\sum_{i,j=1}^{n} Y^*_{ij} = n^2/k$ and $\sum_{j=1}^{n} Y^*_{ij} = n/k$ are fixed and known. Therefore, the formulation (6) has the advantage that it does not require knowledge of the edge probabilities p and q, but instead the number of clusters k.[2]

The optimization problems in (1) and (6) can be solved in polynomial time using any general-purpose SDP solvers or first-order algorithms. Moreover, this SDP approach continues to motivate, and benefit from, the rapid development of efficient algorithms for solving structured SDPs. For example, the algorithms considered in [42] and [45] can solve a problem involving $n = 10^5$ nodes within seconds on a laptop. In addition to computational efficiency, the SDP approach also enjoys several other desired properties including robustness, applicability to sparse graphs and conceptual simplicity, making it an attractive option among other clustering and community detection algorithms. The empirical performance of SDP has been extensively studied, both under SBM and with real data; see for example the work in [10], [15], [16], [42], and [45]. Here we focus on the theoretical guarantees of this SDP approach.

C. Explicit Clustering by k-Medians

After solving the SDP formulations (1) or (6), an estimate of the cluster membership can be extracted from the solution $\widehat{\mathbf{Y}}$. This can be done using many simple procedures. For example, when $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 < \frac{1}{2}$, simply rounding the entries of $\widehat{\mathbf{Y}}$ will exactly recover $\mathbf{Y}^*$, which immediately reveals the true clusters. In the case with $k = 2$ clusters, one may use the signs of the entries of first eigenvector of $\widehat{\mathbf{Y}} - \frac{1}{2}\mathbf{J}$, a procedure analyzed in [4] and [17] among others. More generally, our theoretical results guarantee that the SDP solution $\widehat{\mathbf{Y}}$ is already close to true cluster matrix $\mathbf{Y}^*$; in this case, we expect that many local rounding/refinement procedures, such as Lloyd's-style greedy algorithms [46], will be able to extract a high-quality clustering.

For the sake of retaining focus on the SDP formulation itself, we choose not to separately analyze these possible extraction procedures, but instead consider the following more unified approach. In particular, we view the rows of $\widehat{\mathbf{Y}}$ as n points in $\mathbb{R}^n$, and apply k-medians clustering to these points to find the clusters. While exactly solving the k-medians problem is computationally hard, there exist polynomial-time constant-factor approximation schemes, such as the $6\frac{2}{3}$-approximation algorithm in [47], which suffices for our purpose. This algorithm may not be the most efficient way to extract an explicit clustering from $\widehat{\mathbf{Y}}$; rather, it is intended as a simple venue for deriving a clustering error bound that can be readily compared with existing results.

Formally, we use $\rho\text{-}\mathtt{kmed}(\widehat{\mathbf{Y}})$ to denote a ρ-approximate k-median procedure applied to the rows of $\widehat{\mathbf{Y}}$; details are provided in Appendix I. The output

$$\widehat{\boldsymbol{\sigma}} := \rho\text{-}\mathtt{kmed}(\widehat{\mathbf{Y}})$$

is a vector in $[k]^n$ such that node i is assigned to the $\widehat{\sigma}_i$-th cluster by the procedure. We are interested in bounding the clustering error of $\widehat{\boldsymbol{\sigma}}$ relative to the ground-truth $\boldsymbol{\sigma}^*$. Let S_k denote the symmetric group consisting of all permutations of $[k]$; we consider the error metric

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) := \min_{\pi\in S_k} \frac{1}{n} \left|\left\{i \in [n] : \widehat{\sigma}_i \neq \pi(\sigma^*_i)\right\}\right|, \tag{7}$$

which is the proportion of nodes that are mis-classified, modulo permutations of the cluster labels.

Before proceeding, we briefly mention several possible extensions of the setting discussed above. The number $\frac{p+q}{2}$ in the SDP (1) can be replaced by a tuning parameter λ; as would become evident from the proof, our theoretical results in fact hold for an entire range of λ values, for example $\lambda \in [\frac{1}{4}p + \frac{3}{4}q, \frac{3}{4}p + \frac{1}{4}q]$. Our theory also generalizes to the setting with unequal cluster sizes; in this case the same theoretical guarantees hold with k replaced by $n/\underline{\ell}$, where $\underline{\ell}$ is any lower bound of the cluster sizes.

III. Main Results

We present in Section III-A our main theorems, which provide exponentially-decaying error bounds for the SDP formulation under SBM. We also discuss the consequences of our results, including their implications for robustness in Section III-B and applications to the Censored Block Model in Section III-C. In the sequel, $\widehat{\mathbf{Y}}$ denotes any optimal solution to the SDP formulation in either (1) or (6).

A. Error Rates Under Standard SBM

In this section, we consider the standard SBM setting in Model 1. Recall that n and k are respectively the numbers of nodes and clusters, and $\mathbf{Y}^*$ is the ground-truth cluster matrix (defined in Section I) with $\boldsymbol{\sigma}^*$ being the corresponding vector of true cluster labels. Our results are stated in terms of the SNR measure s given in equation (5).

The first theorem, proved in Section IV, shows that the SDP solution $\widehat{\mathbf{Y}}$ achieves an exponential error rate.

Theorem 1 (Exponential Error Rate): Under Model 1, there exist universal constants $C_s, C_g, C_e > 0$ for which the

[2]Note that the constraint $\mathbf{Y} \le \mathbf{J}$ in the formulations (1) and (6) is in fact redundant, as it is implied by the constraints $\mathbf{Y} \succeq 0$ and $Y_{ii} = 1, \forall i$. We still keep this constraint, as the property $\widehat{\mathbf{Y}} \le \mathbf{J}$ plays a crucial role in our analysis.

following holds. If $s \geq C_s k^2/n$, then we have

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1} \leq C_g \exp\left[-\frac{sn}{C_e k}\right]$$

with probability at least $1 - \frac{7}{n} - 7e^{-\Omega(\sqrt{n})}$.

Our next result concerns the explicit clustering $\widehat{\boldsymbol{\sigma}} := \rho\text{-}\mathtt{kmed}(\widehat{\mathbf{Y}})$ extracted from $\widehat{\mathbf{Y}}$ using the approximate k-medians procedure given in Section II-C, where $\rho = 6\frac{2}{3}$. As we show in the proof of the following theorem, the error rate in $\widehat{\boldsymbol{\sigma}}$ is (deterministically) upper-bounded by the error in $\widehat{\mathbf{Y}}$:

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \leq \frac{86}{3} \cdot \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1};$$

cf. Proposition 3. Consequently, the number of misclassified nodes also exhibits an exponential decay.

Theorem 2 (Clustering Error): Under Model 1, there exist universal constants $C_s, C_m, C_e > 0$ for which the following holds. If $s \geq C_s k^2/n$, then we have

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \leq C_m \exp\left[-\frac{sn}{C_e k}\right]$$

with probability at least $1 - \frac{7}{n} - 7e^{-\Omega(\sqrt{n})}$.

We prove this theorem in Appendix III.

Theorems 1 and 2 are applicable in the sparse graph regime with bounded expected degrees. For example, suppose that $k = 2$, $p = \frac{a}{n}$ and $q = \frac{b}{n}$ for two constants a, b; the results above guarantee non-trivial accuracy for the SDP (i.e., $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 < \frac{1}{2}\|\mathbf{Y}^*\|_1$ or $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) < \frac{1}{2}$) as long as $(a-b)^2 \geq C(a+b)$ for some constant C. Another interesting regime to which our results apply, is when there is a large number of clusters. For example, for any constant $\epsilon \in (0, \frac{1}{2})$, if $k = n^{1/2-\epsilon}$ and $p = 2q$, then SDP achieves exact recovery (i.e.,$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) < \frac{1}{n}$) provided that $p \gtrsim n^{-2\epsilon}$.

Below we provide additional discussion of our results, and compare with existing work.

1) Consequences and Optimality: Theorems 1 and 2 immediately imply sufficient conditions for the various recovery types discussed in Section I-B.

- *Exact Recovery (Strong Consistency):* When $s \gtrsim \frac{k^2 + k\log n}{n}$, Theorem 1 guarantees that $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 < \frac{1}{2}$ with high probability, in which case element-wise rounding $\widehat{\mathbf{Y}}$ exactly recovers the true cluster matrix $\mathbf{Y}^*$.[3] This result matches the best known exact recovery guarantees for SDP (and other polynomial-time algorithms) when k is allowed to grow with n; see [7], [10] for a review of these results.
- *Almost Exact Recovery (Weak Consistency):* Under the condition $s = \omega(\frac{k^2}{n})$, Theorem 2 ensures that $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = o(1)$ with high probability as $n \to \infty$, hence SDP achieves weak consistency. This condition is optimal — necessary for *any* algorithms — as has been proved in [22] and [23].
- *Weak Recovery:* When $s \gtrsim \frac{k^2}{n}$, Theorem 2 ensures that $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) < 1 - \frac{1}{k}$ with high probability, hence SDP achieves weak recovery. In particular, in the setting with $k = 2$ clusters, SDP recovers a binary clustering that is positively correlated with the ground-truth clustering under the condition $s \gtrsim \frac{1}{n}$. This condition matches up to constants the so-called Kesten-Stigum (KS) threshold $s > \frac{1}{n}$, which is known to be optimal [8], [18]–[21].
- *Recovery With δ Error:* More generally, for each number $\delta \in (0, 1)$, Theorem 2 implies that if $s \gtrsim \max\{\frac{k^2}{n}, \frac{k}{n}\log\frac{1}{\delta}\}$, then $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) < \delta$ with high probability. In the case with $k = 2$, the minimax rate result in [24] and [48] implies that $s \gtrsim \frac{1}{n}\log\frac{1}{\delta}$ is necessary for *any* algorithm to achieve a δ clustering error. Our results are thus optimal up to a multiplicative constant.

Our results therefore cover these different recovery regimes via a unified error bound, using a single algorithm. This can be contrasted with the existing error bound (2) proved using the Grothendieck's inequality approach, which fails to identify the exact recovery condition above. In particular, the bound (2) decays polynomially with the SNR measure s; since s is at most k and $\|\mathbf{Y}^*\|_1 = n^2/k$, the smallest possible error that can be derived from this bound is $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 = O(\sqrt{n^3/k})$.

Our results apply to general values of k that is allowed to scale with n, hence the size of the clusters can be sublinear in n. We note that in this regime, a computational-barrier phenomenon seems to take place: there may exist parameter regimes of SBM for which cluster recovery is information-theoretically possible but cannot be achieved by computationally efficient algorithms. For example, the work in [7] proves that the intractable maximum likelihood estimator succeeds in exact recovery when $s \gtrsim \frac{k\log n}{n}$; it also provides evidences suggesting that all efficient algorithms fail unless $s \gtrsim \frac{k^2 + k\log n}{n}$. Note that the latter is consistent with the exact recovery condition derived above from our theorems.

The above discussion has the following implications for the optimality of Theorems 1 and 2. On the one hand, the general minimax rate result in [24] and [48] suggests that all algorithms (regardless of their computational complexity) incur at least $\exp[-\Theta(sn/k)]$ error. Our exponential error rate matches this information-theoretic lower bound. On the other hand, in view of the computational barrier discussed in the last paragraph, our SNR assumption $s \gtrsim k^2/n$ is likely to be unimprovable when efficient algorithms are considered.

2) Comparison With Existing Results: We discuss some prior work that also provides efficient algorithms attaining an exponentially-decaying rate for the clustering error $\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$. To be clear, these algorithms are very different from ours, often involving a two-step procedure that first computes an accurate initial estimate (typically by spectral clustering) followed by a "clean-up" process to obtain the final solution. Some of them require additional steps of sample splitting and graph trimming/regularization. As discussed in Section III-B to follow, many of these procedures rely on delicate properties of the standard SBM, and therefore are not robust against model deviation.

Most relevant to us is the work in [31], which develops a spectral algorithm with sample splitting. As stated in their main theorem, their algorithm achieves the error rate

[3] In fact, a simple modification of our analysis proves that $\widehat{\mathbf{Y}} = \mathbf{Y}^*$ in this case. We omit the details of such refinement for a more streamlined presentation of the analysis.

$\exp\left[-\Omega(sn/k^2)\right]$ when $s \gtrsim k^2/n$, as long as k is a fixed constant when $n \to \infty$. The work in [25] and [32] also considers spectral algorithms, which attain exponential error rates assuming that k is a constant and $pn \to \infty$. The algorithms in [24] and [49] involves obtaining an initial clustering using spectral algorithms, which require $s \gg k^3/n$; a post-processing step (e.g., using a Lloyd's-style algorithm [46]) then outputs a final solution that asymptotically achieves the minimax error rate $\exp\left[-I^* \cdot n/k\right]$, where I^* is an appropriate form of Renyi divergence and satisfies $I^* \asymp s$. The work in [23] proposes an efficient algorithm called Sphere Comparison, which achieves an exponential error rate in the constant degree regime $p = \Theta(1/n)$ when $s \geq k^2/n$. The work [33] uses SDP to produce an initial clustering solution to be fed to another clustering algorithm; their analysis generalizes the techniques in [4] to the setting with corrupted observations, and their overall algorithm attains an exponential error rate assuming that $s \gtrsim k^4/n$.

B. Robustness

Compared to other clustering algorithms, one notable advantage of the SDP approach is its robustness under various challenging settings of SBM. For instance, standard spectral clustering is known to be inconsistent in the sparse graph regime with $p, q = O(1/n)$ due to the existence of atypical node degrees, and alleviating this difficulty generally requires sophisticated algorithmic techniques. In contrast, as shown in Theorem 1 as well as other recent work [4], [15], [17], the SDP approach is applicable without change to this sparse regime. SDP is also robust against the existence of $o(n)$ outlier nodes and/or edge modifications, while standard spectral clustering is fairly fragile in these settings [17], [28], [33], [39], [42], [44].

Here we focus on another remarkable form of robustness enjoyed by SDP with respect to heterogeneous edge probabilities and monotone attack, which is captured in the following generalization of the standard SBM.

Model 2 (Heterogeneous Stochastic Block Model): Given the ground-truth clustering $\boldsymbol{\sigma}^*$ (encoded in the cluster matrix $\mathbf{Y}^*$), the entries $\{A_{ij}, i < j\}$ of the graph adjacency matrix $\mathbf{A}$ are generated independently as follows:

$$\begin{cases} A_{ij} \text{ is Bernoulli with rate } \textit{at least } p & \text{if } Y^*_{ij} = 1, \\ A_{ij} \text{ is Bernoulli with rate } \textit{at most } q & \text{if } Y^*_{ij} = 0, \end{cases}$$

where $p > q$.

The above model imposes no constraint on the edge probabilities besides the upper/lower bounds; in particular the probabilities can be non-uniform inside and outside the clusters. This model encompasses a variant of the so-called *monotone attack* studied extensively in the computer science literature [14], [41], [43]: here an adversary can arbitrarily set some edge probabilities to 1 or 0, which is equivalent to adding edges to node pairs in the same cluster and removing edges across clusters.[4] Note that the adversary can make far more than $o(n)$ edge modifications — $O(n^2)$ to be precise — in a restrictive way that seems to strengthen the clustering structure (hence the name). However, monotone attack does not necessarily make the clustering problem easier. On the contrary, the adversary can significantly alter some predictable structures that arise in standard SBM (such as the graph spectrum, node degrees, subgraph counts and the non-existence of dense spots [41]), and hence foil algorithms that over-exploit such structures. Indeed, some spectral algorithms provably fail in this setting [42], [43]. More generally, Model 2 allows for unpredictable, non-random deviations (not necessarily due to a malicious adversary) from the standard SBM setting, which has statistical properties that are rarely possessed by real-world graphs.

It is straightforward to show that when *exact recovery* is concerned, SDP is unaffected by the heterogeneity in Model 2; see [16], [27], [41]. The following theorem, proved in Section V, shows that SDP in fact achieves the same exponential *error rates* in the presence of heterogeneity.

Theorem 3 (Robustness): The conclusions in Theorems 1 and 2 continue to hold under Model 2.

Consequently, under the same conditions given in Section III-A.1, the SDP approach achieves exact recovery, almost exact recovery, weak recovery and a δ-error in the more general Model 2**.**

As a passing note, the results in [44] show that when exact constant values are concerned, the optimal threshold for weak recovery changes in the presence of monotone attack, and there may exist a fundamental tradeoff between achieving optimal recovery in standard SBM and robustness against model deviation.

C. Censored Block Model

The Censored Block Model [34] is a variant of the standard SBM that represents the scenario with partially observed data, akin to the settings of matrix completion [50] and graph clustering with measurement budgets [7]. In this section, we show that Theorems 1 and 2 immediately yield recovery guarantees for an SDP formulation of this model.

Concretely, again assume a ground-truth set of k equal-size clusters over n nodes, with the corresponding label vector $\boldsymbol{\sigma}^* \in [k]^n$. These clusters can be encoded by the cluster matrix $\mathbf{Y}^* \in \{0,1\}^{n\times n}$ as defined in Section I, but it is more convenient to work with its ± 1 version $2\mathbf{Y}^* - \mathbf{J}$. Under the Censored Block Model, one observes the entries of the matrix $2\mathbf{Y}^* - \mathbf{J}$ restricted to the edges of an Erdos-Renyi graph $G(n, \alpha)$, but with each entry flipped with probability $\epsilon < \frac{1}{2}$. The model is described formally below.

Model 3 (Censored Block Model): The observed matrix $\mathbf{Z} \in \{-1, 0, 1\}^{n\times n}$ is symmetric and has entries generated independently across all $i < j$ with

$$Z_{ij} = \begin{cases} 0, & \text{with probability } 1-\alpha, \\ 2Y^*_{ij} - 1 & \text{with probability } \alpha(1-\epsilon), \\ -(2Y^*_{ij} - 1) & \text{with probability } \alpha\epsilon, \end{cases}$$

where $0 < \alpha \leq 1$ and $0 < \epsilon < \frac{1}{2}$.

[4] We do note that here the addition/removal of edges are determined before the realization of the random edge connections, which is more restrictive than the standard monotone attack model. We believe that this restriction is an artifact of the analysis, and leave further improvements to future work.

The goal is again to estimate $\mathbf{Y}^*$ (equivalently $2\mathbf{Y}^* - \mathbf{J}$) given the observed matrix $\mathbf{Z}$.

We can reduce this problem to the standard SBM by constructing an adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ with $A_{ij} = |Z_{ij}| \cdot (Z_{ij}+1)/2$; that is, we take the binary representation $(\mathbf{Z}+\mathbf{J})/2$ of $\mathbf{Z}$ and zero out its unobserved entries. The upper-triangular entries of $\mathbf{A}$ are independent Bernoulli variables with

$$\mathbb{P}\left(A_{ij}=1\right) = \mathbb{P}\left(Z_{ij}=1\right) = \begin{cases} \alpha(1-\epsilon) & \text{if } Y^*_{ij}=1, \\ \alpha\epsilon & \text{if } Y^*_{ij}=0. \end{cases}$$

Therefore, the matrix $\mathbf{A}$ can be viewed as generated from the standard SBM (Model 1) with $p=\alpha(1-\epsilon)$ and $q=\alpha\epsilon$. We can then obtain an estimate $\widehat{\mathbf{Y}}$ of $\mathbf{Y}^*$ by solving the SDP formulation (1) or (6) with $\mathbf{A}$ as the input, possibly followed by the approximate k-medians procedure to obtain an explicit clustering $\widehat{\boldsymbol{\sigma}}$. Bounds on the error rates of $\widehat{\mathbf{Y}}$ and $\widehat{\boldsymbol{\sigma}}$ can be derived as a corollary of Theorems 1 and 2.

Corollary 1 (Censored Block Model): Under Model 3, there exist universal constants $C_s, C_g, C_e > 0$ for which the following holds. If $\alpha(1-2\epsilon)^2 \geq C_s\frac{k^2}{n}$, then

$$\begin{aligned}\frac{3}{86}\,\mathtt{err}(\widehat{\boldsymbol{\sigma}},\boldsymbol{\sigma}^*) &\leq \frac{\|\widehat{\mathbf{Y}}-\mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1} \\ &\leq C_g \exp\left[-\alpha(1-2\epsilon)^2\cdot\frac{n}{C_e k}\right]\end{aligned}$$

with probability at least $1-\frac{7}{n}-7e^{-\Omega(\sqrt{n})}$.

Specializing this corollary to the different types of recovery defined in Section I-B, we immediately obtain the following sufficient conditions for SDP under the Censored Block Model:

- Exact recovery is achieved when $\alpha \gtrsim \frac{\max\{k\log n, k^2\}}{n(1-2\epsilon)^2}$.
- Almost exact recovery is achieved when $\alpha = \omega\left(\frac{k^2}{n(1-2\epsilon)^2}\right)$.
- Weak recovery is achieved when $\alpha \gtrsim \frac{k^2}{n(1-2\epsilon)^2}$.
- A δ clustering error is achieved when $\alpha \gtrsim \frac{\max\{k^2, k\log(1/\delta)\}}{n(1-2\epsilon)^2}$.

Several existing results focus on the Censored Block Model with $k=2$ clusters in the asymptotic regime $n\to\infty$. In this setting, the work in [34] proves that exact recovery is possible if and only if $\alpha > \frac{2\log n}{n(1-2\epsilon)^2}$ in the limit $\epsilon\to 1/2$, and provides an SDP-based algorithm that succeeds twice the above threshold; a more precise threshold $\alpha > \frac{\log n}{n\left(\sqrt{1-\epsilon}-\sqrt{\epsilon}\right)^2}$ is given in [35]. For weak recovery with a sparse graph $\alpha=\Theta(1/n)$, it is conjectured in [36] that the problem is solvable if and only if $\alpha > \frac{1}{n(1-2\epsilon)^2}$. The converse and achievability parts of the conjecture are proved in [37] and [38], respectively. Corollary 1 shows that SDP achieves (up to constants) the above exact and weak recovery thresholds; moreover, our results apply to the more general setting with $k\geq 2$ clusters.

IV. Proof of Theorem 1

In this section we prove our main theoretical results in Theorem 1 for Model 1 (Standard SBM). While Model 1 is a special case of Model 2 (Heterogeneous SBM), we choose not to deduce Theorem 1 as a corollary of Theorem 3 which concerns the more general model. Instead, to highlight the main ideas of the analysis and avoid technicalities, we provide a self-contained proof of Theorem 1. In Section V to follow, we show how to adapt the proof to Model 2.

Before going into the details, we make a few observations that simplify the subsequent proof. Firstly, it suffices to prove the theorem for the first SDP formulation (1). To see this, note that the ground-truth matrix $\mathbf{Y}^*$ is also feasible to second formulation (6); moreover, thanks to the equality constraint $\sum_{i,j=1}^n Y_{ij} = \sum_{i,j=1}^n Y^*_{ij}$, subtracting the constant-valued term $\left\langle \mathbf{Y}, \frac{p+q}{2}\mathbf{J}\right\rangle$ from the objective of (6) does not affect its optimal solutions. The two formulations are therefore identical except for the above equality constraint, which is never used in the proof below. Secondly, under the assumption $cp\leq q\leq p$ with a universal constant $c>0$, we have $\frac{1}{k}p + (1-\frac{1}{k})q \geq cp$. Therefore, it suffices to prove the theorem with the SNR redefined as $s=\frac{(p-q)^2}{p}$, doing which only affects the values of universal constants C_s and C_e in the theorem statement. Thirdly, it is in fact sufficient to prove the bound

$$\|\widehat{\mathbf{Y}}-\mathbf{Y}^*\|_1 \leq C_g n^2 \exp\left[-\frac{sn}{2C_e k}\right]. \tag{8}$$

Suppose that this bound holds; under the premise $s\geq C_s k^2/n$ of the theorem with the constant C_s sufficiently large, we have $\exp\left[-\frac{sn}{4C_e k}\right] \leq \exp\left[-\frac{C_s}{4C_e}\cdot k\right] \leq \frac{1}{k}$, hence the RHS of the bound (8) is at most

$$C_g n^2 \cdot \frac{1}{k}\cdot \exp\left[-\frac{sn}{4C_e k}\right] = C_g\|\mathbf{Y}^*\|_1 \exp\left[-\frac{sn}{4C_e k}\right],$$

which implies the error bound in theorem statement again up to a change in the universal constant C_e. To summarize, the desired Theorem 1 is implied by the following statement.

Theorem 4: Under Model 1 with $0\leq q<p\leq 1$, there exist universal constants $C_s, C_g, C_e > 0$ for which the following holds. If $s := (p-q)^2/p \geq C_s k^2/n$, then with probability at least $1-\frac{7}{n}-7e^{-\Omega(\sqrt{n})}$, any optimal solution $\widehat{\mathbf{Y}}$ of SDP (1) satisfies the bound (8).

We prove this theorem in the rest of the section. Throughout the proof we make use of the convenient shorthands $\gamma := \|\widehat{\mathbf{Y}}-\mathbf{Y}^*\|_1$ for the error and $\ell := \frac{n}{k}$ for the cluster size.

Our proof begins with a basic inequality using optimality. Since $\mathbf{Y}^*$ is feasible to the SDP (1) and $\widehat{\mathbf{Y}}$ is optimal, we have

$$\begin{aligned} 0 &\leq \left\langle \widehat{\mathbf{Y}}-\mathbf{Y}^*, \mathbf{A}-\frac{p+q}{2}\mathbf{J}\right\rangle \\ &= \left\langle \widehat{\mathbf{Y}}-\mathbf{Y}^*, \mathbb{E}\mathbf{A}-\frac{p+q}{2}\mathbf{J}\right\rangle + \left\langle \widehat{\mathbf{Y}}-\mathbf{Y}^*, \mathbf{A}-\mathbb{E}\mathbf{A}\right\rangle. \end{aligned} \tag{9}$$

A simple observation is that the entries of the error matrix $\mathbf{Y}^*-\widehat{\mathbf{Y}}$ have matching signs with those of $\mathbb{E}\mathbf{A}-\frac{p+q}{2}\mathbf{J}$. This observation implies the following relationship between the first term on the RHS of equation (9) and the error $\gamma := \|\widehat{\mathbf{Y}}-\mathbf{Y}^*\|_1$.

Fact 1: We have the inequality

$$\left\langle \mathbf{Y}^*-\widehat{\mathbf{Y}}, \mathbb{E}\mathbf{A}-\frac{p+q}{2}\mathbf{J}\right\rangle \geq \frac{p-q}{2}\gamma. \tag{10}$$

The proof of this fact is deferred to Appendix II-A. Taking this fact as given and combining with inequality (9), we obtain that

$$\frac{p-q}{2}\gamma \leq \left\langle \widehat{\mathbf{Y}}-\mathbf{Y}^*, \mathbf{A}-\mathbb{E}\mathbf{A}\right\rangle. \tag{11}$$

To bound the error γ, it suffices to control the RHS of equation (11). This is where we depart from existing analysis. The seminal work in [4] bounds the RHS by a direct application of the Grothendieck's inequality. As we discuss below, this argument fails to expose the fast, exponential decay of the error γ. Our analysis establishes a more precise bound. To describe our approach, some additional notation is needed. Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ be the matrix whose columns are the left singular vectors of $\mathbf{Y}^*$ corresponding to non-zero singular values. Define the projection $\mathcal{P}_T(\mathbf{M}) := \mathbf{U}\mathbf{U}^\top\mathbf{M} + \mathbf{M}\mathbf{U}\mathbf{U}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{M}\mathbf{U}\mathbf{U}^\top$ and its orthogonal complement $\mathcal{P}_{T^\perp}(\mathbf{M}) = \mathbf{M} - \mathcal{P}_T(\mathbf{M})$ for any $\mathbf{M} \in \mathbb{R}^{n\times n}$. Our crucial observation is that one should control $\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A}\rangle$ by separating the contributions from the two projected components of $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ defined by $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$. In particular, we rewrite the inequality (11) as

$$\frac{p-q}{2}\gamma \le \underbrace{\left\langle \mathcal{P}_T(\widehat{\mathbf{Y}} - \mathbf{Y}^*), \mathbf{A} - \mathbb{E}\mathbf{A}\right\rangle}_{S_1} + \underbrace{\left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}} - \mathbf{Y}^*), \mathbf{A} - \mathbb{E}\mathbf{A}\right\rangle}_{S_2}. \tag{12}$$

The first term S_1 involves the component of error $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ that is "aligned" with $\mathbf{Y}^*$; in particular, the matrix $\mathcal{P}_T(\widehat{\mathbf{Y}} - \mathbf{Y}^*)$ lies in the subspace spanned by matrices with the same column or row space as $\mathbf{Y}^*$. The second term S_2 involves the orthogonal component $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}} - \mathbf{Y}^*)$, whose column and row spaces are orthogonal to those of $\mathbf{Y}^*$. The main step of our analysis consists of controlling S_1 and S_2 separately.

The following proposition bounds the term S_1 and is proved in Section IV-B to follow.

Proposition 1: Under the conditions of Theorem 4, with probability at least $1 - \frac{6}{nk} - 2(\frac{e}{2})^{-2n}$, at least one of the following inequalities hold:

$$\gamma \le C_g n^2 e^{-sn/(2C_e k)},$$
$$S_1 \le D_1 \gamma \sqrt{\frac{p\log\left(3n^2/\gamma\right)}{\ell}}, \tag{13}$$

where $D_1 = 12D = 12\sqrt{10}$.

The next proposition, proved in Section IV-C to follow, controls the term S_2.

Proposition 2: Under the conditions of Theorem 4, with probability $1 - \frac{1}{n^2} - e^{-(3-\ln 2)n} - 4e^{-c'\sqrt{n}}$, at least one of the following inequalities hold:

$$\gamma \le C_g n^2 e^{-sn/(2C_e k)},$$
$$S_2 \le D_2\sqrt{pn}\frac{\gamma}{\ell} + \frac{1}{8}(p-q)\gamma. \tag{14}$$

where $D_2 > 0$ is a universal constant.

Given these two propositions, the desired bound (8) follows easily. If the first inequality in the two propositions holds, then we are already done. Otherwise, there must hold the inequalities (13) and (14), which can be plugged into the RHS of equation (12) to get

$$\frac{p-q}{2}\gamma \le D_1\gamma\sqrt{\frac{p\log\left(3n^2/\gamma\right)}{\ell}} + D_2\sqrt{\frac{pk^2}{n}}\gamma + \frac{1}{8}(p-q)\gamma.$$

Under the premise $s \ge C_s k^2/n$ of Theorem 4, we know that $D_2\sqrt{\frac{pk^2}{n}} \le \frac{p-q}{8}$. It follows that

$$\frac{p-q}{4}\gamma \le D_1\gamma\sqrt{\frac{p\log\left(3n^2/\gamma\right)}{\ell}}.$$

Doing some algebra yields the inequality $\gamma \le 3n^2\exp[-sn/(16D_1^2k)]$, so the desired bound (8) again holds.

The rest of this section is devoted to establishing Propositions 1 and 2. Before proceeding to the proofs, we remark on the above arguments and contrast them with alternative approaches.

Comparison With the Grothendieck's Inequality Approach: The proof in [4] also begins with a version of the inequality (11), and proceeds by observing that

$$\frac{p-q}{2}\gamma \le \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A}\right\rangle \overset{(i)}{\le} 2\sup_{\mathbf{Y}\ge 0, \mathrm{diag}(\mathbf{Y})\le \mathbf{1}} \left|\langle \mathbf{Y}, \mathbf{A} - \mathbb{E}\mathbf{A}\rangle\right|, \tag{15}$$

where step (i) follows from the triangle inequality and the feasibility of $\widehat{\mathbf{Y}}$ and $\mathbf{Y}^*$. This argument reduces the problem to bounding the RHS of (15), which can be done using the celebrated Grothendieck's inequality. One can already see at this point that this approach yields sub-optimal bounds. For example, SDP is known to achieve exact recovery ($\gamma = 0$) under certain conditions, yet the inequality (15) can never guarantee a zero γ. Sub-optimality arises in step (i): the quantity $\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} - \mathbb{E}\mathbf{A}\right\rangle$ diminishes when $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ is small, but the triangle inequality and the worse-case bound used in (i) are too crude to capture such behaviors. In comparison, our proof takes advantage of the structures of the error matrix $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ as well as its interplay with the noise matrix $\mathbf{A} - \mathbb{E}\mathbf{A}$.

Bounding the S_1 Term: A common approach involves using the generalized Holder's inequality

$$S_1 = \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathcal{P}_T(\mathbf{A} - \mathbb{E}\mathbf{A})\right\rangle \le \gamma \cdot \|\mathcal{P}_T(\mathbf{A} - \mathbb{E}\mathbf{A})\|_\infty.$$

Under SBM, one can show that $\|\mathcal{P}_T(\mathbf{A} - \mathbb{E}\mathbf{A})\|_\infty \lesssim \sqrt{\frac{p\log n^2}{\ell}}$ with high probability, hence yielding the bound $S_1 \lesssim \gamma\sqrt{\frac{p\log n^2}{\ell}}$. Variants of this approach are in fact common (sometimes implicitly) in the proofs of exact recovery for SDP [7], [10], [12], [27], [51]. However, when $\sqrt{\frac{p\log n^2}{\ell}} \ge p - q$ (where exact recovery is impossible), applying this bound of S_1 to the inequality (12) would yield a vacuous bound for γ. In comparison, Proposition 1 gives a strictly sharper bound (13), which correctly characterizes the behaviors of S_1 beyond the exact recovery regime.

Bounding the S_2 Term: Note that since $\mathcal{P}_{T^\perp}(\mathbf{Y}^*) = 0$, we have the equality $S_2 = \langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{A} - \mathbb{E}\mathbf{A}\rangle$. It is easy to show that the matrix $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})$ is positive semidefinite and has diagonal entries at most 4 (cf. Fact 3). Given these two properties, one may attempt to control S_2 again using the Grothendieck's inequality, which would yield the bound $S_2 \le 4 \cdot g(\mathbf{A} - \mathbb{E}\mathbf{A})$ for some function g independent of γ. The bound (14) in Proposition 2 is substantially stronger — it depends on γ, which is in turn proportional to the trace of the matrix $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})$ (cf. Fact 2).

A. Preliminaries and Additional Notation

Recall that $\mathbf{U} \in \mathbb{R}^{n\times k}$ is the matrix of the left singular vectors of $\mathbf{Y}^*$. This matrix is explicitly given by $U_{ia} = 1/\sqrt{\ell}$ if node i is in cluster a and $U_{ia} = 0$ otherwise. Consequently, $\mathbf{U}\mathbf{U}^\top$ is a block diagonal matrix where the entries inside each diagonal block are all equal to $1/\ell$.

Define the noise matrix $\mathbf{W} := \mathbf{A} - \mathbb{E}\mathbf{A}$. The matrix $\mathbf{W}$ is symmetric, which introduces some minor dependency among its entries. To handle this, we let $\mathbf{\Psi}$ be the matrix obtained from $\mathbf{W}$ with its entries in the lower triangular part set to zero. Note that $\mathbf{W} = \mathbf{\Psi} + \mathbf{\Psi}^\top$, and $\mathbf{\Psi}$ has independent entries (with zero entries considered Bern(0)). Similarly, we define $\mathbf{\Lambda}$ as the upper triangular part of the adjacency matrix $\mathbf{A}$.

In the proof we frequently use the inequalities $s = \frac{(p-q)^2}{p} < p-q \le p$. Consequently, the assumption $s \ge C_s k^2/n$ implies that $p \ge C_s k^2/n$. We also record an elementary inequality that is used multiple times.

Lemma 1: For each number $\alpha > 0$, there exists a number $C(\alpha) \ge 1$ such that if $s \ge \frac{C(\alpha)k}{n}$, then

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

Proof: Note that $pn \ge (p-q)n \ge sn \ge C(\alpha)k$. As long as $C(\alpha)$ is sufficiently large, we have $\frac{pn}{k} \le e^{pn/(2\alpha k)}$. These inequalities imply that

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

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

B. Proof of Proposition 1

In this section we prove Proposition 1, which controls the quantity S_1. Using the symmetry of $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ and the cyclic invariance of the trace, we obtain the identity

$$\begin{aligned}
S_1 &= \left\langle \mathcal{P}_T(\mathbf{W}), \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&= \left\langle \mathbf{U}\mathbf{U}^\top\mathbf{W}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle + \left\langle \mathbf{W}\mathbf{U}\mathbf{U}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\quad - \left\langle \mathbf{U}\mathbf{U}^\top\mathbf{W}\mathbf{U}\mathbf{U}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&= 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{W}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\quad - \left\langle \mathbf{U}\mathbf{U}^\top\mathbf{W}\mathbf{U}\mathbf{U}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&= 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle + 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\quad - \left\langle \mathbf{U}\mathbf{U}^\top(\mathbf{\Psi} + \mathbf{\Psi}^\top)\mathbf{U}\mathbf{U}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&= 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle + 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\quad - 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}\mathbf{U}\mathbf{U}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&= 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle + 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle \\
&\quad - 2\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, (\widehat{\mathbf{Y}} - \mathbf{Y}^*)\mathbf{U}\mathbf{U}^\top \right\rangle.
\end{aligned}$$

It follows that

$$\begin{aligned}
S_1 \le\ & 2\left|\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle\right| \\
& + 2\left|\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \right\rangle\right| \\
& + 2\left|\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, (\widehat{\mathbf{Y}} - \mathbf{Y}^*)\mathbf{U}\mathbf{U}^\top \right\rangle\right|. \qquad (16)
\end{aligned}$$

Note that $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_\infty \le 1$ since $\widehat{\mathbf{Y}}, \mathbf{Y}^* \in \{0,1\}^{n\times n}$. One can also check that $\|(\widehat{\mathbf{Y}} - \mathbf{Y}^*)\mathbf{U}\mathbf{U}^\top\|_\infty \le 1$. This is due to the fact that each entry of the matrix inside the norm is the mean of ℓ entries in $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ given the block diagonal structure of $\mathbf{U}\mathbf{U}^\top$, and the absolute value of the mean does not exceed 1. With the same reasoning, we see that $\|(\widehat{\mathbf{Y}} - \mathbf{Y}^*)\mathbf{U}\mathbf{U}^\top\|_1 \le \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 = \gamma$.

Key to our proof is a bound on the sum of order statistics. Intuitively, given m i.i.d. random variables, the sum of the β largest of them (in absolute value) scales as $O\big(\beta\sqrt{\log(m/\beta)}\big)$. The following lemma, proved in Appendix II-B, makes the above intuition precise and moreover establishes a uniform bound in β.

Lemma 2: Let $m \ge 8$ and $g \ge 1$ be positive integers. For each $j \in [m]$, define $X_j := \sum_{i=1}^{g}(B_{ij} - \mathbb{E}B_{ij})$, where B_{ij} are independent Bernoulli variables with variance at most ρ. Then for a constant $D = \sqrt{10}$, we have

$$\sum_{j=1}^{\lceil\beta\rceil} \left|X_{(j)}\right| \le D\,\lceil\beta\rceil\sqrt{g\rho\log(3m/\beta)}, \quad \forall \beta \in (3me^{-g\rho/C_e}, m],$$

with probability at least $1 - P_1(m)$, where $P_1(m) \le \frac{3}{m}$.

We are ready to bound the RHS of equation (16) and hence prove Proposition 1. Consider the event

$$\begin{aligned}
\mathcal{E}_1 := \Bigg\{ & \left|\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, \mathbf{M}\rangle\right| \le 2Db\sqrt{\frac{p\log\left(3n^2/b\right)}{\ell}}, \\
& \forall \mathbf{M}, b : \|\mathbf{M}\|_\infty \le 1, \|\mathbf{M}\|_1 \le b, \\
& 3n^2e^{-sn/(2C_ek)} < b \le n^2 \Bigg\},
\end{aligned}$$

and let $\mathcal{E}_2$ be defined similarly with $\mathbf{\Psi}$ replaced by $\mathbf{\Psi}^\top$. Note that the last line of $\mathcal{E}_1$ is valid since the assumption $s \ge C_s k^2/n$ for a sufficiently large $C_s > 0$ implies that $3n^2e^{-sn/(2C_ek)} < n^2$. We will use Lemma 2 to show that $\mathcal{E}_i$ holds with probability at least $1 - P_1 := 1 - P_1(nk)$ for each $i = 1, 2$. Taking this claim as given, we now condition on the intersection of the events $\mathcal{E}_1, \mathcal{E}_2$. Note that the matrix $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ satisfies $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_\infty \le 1$, and therefore $\gamma := \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \le n^2$. If $\gamma \le 3n^2e^{-sn/(2C_ek)}$, then the first inequality in Proposition 1 holds and we are done. Otherwise, on the event $\mathcal{E}_1$, we are guaranteed that

$$\left|\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, \widehat{\mathbf{Y}} - \mathbf{Y}^*\rangle\right| \le 2D\gamma\sqrt{\frac{p\log\left(3n^2/\gamma\right)}{\ell}}.$$

Since the matrix $\mathbf{M} := \left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\mathbf{U}\mathbf{U}^\top$ satisfies $\|\mathbf{M}\|_\infty \le 1$ and $\|\mathbf{M}\|_1 \le \gamma$, we have the bound

$$\left|\left\langle \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}, (\widehat{\mathbf{Y}} - \mathbf{Y}^*)\mathbf{U}\mathbf{U}^\top \right\rangle\right| \le 2D\gamma\sqrt{\frac{p\log\left(3n^2/\gamma\right)}{\ell}}.$$

Similarly, on the event $\mathcal{E}_2$, we have the bound

$$\left|\langle \mathbf{U}\mathbf{U}^\top \mathbf{\Psi}^\top, \widehat{\mathbf{Y}} - \mathbf{Y}^* \rangle\right| \leq 2D\gamma \sqrt{\frac{p \log\left(3n^2/\gamma\right)}{\ell}}.$$

Applying these estimates to the RHS of equation (16), we arrive at the bound $S_1 \leq 12D\gamma\sqrt{\frac{p\log(3n^2/\gamma)}{\ell}}$, which is the second inequality in Proposition 1.

It remains to bound the probability of the event $\mathcal{E}_1$, for which we shall use Lemma 2; the same arguments apply to the event $\mathcal{E}_2$. Let us take a digression to inspect the structure of the random matrix $\mathbf{V} := \mathbf{U}\mathbf{U}^\top\mathbf{\Psi}$. We treat each zero entry in $\mathbf{\Psi}$ as a Bern(0) random variable with its mean subtracted and independent of all other entries. Therefore, all entries of $\mathbf{\Psi}$ are independent. Since $\mathbf{U}\mathbf{U}^\top$ is a block diagonal matrix, $\mathbf{V}$ can be partitioned into k submatrices of size $\ell \times n$ stacked vertically, where rows within the same submatrix are identical and rows from different submatrices are independent with each of their entries equal to $1/\ell$ times the sum of ℓ independent centered Bernoulli random variables. To verify our observations, for $a \in [k]$ we use $\mathcal{R}_a := \{(a-1)\ell + 1, \ldots, a\ell\}$ to denote the set of row indices of the a-th submatrix of $\mathbf{V}$. Consider any $i \in \mathcal{R}_a$, that is, any row index of the a-th submatrix of $\mathbf{V}$. Then for all $j \in [n]$, we have $V_{ij} = \sum_{t=1}^n \left(\mathbf{U}\mathbf{U}^\top\right)_{it} \Psi_{tj} = \ell^{-1} \sum_{u \in \mathcal{R}_a} \Psi_{uj}$. We see that we get the same random variable by varying i within $\mathcal{R}_a$ while fixing j, but independent random variables by fixing i while varying j.

Fix an index $i_a := (a-1)\ell + 1 \in \mathcal{R}_a$ for each $a \in [k]$. Consider any $n \times n$ matrix $\mathbf{M}$ and number b such that $\|\mathbf{M}\|_\infty \leq 1$, $\|\mathbf{M}\|_1 \leq b$ and $3n^2 e^{-sn/(2C_e k)} < b \leq n^2$. We can compute

$$\begin{aligned}
|\langle \mathbf{V}, \mathbf{M}\rangle| &\leq \sum_{i=1}^n \sum_{j=1}^n \left|V_{ij}\right| \left|M_{ij}\right| \\
&= \sum_{a=1}^k \sum_{i \in \mathcal{R}_a} \sum_{j=1}^n \left|V_{ij}\right| \left|M_{ij}\right| \\
&= \sum_{a \in [k], j \in [n]} \left|\ell V_{i_a j}\right| \left[\sum_{i \in \mathcal{R}_a} \frac{\left|M_{ij}\right|}{\ell}\right],
\end{aligned}$$

where the last step follows from the previously established fact that $V_{ij} = V_{i_a j}, \forall i \in \mathcal{R}_a$. Recall that the nk random variables $\{\ell V_{i_a j}, a \in [k], j \in [n]\}$ are independent; moreover, each $\ell V_{i_a j}$ is the sum of ℓ independent Bernoulli variables with variance at most p. Let $X_{(t)}$ denote the element in these nk random variables with the t-th largest absolute value. Define the quantity $w := \|\mathbf{M}\|_1/\ell = \sum_{a \in [k], j \in [n]} \sum_{i \in \mathcal{R}_a} \left|M_{ij}\right|/\ell$. Since $\sum_{i \in \mathcal{R}_a} \ell^{-1} \left|M_{ij}\right| \leq \|\mathbf{M}\|_\infty \leq 1$ and $w \leq b/\ell$, we have

$$|\langle \mathbf{V}, \mathbf{M}\rangle| \leq \begin{cases} \sum_{t=1}^{\lceil w \rceil} \left|X_{(t)}\right| \leq \sum_{t=1}^{\lceil b/\ell \rceil} \left|X_{(t)}\right|, & b/\ell \geq 1 \\ \left|X_{(1)}\right| (b/\ell), & b/\ell < 1. \end{cases} \quad (17)$$

Here we use the fact that for any sequence of numbers $a_1, a_2, \ldots$ in $[0,1]$, $\sum_t |X_t| a_t \leq \sum_{t=1}^{\sum_i a_i} \left|X_{(t)}\right|$. Now note that $b/\ell \in (3nke^{-sn/(2C_e k)}, nk]$, which implies that $b/\ell \in (3nke^{-p\ell/C_e}, nk]$ by Lemma 1. Applying Lemma 2 with $m = nk \geq 8$, $g = \ell$, $\rho = p$ and $\beta = b/\ell$, we are guaranteed that with probability at least $1 - P_1(nk)$,

$$\sum_{t=1}^{\lceil b/\ell \rceil} \left|X_{(t)}\right| \leq D\lceil b/\ell \rceil \sqrt{\ell p \log\left(\frac{3nk}{b/\ell}\right)}$$

and

$$\left|X_{(1)}\right| \leq D\sqrt{\ell p \log(3nk)}$$

simultaneously for all relevant b/ℓ. On this event, we can continue the inequality (17) to conclude that

$$|\langle \mathbf{V}, \mathbf{M}\rangle| \leq \begin{cases} D\lceil b/\ell \rceil \sqrt{\ell p \log\left(\frac{3nk}{b/\ell}\right)}, & b/\ell \geq 1 \\ D(b/\ell)\sqrt{\ell p \log(3nk)}, & b/\ell < 1 \end{cases}$$

simultaneously for all relevant matrices $\mathbf{M}$. It is easy to see that the last RHS is bounded by $2Db\sqrt{\frac{p\log(3nk/w)}{\ell}}$ in either case, hence the event $\mathcal{E}_1$ holds.

We remark that the box constraints in the SDP (1) are crucial to the above arguments, which are ultimately applied to the matrix $\mathbf{M} = \widehat{\mathbf{Y}} - \mathbf{Y}^*$. In particular, the box constraints ensure that $\|\mathbf{M}\|_\infty = \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_\infty \leq 1$, which allows us to establish the inequality (17) and apply the order statistics bound in Lemma 2.

C. Proof of Proposition 2

We begin our proof by re-writing S_2 as

$$S_2 = \left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{W} \right\rangle, \quad (18)$$

which holds since $\mathcal{P}_{T^\perp}(\mathbf{Y}^*) = 0$ by definition of the projection $\mathcal{P}_{T^\perp}$. We can relate the matrix $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})$ appeared above to the quantity of interest, $\gamma = \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$. In particular, observe that

$$\begin{aligned}
&\mathrm{Tr}\left\{\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\} \\
&\quad = \mathrm{Tr}\left\{\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\right\} \\
&\quad \overset{(i)}{=} \mathrm{Tr}\left\{\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\right\} \\
&\quad \overset{(ii)}{=} \mathrm{Tr}\left\{\mathbf{U}\mathbf{U}^\top\left(\mathbf{Y}^* - \widehat{\mathbf{Y}}\right)\right\},
\end{aligned}$$

where step (i) holds since trace is invariant under cyclic permutations and $\mathbf{I} - \mathbf{U}\mathbf{U}^\top$ is a projection matrix, and step (ii) holds since $\mathrm{diag}\left(\widehat{\mathbf{Y}}\right) = \mathrm{diag}\left(\mathbf{Y}^*\right) = \mathbf{1}$ by feasibility of $\widehat{\mathbf{Y}}$ and $\mathbf{Y}^*$ to the SDP (1). Recall that all the non-zero entries of $\mathbf{U}\mathbf{U}^\top$ are in its diagonal block and equal to $1/\ell$. Since the corresponding diagonal-block entries of the matrix $\mathbf{Y}^* - \widehat{\mathbf{Y}}$ are non-negative, we have

$$\mathrm{Tr}\left\{\mathbf{U}\mathbf{U}^\top\left(\mathbf{Y}^* - \widehat{\mathbf{Y}}\right)\right\} = \sum_{(i,j): Y^*_{ij} = 1} \frac{1}{\ell} \cdot \left|\left(\mathbf{Y}^* - \widehat{\mathbf{Y}}\right)_{ij}\right| \leq \frac{\gamma}{\ell}.$$

Combining these pieces gives

Fact 2: $\mathrm{Tr}\left\{\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\} \leq \frac{\gamma}{\ell}$.

Equipped with this fact, we proceed to bound the quantity S_2 given by equation (18). We consider separately the dense case $p \geq c_d \frac{\log n}{n}$ and the sparse case $p \leq c_d \frac{\log n}{n}$, where $c_d > 0$ is a constant given in the statement of Lemma 5.

1) The Dense Case: First assume that $p \geq c_d \frac{\log n}{n}$. In this case bounding S_2 is relatively straightforward, as the graph spectrum is well-behaved. We first recall that the nuclear norm $\|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_*$ is defined as the sum of the singular values of the matrix $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})$. Since $\widehat{\mathbf{Y}} \succeq 0$ by feasibility, we have $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}) = (\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\widehat{\mathbf{Y}}(\mathbf{I}-\mathbf{U}\mathbf{U}^\top) \succeq 0$, whence $\|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_* = \mathrm{Tr}\left\{\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right\}$. Revisiting the expression (18), we obtain that

$$\begin{aligned} S_2 &= \left\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{W} \right\rangle \\ &\leq \mathrm{Tr}\left\{\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}}\right)\right\} \cdot \|\mathbf{W}\|_{\mathrm{op}} \\ &\leq \frac{\gamma}{\ell} \cdot \|\mathbf{W}\|_{\mathrm{op}}, \end{aligned}$$

where the first inequality follows from the duality between the nuclear and spectral norms, and the second inequality follows from Fact 2.

It remains to control the spectral norm $\|\mathbf{W}\|_{\mathrm{op}}$ of the centered adjacency matrix $\mathbf{W} := \mathbf{A} - \mathbb{E}\mathbf{A}$. This can be done in the following lemma, which is proved in Appendix II-C using standard tools from random matrix theory.

Lemma 3: We have $\|\mathbf{A}-\mathbb{E}\mathbf{A}\|_{\mathrm{op}} \leq 8\sqrt{pn}+174\sqrt{\log n}$ with probability at least $1 - P_2$ where $P_2 := n^{-2}$.

Applying the lemma, we obtain that with probability at least $1 - P_2$,

$$S_2 \leq \left(8\sqrt{pn} + 174\sqrt{\log n}\right) \cdot \frac{\gamma}{\ell} \leq D_2\sqrt{pn} \cdot \frac{\gamma}{\ell},$$

where in the last step holds for some constant D_2 sufficiently large under the assumption $p \geq c_d \frac{\log n}{n}$. This completes the proof of Proposition 2 in the dense case.

2) The Sparse Case: Now suppose that $p \leq c_d \frac{\log n}{n}$. In this case we can no longer use the arguments above, because some nodes will have degrees far exceeding their expectation $O(pn)$, and $\|\mathbf{W}\|_{\mathrm{op}}$ will be dominated by these nodes and become much larger than $\sqrt{pn}$. This issue is particularly severe when $p, q \asymp \frac{1}{n}$, in which case the expected degree is a constant, yet the maximum node degree diverges. Addressing this issue requires a new argument. In particular, we show that the matrix $\mathbf{W}$ can be partitioned into two parts, where the first part has a spectral norm bounded as desired, and the second part involves only a small number of edges; the structure of the SDP solution allows us to control the impact of the second part.

As before, to avoid the minor dependency due to symmetry, we focus on the upper triangular parts $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ of the matrices $\mathbf{A}$ and $\mathbf{W} := \mathbf{A} - \mathbb{E}\mathbf{A}$, and later make use of the relations $\mathbf{A} = \mathbf{\Lambda} + \mathbf{\Lambda}^\top$ and $\mathbf{W} = \mathbf{\Psi} + \mathbf{\Psi}^\top$. We define the sets

$$\mathcal{V}_{\mathrm{row}} := \left\{ i \in [n] : \sum_{j\in[n]} (\mathbf{\Lambda})_{ij} > 40pn \right\},$$

$$\mathcal{V}_{\mathrm{col}} := \left\{ j \in [n] : \sum_{i\in[n]} (\mathbf{\Lambda})_{ij} > 40pn \right\},$$

which are the nodes whose degrees (more specifically, row/column sums w.r.t. the halved matrix $\mathbf{\Lambda}$) are large compared to their expectation $O(pn)$. For each matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, we define the matrix $\widetilde{\mathbf{M}}$ such that

$$\widetilde{M}_{ij} = \begin{cases} 0, & \text{if } i \in \mathcal{V}_{\mathrm{row}} \text{ or } j \in \mathcal{V}_{\mathrm{col}}, \\ M_{ij}, & \text{otherwise.} \end{cases}$$

In other words, $\widetilde{\mathbf{M}}$ is obtained from $\mathbf{M}$ by "trimming" the rows/columns corresponding to nodes with large degrees. With this notation, we can write $\mathbf{\Psi} = \widetilde{\mathbf{\Psi}} + (\mathbf{\Psi} - \widetilde{\mathbf{\Psi}})$, which can be combined with the expression (18) to yield

$$\begin{aligned} S_2 &= \langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{\Psi} + \mathbf{\Psi}^\top \rangle \\ &= 2\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{\Psi} \rangle \\ &= 2\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \widetilde{\mathbf{\Psi}} \rangle + 2\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{\Psi} - \widetilde{\mathbf{\Psi}} \rangle \\ &\leq 2\,\mathrm{Tr}\left\{\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right\} \cdot \|\widetilde{\mathbf{\Psi}}\|_{\mathrm{op}} + 2\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{\Psi} - \widetilde{\mathbf{\Psi}} \rangle, \end{aligned} \quad (19)$$

where in the last step we use the fact that $\widehat{\mathbf{Y}} \succeq 0$ and thus $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}) = (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\widehat{\mathbf{Y}}(\mathbf{I} - \mathbf{U}\mathbf{U}^\top) \succeq 0$.

The first term in (19) can be controlled using the fact that the spectrum of the trimmed matrix $\widetilde{\mathbf{\Psi}}$ is well-behaved, for any p. In particular, we prove the following lemma in Appendix II-D as a consequence of known results.

Lemma 4: For some absolute constant $C > 0$, we have $\|\widetilde{\mathbf{\Psi}}\|_{\mathrm{op}} \leq C\sqrt{pn}$, with probability at least $1 - P_3$, where $P_3 := e^{-(3-\ln 2)2n} + (2n)^{-3}$.

One shall compare the bound in Lemma 4 with that in Lemma 3. After trimming, the term $\sqrt{\log n}$ in Lemma 3 (which would dominate in the sparse regime) disappears, and the spectral norm of $\widetilde{\mathbf{\Psi}}$ behaves similarly as a random matrix with Gaussian entries. Such a bound is in fact standard in recent work on spectral algorithms applied to sparse graphs with constant expected degrees [31], [50]. Typically these algorithms proceed by first trimming the graph, and then running standard spectral algorithms with the trimmed graph as the input. We emphasize that in our case trimming is used *only in the analysis*; our algorithm itself does not require any trimming, and the SDP (1) is applied to the original graph. As shown below, we are able to control the contributions from what is not trimmed, namely the second term in (19). Such a bound is made possible by leveraging the structures of the solution $\widehat{\mathbf{Y}}$ induced by the box constraints of the SDP (1) — a manifest of the regularization effect of SDP.

Turning to the second term in equation (19), we first note that the Holder's type inequality

$$\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \mathbf{\Psi} - \widetilde{\mathbf{\Psi}} \rangle \leq \mathrm{Tr}\left\{\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right\} \|\mathbf{\Psi} - \widetilde{\mathbf{\Psi}}\|_{\mathrm{op}}$$

is no longer sufficient, as the residual matrix $\mathbf{\Psi} - \widetilde{\mathbf{\Psi}}$ may have large eigenvalues. Here it is crucial to use an important property of the matrix $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})$, namely the fact that the magnitudes of its entries are $O(1)$, a consequence of the constraint $0 \leq \mathbf{Y} \leq \mathbf{J}$ in the SDP (1). More precisely, we have the following bound, which is proved in Appendix II-E.

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

Intuitively, this bound ensures that the mass of the matrix $\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})$ is spread across its entries, so its eigen-space is unlikely to be aligned with the top eigenvector of the random matrix $\mathbf{\Psi} - \widetilde{\mathbf{\Psi}}$, which by definition is supported on a few columns/rows indexed by $\mathcal{V}_{\mathrm{col}}$ and $\mathcal{V}_{\mathrm{row}}$. Consequently,

the quantity $\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}} \rangle$ is likely to be small. To make this precise, we start with the inequality

$$\langle \mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}}), \boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}} \rangle \le \|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_\infty \cdot \|\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}\|_1 \tag{20}$$

The following lemma bounds the ℓ_1 norm of the residual matrix $\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}$.

Lemma 5: Suppose $p \ge C_p/n$ for some sufficiently large constant $C_p > 0$, and $p \le c_d \log n/n$ for some sufficiently small constant $c_d > 0$. Then there exist some constants $C, c' > 0$ such that

$$\|\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}\|_1 \le Cpn^2 e^{-pn/C_e}$$

with probability at least $1 - P_4$ where $P_4 := 4e^{-c'\sqrt{n}}$.

We prove this lemma in Section IV-D to follow using tail bounds for sub-exponential random variables.

Equipped with the above bounds, we are ready to bound S_2. If $\gamma \le C_g n^2 e^{-sn/(2C_e k)}$, then the first inequality in Proposition 2 holds and we are done. It remains to consider the case with $\gamma > C_g n^2 e^{-sn/(2C_e k)}$, which by Lemma 1 implies that $\gamma > C_g n^2 \cdot \frac{p}{p-q} e^{-pn/C_e}$. The inequalities (19) and (20) together give

$$\begin{aligned} S_2 &\le 2 \operatorname{Tr}\left\{\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\right\} \cdot \|\widetilde{\boldsymbol{\Psi}}\|_{\mathrm{op}} \\ &\quad + \|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_\infty \cdot \|\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}\|_1 \\ &\le 2\frac{\gamma}{\ell}\|\widetilde{\boldsymbol{\Psi}}\|_{\mathrm{op}} + 4\|\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}\|_1, \end{aligned}$$

where we use Facts 2 and 3 in the second step. Applying Lemma 4 to the first RHS term above and Lemma 5 to the second, we obtain that

$$S_2 \le 2C\frac{\sqrt{pn}}{\ell}\gamma + 4Cpn^2 e^{-pn/C_e}$$

with probability at least $1 - P_3 - P_4$. Since $\gamma > C_g n^2 \cdot \frac{p}{p-q} e^{-pn/C_e}$ as shown above, we obtain that

$$S_2 \le 2C\frac{\sqrt{pn}}{\ell}\gamma + \frac{4C}{C_g}(p-q)\gamma \le 2C\frac{\sqrt{pn}}{\ell}\gamma + \frac{1}{8}(p-q)\gamma,$$

where the last step holds provided that the constant C_g is large enough. This proves the second inequality in Proposition 2.

D. Proof of Lemma 5

For any matrix $\mathbf{M}$, let $\mathrm{pos}_{i\bullet}(\mathbf{M})$ and $\mathrm{pos}_{\bullet j}(\mathbf{M})$ denote the number of positive entries in i-th row and j-th column of $\mathbf{M}$, respectively. Recall that the entries of $\boldsymbol{\Lambda}$ are independent Bernoulli random variables with variance at most p, where $C_p/n \le p \le c_d \log n/n$ for some sufficiently large constant $C_p > 0$ and sufficiently small constant $c_p > 0$. By definition, we have $\boldsymbol{\Psi} := \boldsymbol{\Lambda} - \mathbb{E}\boldsymbol{\Lambda}$, whence

$$\begin{aligned} \|\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}\|_1 &= \|(\boldsymbol{\Lambda} - \mathbb{E}\boldsymbol{\Lambda}) - \widetilde{(\boldsymbol{\Lambda} - \mathbb{E}\boldsymbol{\Lambda})}\|_1 \\ &= \|(\boldsymbol{\Lambda} - \mathbb{E}\boldsymbol{\Lambda}) - (\widetilde{\boldsymbol{\Lambda}} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}})\|_1 \\ &\le \|\mathbb{E}\boldsymbol{\Lambda} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}}\|_1 + \|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\|_1. \end{aligned}$$

The two terms on the last RHS are both random quantities depending on the sets $\mathcal{V}_{\mathrm{row}} = \{i : \mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) \ge 40pn\}$ and $\mathcal{V}_{\mathrm{col}} := \{j : \mathrm{pos}_{\bullet j}(\boldsymbol{\Lambda}) \ge 40pn\}$. We bound these two terms separately.

1) Bounding $\|\mathbb{E}\boldsymbol{\Lambda} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}}\|_1$*:* Let $\mathbb{I}(\cdot)$ denote the indicator function. We begin by investigating the indicator variable $G_i := \mathbb{I}(i \in \mathcal{V}_{\mathrm{row}}) = \mathbb{I}(\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) \ge 40pn)$; similar properties hold for $G'_j := \mathbb{I}(j \in \mathcal{V}_{\mathrm{col}})$. Note that $\{G_i, i \in [n]\}$ are independent Bernoulli random variables. To bound their rates $\mathbb{E}G_i$, we observe that each $\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) = \sum_j \Lambda_{ij}$ is the sum of n independent Bernoulli random variables with variance at most $p(1-p)n$, and $\mathbb{E}\,\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) \le pn$. The Bernstein's inequality ensures that for any number $z > pn$,

$$\begin{aligned} &\mathbb{P}\left\{\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) \ge z\right\} \\ &\quad \le \mathbb{P}\left\{\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) - \mathbb{E}\,\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) \ge z - pn\right\} \\ &\quad \le \exp\left\{-\frac{\frac{1}{2}(z-pn)^2}{p(1-p)n + \frac{1}{3}\cdot 1 \cdot (z-pn)}\right\} \\ &\quad \le \exp\left\{-\frac{(z-pn)^2}{2z}\right\}. \end{aligned} \tag{21}$$

Plugging in $z = 40pn$ yields

$$\begin{aligned} \mathbb{E}G_i &= \mathbb{P}\left\{\mathrm{pos}_{i\bullet}(\boldsymbol{\Lambda}) \ge 40pn\right\} \\ &\le \exp\left\{-\frac{(40pn - pn)^2}{2 \cdot 40pn}\right\} \\ &\le e^{-8pn}. \end{aligned}$$

We now turn to the quantity of interest, $\|\mathbb{E}\boldsymbol{\Lambda} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}}\|_1$. Since $\mathbb{E}\Lambda_{ij} \le p$ for all (i, j), we have that

$$\begin{aligned} \|\mathbb{E}\boldsymbol{\Lambda} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}}\|_1 &= \sum_i \sum_j \mathbb{E}\Lambda_{ij} \cdot \mathbb{I}(i \in \mathcal{V}_{\mathrm{row}} \text{ or } j \in \mathcal{V}_{\mathrm{col}}) \\ &\le \sum_i \sum_j p \cdot \mathbb{I}(i \in \mathcal{V}_{\mathrm{row}}) \\ &\quad + \sum_i \sum_j p \cdot \mathbb{I}(j \in \mathcal{V}_{\mathrm{col}}) \\ &= pn \sum_i G_i + pn \sum_j G'_j. \end{aligned}$$

The quantity $\sum_i G_i$ is the sum of independent Bernoulli variables with rates bounded above. Applying the Bernstein's inequality to this sum, we obtain that

$$\begin{aligned} &\mathbb{P}\left\{\sum_i G_i > 2ne^{-pn/C_e}\right\} \\ &\quad \le \mathbb{P}\left\{\sum_i G_i - \sum_i \mathbb{E}G_i > 2ne^{-pn/C_e} - ne^{-pn/C_e}\right\} \\ &\quad \le \exp\left\{-\frac{\frac{1}{2}\left(ne^{-pn/C_e}\right)^2}{n \cdot e^{-pn/C_e} + \frac{1}{3}\cdot 1 \cdot ne^{-pn/C_e}}\right\} \\ &\quad \le \exp\left\{-\frac{3}{8}ne^{-pn/C_e}\right\} \le \exp\left\{-\frac{3}{8}\sqrt{n}\right\}, \end{aligned}$$

where the first two steps hold because $C_e \ge 28$, and the last step holds by the assumption in Lemma 5 that $p \le c_d \log n/n$ for some sufficiently small $c_d > 0$. The same tail bound holds for the sum $\sum_j G'_j$. It follows that with probability at least $1 - 2e^{-\frac{3}{8}\sqrt{n}}$, we have

$$\|\mathbb{E}\boldsymbol{\Lambda} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}}\|_1 \le 4pn^2 e^{-pn/C_e}$$

2) Bounding $\|\mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}}\|_1$*:* Since the matrix $\mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}}$ have non-negative entries, we have the inequality

$$\begin{aligned}\|\mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}}\|_1 &= \sum_i \sum_j \Lambda_{ij}\, \mathbb{I}(i \in \mathcal{V}_{\text{row}} \text{ or } j \in \mathcal{V}_{\text{col}}) \\ &\leq \underbrace{\sum_i \sum_j \Lambda_{ij}\, \mathbb{I}(i \in \mathcal{V}_{\text{row}})}_{Z_i} \\ &\quad + \underbrace{\sum_j \sum_i \Lambda_{ij}\, \mathbb{I}(j \in \mathcal{V}_{\text{col}})}_{Z'_j}. \end{aligned} \tag{22}$$

The variables $\{Z_i, i \in [n]\}$ are independent. Below we show that they are sub-exponential, for which we recall that a variable X is called sub-exponential with parameter λ if

$$\mathbb{E}e^{t(X-\mathbb{E}X)} \leq e^{t^2\lambda^2/2}, \quad \text{for all } t \in \mathbb{R} \text{ with } |t| \leq \frac{1}{\lambda}. \tag{23}$$

It is a standard fact [52] that sub-exponential variables can be equivalently defined in terms of the tail condition

$$\mathbb{P}\{|X| \geq z\} \leq 2e^{-8z/\lambda}, \quad \text{for all } z \geq 0. \tag{24}$$

For the sake of completeness and tracking explicit constant values, we prove in Appendix II-F the one-sided implication (24)⇒(23), which is what we need below.

To verify the condition (24) for each Z_i, we observe that by definition either $Z_i = 0$ or $Z_i \geq 40pn$. Therefore, for each number $z \geq 40pn$, we have the tail bound

$$\begin{aligned}\mathbb{P}\{Z_i \geq z\} &= \mathbb{P}\left\{\text{pos}_{i\bullet}(\mathbf{\Lambda}) \geq z\right\} \\ &\overset{(i)}{\leq} \exp\left\{-\frac{(z-pn)^2}{2z}\right\} \\ &\leq \exp\left\{-\frac{(z-z/40)^2}{2z}\right\} \\ &\leq e^{-z/4},\end{aligned}$$

where in step (i) we use the previously established bound (21). For each $0 < z < 40pn$, we use the bound (21) again to get

$$\begin{aligned}\mathbb{P}\{Z_i \geq z\} &= \mathbb{P}\{Z_i \geq 40pn\} \\ &= \mathbb{P}\left\{\text{pos}_{i\bullet}(\mathbf{\Lambda}) \geq 40pn\right\} \\ &\leq e^{-8pn} \leq e^{-z/4}.\end{aligned}$$

We conclude that $\mathbb{P}\{Z_i \geq z\} \leq e^{-z/4}$ for all $z \geq 0$. Since Z_i is non-negative, it satisfies the tail condition (24) with $\lambda = 32$ and hence is sub-exponential as claimed. Moreover, the non-negative variable Z_i satisfies the expectation bound

$$\begin{aligned}\mathbb{E}Z_i &= \int_0^{20pn} \mathbb{P}\{Z_i > z\}\,\mathrm{d}z + \int_{20pn}^{\infty} \mathbb{P}\{Z_i > z\}\,\mathrm{d}z \\ &\overset{(i)}{\leq} \int_0^{20pn} e^{-8pn}\mathrm{d}z + \int_{20pn}^{\infty} e^{-z/4}\mathrm{d}z \\ &= 20pn \cdot e^{-8pn} + 4e^{-5pn} \\ &\overset{(ii)}{\leq} 24pne^{-5pn},\end{aligned}$$

where step (i) follows the previously established tail bounds for Z_i, and step (ii) follows from the assumption that $p \geq C_p/n$ for some sufficiently large constant $C_p > 0$. It follows that $\mathbb{E}\left[\sum_i Z_i\right] \leq 24pn^2e^{-5pn}$.

To control the sum $\sum_i Z_i$ in equation (22), we use a Bernstein-type inequality for the sum of sub-exponential random variables.

Lemma 6 (Theorem 1.13 of [53]): Suppose that $X_1, \ldots, X_n$ are independent random variables, each of which is sub-exponential with parameter λ in the sense of (23). Then for each $t \geq 0$, we have

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left\{-\frac{1}{2}\min\left(\frac{t^2}{\lambda^2 n}, \frac{t}{\lambda}\right)\right\}.$$

For a sufficiently large constant $C_2 > 1$, we apply the lemma above to obtain the tail bound

$$\begin{aligned}&\mathbb{P}\left\{\sum_i Z_i \geq 2C_2pn^2e^{-5pn}\right\} \\ &\overset{(i)}{\leq} \mathbb{P}\left\{\sum_i (Z_i - \mathbb{E}Z_i) \geq C_2pn^2e^{-5pn}\right\} \\ &\overset{(ii)}{\leq} \exp\left\{-\frac{1}{2}\min\left(\frac{(C_2pn^2e^{-5pn})^2}{(32)^2 n}, \frac{C_2pn^2e^{-5pn}}{32}\right)\right\},\end{aligned}$$

where step (i) follows from our previous bound on $\mathbb{E}\left[\sum_i Z_i\right]$. To proceed, we recall the assumption in Lemma 5 that p satisfies $C_p/n \leq p \leq c_d \log n/n$ for C_p sufficiently large and $c_d > 0$ sufficiently small, which implies that

$$\begin{aligned}\log\left(C_2pn^2e^{-5pn}\right) &= \log C_2 + \log(pn) + \log n - 5pn \\ &\geq \frac{3}{4}\log n,\end{aligned}$$

whence $C_2pn^2e^{-5pn} \geq n^{3/4}$. Plugging into the above tail bound, we get that

$$\begin{aligned}&\mathbb{P}\left\{\sum_i Z_i \geq 2C_2pn^2e^{-5pn}\right\} \\ &\leq \exp\left\{-\frac{1}{2}\min\left(\frac{n^{1/2}}{(32)^2}, \frac{n^{3/4}}{32}\right)\right\} \leq e^{-c'\sqrt{n}},\end{aligned}$$

for some absolute constant c'. The same tail bound holds for the sum $\sum_j Z'_j$. Combining these two bounds with the inequality (22), we obtain that

$$\|\mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}}\|_1 \leq 4C_2pn^2e^{-5pn}$$

with probability at least $1 - 2e^{-c'\sqrt{n}}$.

Putting together the above bounds on $\|\mathbb{E}\mathbf{\Lambda} - \widetilde{\mathbb{E}\mathbf{\Lambda}}\|_1$ and $\|\mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}}\|_1$, we conclude that there exists an absolute constant $C > 0$ such that

$$\|\mathbf{\Psi} - \widetilde{\mathbf{\Psi}}\|_1 \leq \|\mathbb{E}\mathbf{\Lambda} - \widetilde{\mathbb{E}\mathbf{\Lambda}}\|_1 + \|\mathbf{\Lambda} - \widetilde{\mathbf{\Lambda}}\|_1 \leq Cpn^2e^{-pn/C_e}$$

with probability at least $1 - 4e^{-c'\sqrt{n}}$, as desired.

V. Proof of Theorem 3

We only need to show that the conclusion of Theorem 1 (the bound on $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$) holds under the Heterogeneous SBM in Model 2. Once this is established, the conclusion in Theorem 2 follows immediately, as the clustering error is deterministically upper bounded by $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$ (cf. Proposition I).

To achieve the goal above, we first consider a model that bridges Model 1 and Model 2, and prove robustness under this intermediate model.

Model 4 (Semi-Random Stochastic Block Model): Given an *arbitrary* set of node pairs $\mathcal{L} \subseteq \{(i,j) \in [n] \times [n] : i < j\}$ and the ground-truth clustering $\boldsymbol{\sigma}^*$ (encoded in the matrix $\mathbf{Y}^*$), a graph is generated as follows. For each $(i,j) \in \mathcal{L}$, $A_{ij} = Y^*_{ij}$ deterministically. For each $(i,j) \notin \mathcal{L}$, A_{ij} is generated according to the standard SBM in Model 1; that is, $A_{ij} \sim \text{Bern}(p)$ if $Y^*_{ij} = 1$, and $A_{ij} \sim \text{Bern}(q)$ if $Y^*_{ij} = 0$.

Note that the standard SBM in Model 1 is a special case of the Semi-random SBM in Model 4 with an empty set $\mathcal{L}$. Model 4 is in turn a special case of the Heterogeneous SBM in Model 2 where $A_{ij} \sim \text{Bern}(\mathbb{I}(Y^*_{ij} = 1))$ for each $(i,j) \in \mathcal{L}$, and $A_{ij} \sim \text{Bern}\left(p\,\mathbb{I}(Y^*_{ij} = 1) + q\,\mathbb{I}(Y^*_{ij} = 0)\right)$ for each $(i,j) \notin \mathcal{L}$.

We first show that the conclusion of Theorem 1 holds under Model 4. We do so by verifying the steps of the proof of Theorem 1 given in Section IV. The proof consists of Fact 1 concerning the expected graph $\mathbb{E}\mathbf{A}$, and Propositions 1 and 2, which bound the deviation terms S_1 and S_2. Under Model 4, Fact 1 remains valid: each entry of $\mathbf{A}$ in the set $\mathcal{L}$ is changed in the direction of the sign of the corresponding entries of $\widehat{\mathbf{Y}} - \mathbf{Y}^*$, which only increases the LHS of equation (10). Both terms S_1 and S_2 involve the noise matrix $\mathbf{W} := \mathbf{A} - \mathbb{E}\mathbf{A}$. Examining the proofs of Propositions 1 and 2, we see that they only rely on the independence of the entries of $\boldsymbol{\Psi} := \boldsymbol{\Lambda} - \mathbb{E}\boldsymbol{\Lambda}$ (that is, the upper triangular entries of $\mathbf{W} := \mathbf{A} - \mathbb{E}\mathbf{A}$) as well as upper bounds on their expectations and variances. Under Model 4, the entries of $\mathbf{A}$ indexed by $\mathcal{L}$ are deterministically set to 1 or 0; the corresponding entries of $\boldsymbol{\Psi}$ become Bern(0), in which case they remain independent, with their expectations and variances decreased to zero.[5] Therefore, Propositions 1 and 2, and hence all the steps in the proof of Theorem 1, continue to hold under Model 4.

We now turn to the Heterogeneous SBM in Model 2, and show that this model can be reduced to Model 4 via a coupling argument. Suppose that under Model 2, $A_{ij} \sim \text{Bern}(p_{ij})$ for each (i,j) with $Y^*_{ij} = 1$, and $A_{ij} \sim \text{Bern}(q_{ij})$ for each (i,j) with $Y^*_{ij} = 0$, where by assumption $p_{ij} \geq p$ and $q_{ij} \leq q$. Model 2 is equivalent to the following 3-step process:

Step 1: We first generate a set of edge pairs $\mathcal{L}$ as follows: independently for each $(i,j), i < j$, if $Y^*_{ij} = 1$, then (i,j) is included in $\mathcal{L}$ with probability $1 - \frac{1-p_{ij}}{1-p}$; if $Y^*_{ij} = 0$, then (i,j) is included in $\mathcal{L}$ with probability $1 - \frac{q_{ij}}{q}$.

Step 2: Independently of above, we sample a graph from Model 1; let $\mathbf{A}^0$ denote its adjacency matrix.

Step 3: The final graph $\mathbf{A}$ is constructed as follows: for each (i,j), $i < j$ with $Y^*_{ij} = 1$, $A_{ij} = \mathbb{I}\left(A^0_{ij} = 1 \text{ or } (i,j) \in \mathcal{L}\right)$; for each (i,j), $i < j$ with $Y^*_{ij} = 0$, $A_{ij} = \mathbb{I}\left(A^0_{ij} = 1 \text{ and } (i,j) \notin \mathcal{L}\right)$.

Note that the assumption $p_{ij} \geq p$ and $q_{ij} \leq q$ ensures that the probabilities in step 1 are in $[0,1]$. One can verify that the distribution of $\mathbf{A}$ is the same as in Model 2; in particular, for each (i,j), $i < j$, we have

$$\begin{aligned}
&\mathbb{P}(A_{ij} = 1) \\
&\quad = \begin{cases} \mathbb{P}\left(A^0_{ij} = 1 \text{ or } (i,j) \in \mathcal{L}\right), & \text{if } Y^*_{ij} = 1, \\ \mathbb{P}\left(A^0_{ij} = 1 \text{ and } (i,j) \notin \mathcal{L}\right), & \text{if } Y^*_{ij} = 0. \end{cases} \\
&\quad = \begin{cases} 1 - (1-p) \cdot \frac{1-p_{ij}}{1-p} = p_{ij}, & \text{if } Y^*_{ij} = 1, \\ q \cdot \frac{q_{ij}}{q} = q_{ij}, & \text{if } Y^*_{ij} = 0. \end{cases}
\end{aligned}$$

Now, conditioned on the set $\mathcal{L}$, the distribution of $\mathbf{A}$ follows Model 4, since steps 1 and 2 are independent. We have established above that under Model 4, the error bound in Theorem 1 holds with high probability. Integrating out the randomness of the set $\mathcal{L}$ proves that the error bound holds with the same probability under Model 2.

Appendix I
Approximate k-Medians Algorithm

We describe the procedure for extracting an explicit clustering $\widehat{\boldsymbol{\sigma}}$ from the solution $\widehat{\mathbf{Y}}$ of the SDP (1) or (6). Our procedure builds on the idea in [15], and applies a version of the k-median algorithm to the rows of $\widehat{\mathbf{Y}}$, viewed as n points in $\mathbb{R}^n$. The k-median algorithm seeks a partition of these n points into k clusters, and a center associated with each cluster, such that the sum of the ℓ_1 distances of each point to its cluster center is minimized. Note that this procedure differs from the standard k-means algorithm: the latter minimizes the sum of *squared* distance, and uses the ℓ_2 distance rather than ℓ_1.

To formally specify the algorithm, we need some additional notations. Let $\text{Rows}(\mathbf{M})$ be the set of rows in the matrix $\mathbf{M}$. Define $\mathbb{M}_{n,k}$ to be the set of membership matrices corresponding to all k-partitions of $[n]$; that is, $\mathbb{M}_{n,k}$ is the set of matrices in $\{0,1\}^{n \times k}$ such that each of their rows has exactly one entry equal to 1 and each of their columns has at least one entry equal to 1. By definition each element in $\mathbb{M}_{n,k}$ has exactly k distinct rows. The k-medians algorithm described above can be written as an optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{\Psi}, \mathbf{X}} \ & \|\boldsymbol{\Psi}\mathbf{X} - \widehat{\mathbf{Y}}\|_1 \\
\text{s.t.} \ & \boldsymbol{\Psi} \in \mathbb{M}_{n,k}, \\
& \mathbf{X} \in \mathbb{R}^{k \times n}, \ \text{Rows}(\mathbf{X}) \subset \text{Rows}(\widehat{\mathbf{Y}})
\end{aligned} \tag{25}$$

Here the optimization variable $\boldsymbol{\Psi}$ encodes a partition of n points, assigning the i-th point to the cluster indexed by the unique non-zero element of the i-th row of $\boldsymbol{\Psi}$; the rows of the

[5] We illustrate this argument using Section IV-D as an example. There we would like to upper bound the quantity $\|\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}}\|_1$. Suppose that under Model 4, the entry $(i,j) \in \mathcal{L}$ is such that $A_{ij} = Y^*_{ij} = 1$ surely. In this case $\Psi_{ij} \sim \text{Bern}(0)$, which can be written as $\Psi_{ij} = \Lambda_{ij} - \mathbb{E}\Lambda_{ij}$ with $\mathbb{E}\Lambda_{ij} = 0 \leq p$ and $\text{Var}(\Lambda_{ij}) = 0 \leq p(1-p)$. This is all that is required when we proceed to bound $\|\mathbb{E}\boldsymbol{\Lambda} - \widetilde{\mathbb{E}\boldsymbol{\Lambda}}\|_1$ and $\|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\|_1$.

Algorithm 1 Approximate k-Median ρ-kmed

Input: data matrix $\widehat{\mathbf{Y}} \in \mathbb{R}^{n\times n}$; approximation factor $\rho \geq 1$.
1. Use a ρ-approximation algorithm to solve the optimization problem (25). Denote the solution as $(\check{\mathbf{\Psi}}, \check{\mathbf{X}})$.
2. For each $i \in [n]$, set $\widehat{\sigma}_i \in [k]$ to be the index of the unique non-zero element of $\left\{\check{\Psi}_{1i}, \check{\Psi}_{2i}, \ldots, \check{\Psi}_{ki}\right\}$; assign node i to cluster $\widehat{\sigma}_i$.
Output: Clustering assignment $\widehat{\boldsymbol{\sigma}} \in [k]^n$.

variable $\mathbf{X}$ represent the r cluster centers. Note that the last constraint in (25) stipulates that the cluster centers be elements of the input data points represented by the rows of $\widehat{\mathbf{Y}}$, so this procedure is sometimes called k-medoids. Let $(\overline{\mathbf{\Psi}}, \overline{\mathbf{X}})$ be the optimal solution of the problem (25).

Finding the exact solution $(\overline{\mathbf{\Psi}}, \overline{\mathbf{X}})$ is in general computationally intractable, but polynomial-time constant-factor approximation algorithms exist. In particular, we will make use of the polynomial-time procedure in [47], which produces an approximate solution $(\check{\mathbf{\Psi}}, \check{\mathbf{X}}) \in \mathbb{M}_{n,k} \times \mathbb{R}^{k\times n}$ that is guaranteed to satisfy Rows($\check{\mathbf{X}}$) $\subset$ Rows($\widehat{\mathbf{Y}}$) and

$$\|\check{\mathbf{\Psi}}\check{\mathbf{X}} - \widehat{\mathbf{Y}}\|_1 \leq \rho \|\overline{\mathbf{\Psi}}\,\overline{\mathbf{X}} - \widehat{\mathbf{Y}}\|_1 \tag{26}$$

with an approximation ratio $\rho = \frac{20}{3}$. The variable $\check{\mathbf{\Psi}}$ encodes our final clustering assignment.

The complete procedure ρ-kmed is given in Algorithm 1, which takes as input a data matrix $\widehat{\mathbf{Y}}$ (which in our case is the solution to the SDP (1) or (6)) and outputs an explicit clustering $\widehat{\boldsymbol{\sigma}} = \rho$-kmed$(\widehat{\mathbf{Y}})$. We assume that the number of clusters k is known.

Appendix II
Technical Lemmas

In this section we provide the proofs of the technical lemmas used in the main text.

A. Proof of Fact 1

For each pair $i \neq j$, if nodes i and j belong to the same cluster, then $Y^*_{ij} = 1 \geq \widehat{Y}_{ij}$ and $\mathbb{E}A_{ij} - \frac{p+q}{2} = p - \frac{p+q}{2} = \frac{p-q}{2}$; if nodes i and j belong to different clusters, then $Y^*_{ij} = 0 \leq \widehat{Y}_{ij}$ and $\mathbb{E}A_{ij} - \frac{p+q}{2} = q - \frac{p+q}{2} = -\frac{p-q}{2}$. For $i = j$, we have $\widehat{Y}_{ij} = Y^*_{ij}$. In each case, we have the expression

$$\left(Y^*_{ij} - \widehat{Y}_{ij}\right)\left(\mathbb{E}A_{ij} - \frac{p+q}{2}\right) = \frac{p-q}{2}\left|\widehat{Y}_{ij} - Y^*_{ij}\right|.$$

Summing up the above equation for all $i, j \in [n]$ gives the desired equation (10).

B. Proof of Lemma 2

To establish a uniform bound in β, we apply a discretization argument to the possible values of β. Define the shorthand $E := (3me^{-g\rho/C_e}, m]$. We can cover E by the subintervals $E_t := (t-1, t]$ for integers $t \in [\lceil 3me^{-g\rho/C_e}\rceil, m]$. By construction we have

$$\frac{g\rho}{C_e} = \log\frac{3m}{3me^{-g\rho/C_e}} \geq \log\frac{3m}{t}. \tag{27}$$

We also know that our choice of $D = \sqrt{10}$ satisfies $\frac{1}{2}D^2 \geq 4\left(1 + \frac{1}{3\sqrt{C_e}}D\right)$ since $C_e \geq 28$.

We define the shorthand

$$R_\beta := D\,\lceil\beta\rceil\sqrt{g\rho\log(3m/\beta)}.$$

For each $t \in [\lceil 3me^{-g\rho/C_e}\rceil, m]$ we define the probabilities

$$\alpha_t := \mathbb{P}\left\{\exists\beta \in E_t : \sum_{j=1}^{\lceil\beta\rceil}\left|X_{(j)}\right| > R_\beta\right\}.$$

We bound each of these probabilities:

$$\begin{aligned}
\alpha_t &\overset{(i)}{\leq} \mathbb{P}\left\{\sum_{j=1}^{t}\left|X_{(j)}\right| > R_t\right\} \\
&= \mathbb{P}\left\{\bigcup_{i_1<\cdots<i_t}\left\{\sum_{j=1}^{t}\left|X_{i_j}\right| > R_t\right\}\right\} \\
&\leq \sum_{i_1<\cdots<i_t}\mathbb{P}\left\{\sum_{j=1}^{t}\left|X_{i_j}\right| > R_t\right\} \\
&= \sum_{i_1<\cdots<i_t}\mathbb{P}\left\{\max_{\mathbf{u}\in\{\pm1\}^t}\sum_{j=1}^{t}X_{i_j}u_j > R_t\right\} \\
&\leq \sum_{i_1<\cdots<i_t}\sum_{\mathbf{u}\in\{\pm1\}^t}\mathbb{P}\left\{\sum_{j=1}^{t}X_{i_j}u_j > R_t\right\},
\end{aligned} \tag{28}$$

where step (i) holds since $\beta \in E_t$ implies $\beta \leq \lceil\beta\rceil = t$.

For each positive integer t, fix an index set $(i_j)_{j=1}^t$ and a sign pattern $\mathbf{u} \in \{\pm1\}^t$. Note that $\sum_{j=1}^t X_{i_j}u_j$ is the sum of tg centered Bernoulli random variables with $\mathrm{Var}(B_{ij} - \mathbb{E}B_{ij}) \leq \rho$ and $\left|B_{ij} - \mathbb{E}B_{ij}\right| \leq 1$ for each $j \in [t]$ and $i \in [g]$. We apply Bernstein inequality to bound the probability on the RHS of equation (28):

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{j=1}^{t}X_{i_j}u_j > R_t\right\} \\
&\quad\leq \exp\left\{-\frac{\frac{1}{2}D^2t^2g\rho\log(3m/t)}{tg\rho + \frac{1}{3}Dt\sqrt{g\rho\log(3m/t)}}\right\} \\
&\quad\overset{(a)}{\leq} \exp\left\{-\frac{\frac{1}{2}D^2t^2g\rho\log(3m/t)}{\left(1 + \frac{1}{3\sqrt{C_e}}D\right)tg\rho}\right\} \\
&\quad\overset{(b)}{\leq} \exp\left\{-4t\log(3m/t)\right\},
\end{aligned}$$

where step (a) holds by equation (27) and step (b) holds by our choice of the constant D. Plugging this back to

equation (28), we get that for each $\lceil 3me^{-g\rho/C_e} \rceil \le t \le m$,

$$\begin{aligned}\alpha_t &\le \sum_{i_1<\cdots<i_t} \sum_{\mathbf{u}\in\{\pm 1\}^t} \exp\{-4t\log(3m/t)\} \\ &= \binom{m}{t} \cdot 2^t \cdot \exp\{-4t\log(3m/t)\} \\ &\le \left(\frac{me}{t}\right)^t e^t \exp\{-4t\log(3m/t)\} \\ &= \exp\{t\log(m/t) + 2t - 4t\log(3m/t)\} \\ &\le \exp\{-t\log(3m/t)\} = \left(\frac{t}{3m}\right)^t, \end{aligned} \tag{29}$$

where the last inequality follows from $t \le t\log(3m/t)$ since $m \ge t$. It follows that

$$\begin{aligned}&\mathbb{P}\left\{\exists \beta \in E : \sum_{j=1}^{\lceil \beta \rceil} \left|X_{(j)}\right| > R_\beta\right\} \\ &\quad\le \sum_{t=\lceil 3me^{-\rho g/C_e}\rceil}^{m} \alpha_t \\ &\quad\le \sum_{t=\lceil 3me^{-\rho g/C_e}\rceil}^{m} \left(\frac{t}{3m}\right)^t \\ &\quad\le \sum_{t=1}^{m} \left(\frac{t}{3m}\right)^t =: P_1(m).\end{aligned}$$

It remains to show that $P_1(m) \le \frac{3}{m}$. Since

$$\begin{aligned}P_1(m) &\le \frac{1}{3m} + \sum_{t=2}^{m} \left(\frac{t}{3m}\right)^t \\ &\le \frac{2}{m} + m \cdot \max_{t=2,3,\ldots,m} \left(\frac{t}{3m}\right)^t,\end{aligned}$$

the proof is completed if for each integer $t = 2, 3, \ldots, m$, we can show the bound $\left(\frac{t}{3m}\right)^t \le \frac{1}{m^2}$, or equivalently $f(t) := t(\log 3m - \log t) \ge 2\log m$. Since $t \le m$, $f(t)$ has derivative

$$f'(t) = \log 3m - \log t - 1 \ge \log 3 - 1 \ge 0.$$

Therefore, $f(t)$ is non-decreasing for $2 \le t \le m$ and hence $f(t) \ge f(2) = 2\log 3m - 2\log 2 \ge 2\log m$. Hence, we have $\left(\frac{t}{3m}\right)^t \le \frac{1}{m^2}$ and $P_1(m) \le \frac{3}{m}$. □

C. Proof of Lemma 3

Our proof follows similar lines as that of [15, Lemma 3]. We first bound $\mathbb{E}\|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\text{op}}$ using a standard symmetrization argument. Let $\mathbf{A}'$ be an independent copy of $\mathbf{A}$, and $\mathbf{R}$ be an $n \times n$ symmetric matrix, independent of both $\mathbf{A}$ and $\mathbf{A}'$, with i.i.d. Rademacher entries on and above its diagonal. We can compute

$$\begin{aligned}\mathbb{E}\|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\text{op}} &= \mathbb{E}\|\mathbf{A} - \mathbb{E}\mathbf{A}'\|_{\text{op}} \\ &\overset{(i)}{\le} \mathbb{E}\|\mathbf{A} - \mathbf{A}'\|_{\text{op}} \\ &\overset{(ii)}{=} \mathbb{E}\|\left(\mathbf{A} - \mathbf{A}'\right) \circ \mathbf{R}\|_{\text{op}} \\ &\overset{(iii)}{\le} 2\mathbb{E}\|\mathbf{A} \circ \mathbf{R}\|_{\text{op}},\end{aligned}$$

where step (i) follows from convexity of the spectral norm, step (ii) holds since the matrices $\mathbf{A} - \mathbf{A}'$ and $\left(\mathbf{A} - \mathbf{A}'\right) \circ \mathbf{R}$ are identically distributed, and step (iii) follows from the triangle inequality.

We proceed by bounding $\mathbb{E}\|\mathbf{A} \circ \mathbf{R}\|_{\text{op}}$. Write $\|\zeta\|_a := \mathbb{E}[\zeta^a]^{1/a}$ for a random variable ζ and an integer $a \ge 0$. Define the scalars $(b_{ij})_{i\ge j}$ by $b_{ij} = \sqrt{\mathbb{E}A_{ij}}$. Also define the independent random variables $(\xi_{ij})_{i\ge j}$ as follows: if $\mathbb{E}A_{ij} = 0$, then ξ_{ij} is a Rademacher random variable; if $\mathbb{E}A_{ij} > 0$,

$$\xi_{ij} = \begin{cases} \frac{1}{\sqrt{\mathbb{E}A_{ij}}}, & \text{with probability } \mathbb{E}A_{ij}/2, \\ -\frac{1}{\sqrt{\mathbb{E}A_{ij}}}, & \text{with probability } \mathbb{E}A_{ij}/2, \\ 0, & \text{with probability } 1 - \mathbb{E}A_{ij}. \end{cases}$$

It can be seen that the lower triangular entries of the symmetric matrix $\mathbf{A} \circ \mathbf{R}$ can be written as $A_{ij}R_{ij} = \xi_{ij}b_{ij}$. By definition, the random variables $(\xi_{ij})_{i\ge j}$ are symmetric and have zero mean and unit variance. We can therefore make use of the following known result:

Lemma 7 (Corollary 3.6 in [54]): Let $\mathbf{X}$ be the $n \times n$ symmetric random matrix with $X_{ij} = \xi_{ij}b_{ij}$, where $\{\xi_{ij}, i \ge j\}$ are independent symmetric random variables with unit variance and $\{b_{ij} : i \ge j\}$ are given scalars. Then we have for any $\alpha \ge 3$,

$$\mathbb{E}\|\mathbf{X}\|_{\text{op}} \le e^{2/\alpha}\left\{2\sigma + 14\alpha \max_{ij} \|\xi_{ij}b_{ij}\|_{2\lceil \alpha \log n\rceil}\sqrt{\log n}\right\},$$

where $\sigma := \max_i \sqrt{\sum_j b_{ij}^2}$.

For every pair $i \ge j$, the random variable $\xi_{ij}b_{ij}$ is surely bounded by 1 in absolute value and thus satisfies $\|\xi_{ij}b_{ij}\|_{2\lceil\alpha\log n\rceil} \le 1$ for any $\alpha \ge 3$. The scalars $(b_{ij})_{i\ge j}$ are all bounded by $\sqrt{p}$, so $\sigma \le \sqrt{pn}$. Applying the above lemma with $\alpha = 3$ gives

$$\begin{aligned}\mathbb{E}\|\mathbf{A} \circ \mathbf{R}\|_{\text{op}} &\le e^{2/3}\left(2\sqrt{pn} + 42\sqrt{\log n}\right) \\ &\le 4\sqrt{pn} + 84\sqrt{\log n},\end{aligned}$$

whence

$$\mathbb{E}\|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\text{op}} \le 2\mathbb{E}\|\mathbf{A} \circ \mathbf{R}\|_{\text{op}} \le 8\sqrt{pn} + 168\sqrt{\log n}.$$

We complete the proof by bounding the deviation of $\|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\text{op}}$ from its expectation. This can be done using a standard Lipschitz concentration inequality:

Lemma 8 (Theorem 6.10 in [55], Generalized in Exercise 6.5 Therein): Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex compact set with diameter B. Let $X_1, \ldots, X_N$ be independent random variables taking values in $\mathcal{X}$ and assume that $f : \mathcal{X}^N \to \mathbb{R}$ is separately convex and 1-Lipschitz, that is, $|f(x) - f(y)| \le \|x - y\|$ for all $x, y \in \mathcal{X}^N \subset \mathbb{R}^{dN}$. Then $Z = f(X_1, \ldots, X_N)$ satisfies, for all $t > 0$,

$$\mathbb{P}\{Z > \mathbb{E}Z + t\} \le e^{-t^2/(2B^2)}.$$

To use this result for our purpose, we note that $Z = \|\mathbf{A} - \mathbb{E}\mathbf{A}\|_{\text{op}}/\sqrt{2}$ is a function of the $N = n(n-1)/2$ independent lower triangular entries of the symmetric matrix $\mathbf{A} - \mathbb{E}\mathbf{A}$. Moreover, this function is separately convex and

1-Lipschitz. In our setting, each entry of $\mathbf{A}-\mathbb{E}\mathbf{A}$ takes values in the interval $\mathcal{X}=[-p,1-q]$, which is convex compact with diameter $B=1-q+p\le 2$. Applying the above lemma yields that for each $t\ge 0$,

$$\mathbb{P}\left\{\frac{1}{\sqrt{2}}\|\mathbf{A}-\mathbb{E}\mathbf{A}\|_{\mathrm{op}}>\frac{1}{\sqrt{2}}\mathbb{E}\|\mathbf{A}-\mathbb{E}\mathbf{A}\|_{\mathrm{op}}+t\right\}\le e^{-t^2/(2B^2)}\le e^{-t^2/8}.$$

Choosing $t=3\sqrt{2\log n}$ and combining with the previous bound on $\mathbb{E}\|\mathbf{A}-\mathbb{E}\mathbf{A}\|_{\mathrm{op}}$, gives the desired inequality

$$\mathbb{P}\left\{\|\mathbf{A}-\mathbb{E}\mathbf{A}\|_{\mathrm{op}}>8\sqrt{pn}+174\sqrt{\log n}\right\}\le n^{-2}.$$

D. Proof of Lemma 4

For future reference, we state below a more general result; Lemma 4 is a special case with $\sigma^2=p$.

Lemma 9: Suppose that $\mathbf{X}\in\mathbb{R}^{n\times n}$ is a random matrix with independent entries of the following distributions

$$X_{ij}=\begin{cases}1-p_{ij} & \text{with probability } p_{ij}\\ -p_{ij} & \text{with probability } 1-p_{ij}.\end{cases}$$

Let σ be a number that satisfies $p_{ij}\le\sigma^2$ for all (i,j), and $\overleftrightarrow{\mathbf{X}}$ be the matrix obtained from $\mathbf{X}$ by zeroing out all the rows and columns having more than $40\sigma^2 n$ positive entries. Then with probability at least $1-P_3=1-e^{-(3-\ln 2)2n}-(2n)^{-3}$, $\|\overleftrightarrow{\mathbf{X}}\|_{\mathrm{op}}\le C\sigma\sqrt{n}$ for some absolute constant $C>0$.

We now prove this lemma. Note that the matrix $\mathbf{X}$ is not symmetric. To make use of a known result that requires symmetry, we employ a standard dilation argument. We construct the matrix

$$\mathbf{D}:=\begin{bmatrix}\mathbf{O}_1 & \mathbf{X}\\ \mathbf{X}^\top & \mathbf{O}_2\end{bmatrix}\in\mathbb{R}^{2n\times 2n},$$

where $\mathbf{O}_1,\mathbf{O}_2\in\mathbb{R}^{n\times n}$ are all-zero matrices. The matrices $\mathbf{O}_1$ and $\mathbf{O}_2$ can be viewed as random symmetric matrices whose entries above the diagonal are independent centered Bernoulli random variable of rate zero. The matrix $\mathbf{D}$ is symmetric and satisfies the assumptions in the following known result with $N=2n$.

Lemma 10 (Lemma 12 of [31]): Suppose that $\mathbf{D}\in\mathbb{R}^{N\times N}$ is a random symmetric matrix with zero on the diagonal whose entries above the diagonal are independent with the following distributions

$$D_{ij}=\begin{cases}1-p_{ij} & \text{with probability } p_{ij}\\ -p_{ij} & \text{with probability } 1-p_{ij}.\end{cases}$$

Let σ be a quantity such that $p_{ij}\le\sigma^2$ for all $(i,j)\in[N]\times[N]$, and $\mathbf{D}_1$ be the matrix obtained from $\mathbf{D}$ by zeroing out all the rows and columns having more than $20\sigma^2 N$ positive entries. Then with probability $1-o(1)$, $\|\mathbf{D}_1\|_{\mathrm{op}}\le C'\sigma\sqrt{N}$ for some absolute constant $C'>0$.

Remark 1: The probability above is in fact $1-o(1)=1-P_3=1-e^{-(3-\ln 2)N}-N^{-3}$, as can be seen by inspecting the proof of the lemma. Moreover, since $\mathbf{D}$ is symmetric, the indices of the rows and columns being zeroed out are identical.

Lemma 10 ensures that $\|\mathbf{D}_1\|_{\mathrm{op}}\le C'\sigma\sqrt{2n}=C\sigma\sqrt{n}$ with probability at least $1-P_3$, where $C=\sqrt{2}C'$. The lemma at the beginning of this sub-section then follows from the claim that $\|\overleftrightarrow{\mathbf{X}}\|_{\mathrm{op}}=\|\mathbf{D}_1\|_{\mathrm{op}}$.

It remains to prove the above claim. Recall that $\mathrm{pos}_{i\bullet}(\mathbf{M})$ and $\mathrm{pos}_{\bullet j}(\mathbf{M})$ are the numbers of positive entries in the i-th row and j-th column of a matrix $\mathbf{M}$, respectively. For each $(i,j)\in[n]\times[n]$, we observe that by construction,

$$\mathrm{pos}_{i\bullet}(\mathbf{D})>20\sigma^2N\Leftrightarrow\mathrm{pos}_{i\bullet}(\mathbf{X})>40\sigma^2n$$

$$\mathrm{pos}_{\bullet(j+n)}(\mathbf{D})>20\sigma^2N\Leftrightarrow\mathrm{pos}_{\bullet j}(\mathbf{X})>40\sigma^2n$$

and similarly

$$\mathrm{pos}_{(i+n)\bullet}(\mathbf{D})>20\sigma^2N\Leftrightarrow\mathrm{pos}_{i\bullet}(\mathbf{X}^\top)>40\sigma^2n$$

$$\mathrm{pos}_{\bullet j}(\mathbf{D})>20\sigma^2N\Leftrightarrow\mathrm{pos}_{\bullet j}(\mathbf{X}^\top)>40\sigma^2n.$$

It then follows from the definitions of $\mathbf{D}$ and $\mathbf{D}_1$ that

$$\mathbf{D}_1=\begin{bmatrix}\mathbf{O}_1 & \overleftrightarrow{\mathbf{X}}\\ \overleftrightarrow{\mathbf{X}}^\top & \mathbf{O}_2\end{bmatrix}=\begin{bmatrix}\mathbf{O}_1 & \overleftrightarrow{\mathbf{X}}\\ \overleftrightarrow{\mathbf{X}}^\top & \mathbf{O}_2\end{bmatrix},$$

whence $\|\mathbf{D}_1\|_{\mathrm{op}}=\|\overleftrightarrow{\mathbf{X}}\|_{\mathrm{op}}$.

E. Proof of Fact 3

Because $\widehat{\mathbf{Y}}$ is feasible to the SDP (1), we know that $\|\widehat{\mathbf{Y}}\|_\infty\le 1$. Let $c(i)$ be the index of the cluster that contains node i. For each $(i,j)\in[n]\times[n]$, we have the bound

$$|(\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}})_{ij}|=\left|\sum_{t:c(t)=c(i)}(\mathbf{U}\mathbf{U}^\top)_{it}\widehat{Y}_{tj}\right|\le\ell\cdot\frac{1}{\ell}\cdot\|\widehat{\mathbf{Y}}\|_\infty\le 1$$

thanks to the structure of the matrix $\mathbf{U}\mathbf{U}^\top$. The same bound holds for the matrices $\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top$ and $\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top$ by similar arguments. It follows that

$$\begin{aligned}&\|\mathcal{P}_{T^\perp}(\widehat{\mathbf{Y}})\|_\infty\\ &\quad=\|\widehat{\mathbf{Y}}-\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}-\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top+\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\|_\infty\\ &\quad\le\|\widehat{\mathbf{Y}}\|_\infty+\|\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\|_\infty\\ &\qquad+\|\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\|_\infty+\|\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{Y}}\mathbf{U}\mathbf{U}^\top\|_\infty\\ &\quad\le 4.\end{aligned}$$

F. Proof of the Implication (24)⇒(23)

We first show that the tail condition (24) of a random variable X implies a bound for its moments. Note that we do not require X to be centered.

Lemma 11: Suppose that for some $\lambda>0$, the random variable X satisfies $\mathbb{P}(|X|>z)\le 2e^{-8z/\lambda},\forall z\ge 0$. Then for each positive integer $m\ge 1$,

$$\left(\mathbb{E}\left[|X|^m\right]\right)^{1/m}\le\frac{1}{4}\lambda\,(m!)^{1/m}.$$

Proof: Let $\Gamma(\cdot)$ denote the Gamma function. We have the bound

$$\begin{aligned}\mathbb{E}\left[|X|^m\right] &= \int_0^\infty \mathbb{P}\left(|X|^m > z\right) \mathrm{d}z \\ &= \int_0^\infty \mathbb{P}\left(|X| > z^{1/m}\right) \mathrm{d}z \\ &\leq \int_0^\infty 2\exp\left(-\frac{8z^{1/m}}{\lambda}\right) \mathrm{d}z \\ &\overset{(i)}{=} 2\cdot(\lambda/8)^m \cdot m\int_0^\infty e^{-u}u^{m-1}\,\mathrm{d}u \\ &\leq (\lambda/4)^m \cdot m\Gamma(m) = (\lambda/4)^m m!,\end{aligned}$$

where step (i) follows from a change of variable $u = 8z^{1/m}/\lambda$. Taking the m-th root of both sides proves the result. ∎

Using Minkowski's inequality, we can see that the above moment bounds are sub-additive:

Lemma 12: Suppose that for some $\lambda_1, \lambda_2 > 0$, the random variables X_1 and X_2 satisfy

$$\left(\mathbb{E}\left[|X_1|^m\right]\right)^{1/m} \leq \lambda_1 (m!)^{1/m},$$

$$\left(\mathbb{E}\left[|X_2|^m\right]\right)^{1/m} \leq \lambda_2 (m!)^{1/m}$$

for each positive integer m. Then for each positive integer m,

$$\left(\mathbb{E}\left[|X_1 + X_2|^m\right]\right)^{1/m} \leq (\lambda_1 + \lambda_2)(m!)^{1/m}.$$

Consequently, centering a random variable does not affect its sub-exponentiality up to constant factors. In particular, suppose that X satisfies the moment bounds in Lemma 11. Then for every positive integer m,

$$\begin{aligned}\left(\mathbb{E}\left[|\mathbb{E}X|^m\right]\right)^{1/m} &= \left(|\mathbb{E}X|^m\right)^{1/m} \\ &\overset{(i)}{\leq} \left(\mathbb{E}\,|X|^m\right)^{1/m} \\ &\leq \frac{\lambda}{4}(m!)^{1/m},\end{aligned}$$

where step (i) uses the Jensen's inequality. Applying Lemma 12 gives

$$\left(\mathbb{E}\left[|X - \mathbb{E}X|^m\right]\right)^{1/m} \leq \frac{\lambda}{2}(m!)^{1/m}.$$

The next lemma shows that the moment bound implies the bound (23) on the moment generating function, hence completing the proof of the implication (24)⇒(23).

Lemma 13: Suppose that for some number $\lambda > 0$, the random variable X satisfies

$$\left(\mathbb{E}\left[|X - \mathbb{E}X|^m\right]\right)^{1/m} \leq \frac{\lambda}{2}\cdot(m!)^{1/m}$$

for each positive integer m. Then we have

$$\mathbb{E}\left[e^{t(X-\mathbb{E}X)}\right] \leq e^{t^2\lambda^2/2}, \qquad \forall\, |t| \leq \frac{1}{\lambda}.$$

Proof: For each t such that $|t| \leq 1/\lambda$, we have

$$\begin{aligned}\mathbb{E}\left[e^{t(X-\mathbb{E}X)}\right] &= 1 + t\mathbb{E}\left[X - \mathbb{E}X\right] \\ &\quad + \sum_{m=2}^\infty \frac{t^m\mathbb{E}\left[(X-\mathbb{E}X)^m\right]}{m!} \\ &\overset{(i)}{\leq} 1 + \sum_{m=2}^\infty \frac{|t|^m\,\mathbb{E}\left[|X-\mathbb{E}X|^m\right]}{m!} \\ &\leq 1 + \sum_{m=2}^\infty \left(|t|\cdot\frac{\lambda}{2}\right)^m \\ &= 1 + \frac{t^2\lambda^2}{4}\sum_{m=0}^\infty \left(|t|\cdot\frac{\lambda}{2}\right)^m \\ &\leq 1 + \frac{t^2\lambda^2}{2} \leq e^{t^2\lambda^2/2},\end{aligned}$$

where step (i) holds since $\mathbb{E}\left[X - \mathbb{E}X\right] = 0$. ∎

Appendix III
Proof of Theorem 2

As mentioned in Appendix I, we use an approximate k-medians clustering algorithm with approximation ratio $\rho = \frac{20}{3}$. Theorem 2 follows immediately by combining Theorem 1 and the proposition below.

Proposition 3: The clustering assignment $\widehat{\boldsymbol{\sigma}} = \rho\text{-}\mathtt{kmed}(\widehat{\mathbf{Y}})$ produced by the ρ-approximate k-median algorithm (Algorithm 1) satisfies the error bound

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \leq 2(1+2\rho)\cdot\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1}.$$

We note that a similar result appeared in [15], specifically their Theorem 2 restricted to the non-degree-corrected setting. The proposition provides a moderate generalization, establishing a general relationship between clustering error of the ρ-approximation k-median procedure and the error of its input $\widehat{\mathbf{Y}}$. The rest of this section is devoted to the proof of Proposition 3.

Recall that $(\overline{\boldsymbol{\Psi}}, \overline{\mathbf{X}})$ is the optimal solution of the k-medians problem (25), and $(\check{\boldsymbol{\Psi}}, \check{\mathbf{X}})$ is a ρ-approximate solution. Set $\overline{\mathbf{Y}} := \overline{\boldsymbol{\Psi}}\overline{\mathbf{X}}$ and $\check{\mathbf{Y}} := \check{\boldsymbol{\Psi}}\check{\mathbf{X}}$. Note that the solution $(\check{\boldsymbol{\Psi}}, \check{\mathbf{X}})$ corresponds to at most k clusters. Without loss of generality we may assume that it actually contains exactly k clusters, and thus the cluster membership matrix $\check{\boldsymbol{\Psi}}$ is in $\mathbb{M}_{n,k}$ and has exactly k distinct rows. If this is not true, we can always move an arbitrary point from the n input points to a new cluster, without changing the k-medians objective value of the approximation solution $(\check{\boldsymbol{\Psi}}, \check{\mathbf{X}})$.

We next rewrite the cluster error metric (7) in matrix form. Let $\boldsymbol{\Psi}^* \in \mathbb{M}_{n,k}$ be the membership matrix corresponding to the ground truth clusters; that is, for each $i \in [n]$, $\Psi^*_{i\sigma^*_i}$ is the only non-zero element of the i-th row of $\boldsymbol{\Psi}^*$, and thus $\boldsymbol{\Psi}^*(\boldsymbol{\Psi}^*)^\top = \mathbf{Y}^*$. Let $\mathcal{S}_k$ be the set of $k\times k$ permutation matrices. The set of misclassified nodes with respect to a permutation $\boldsymbol{\Pi} \in \mathcal{S}_k$, is then given by

$$\mathcal{E}(\boldsymbol{\Pi}) := \left\{i \in [n] : \left(\check{\boldsymbol{\Psi}}\boldsymbol{\Pi}\right)_{i\bullet} \neq \boldsymbol{\Psi}^*_{i\bullet}\right\}.$$

With this notation, the error metric (7) can be expressed as $\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = \min_{\boldsymbol{\Pi}\in\mathcal{S}_k} n^{-1}\left|\mathcal{E}(\boldsymbol{\Pi})\right|$, and it remains to bound the RHS.

To this end, we construct several useful sets. For each $a \in [k]$, let $C_a^* = \{i \in [n] : \sigma_i^* = a\}$ be the a-th cluster, and we define the node index sets

$$T_a := \left\{i \in C_a^* : \|\check{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 < \ell\right\}$$

and $S_a := C_a^* \backslash T_a$. Let $T := \bigcup_{a\in[k]} T_a$ and $S := \bigcup_{a\in[k]} S_a$. Note that $S_1, \ldots, S_k, T_1, \ldots, T_k$ are disjoint with $T \cup S = [n]$. Further define the cluster index sets

$$\begin{aligned}
R_1 &:= \{a \in [k] : T_a = \emptyset\}, \\
R_2 &:= \left\{a \in [k] : T_a \neq \emptyset;\, \check{\boldsymbol{\Psi}}_{i\bullet} = \check{\boldsymbol{\Psi}}_{j\bullet}, \forall i, j \in T_a\right\}, \\
R_3 &:= [k] \setminus (R_1 \cup R_2) \\
&= \left\{a \in [k] : T_a \neq \emptyset;\, \check{\boldsymbol{\Psi}}_{i\bullet} \neq \check{\boldsymbol{\Psi}}_{j\bullet}, \exists i, j \in T_a\right\}.
\end{aligned}$$

Note that R_1, R_2 and R_3 are disjoint sets. With the above notations, we have the following claims.

Claim 1: $\min_{\boldsymbol{\Pi}\in\mathcal{S}_k} |\mathcal{E}(\boldsymbol{\Pi})| \le |S| + |R_3|\,\ell$.

Claim 2: $|R_3| \le |R_1|$.

Claim 3: $|R_1|\,\ell \le |S| \le \frac{1}{\ell}\|\check{\mathbf{Y}} - \mathbf{Y}^*\|_1$.

Claim 4: $\|\check{\mathbf{Y}} - \mathbf{Y}^*\|_1 \le (1+2\rho)\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$.

Applying the above claims in order, we obtain that

$$\begin{aligned}
\min_{\boldsymbol{\Pi}\in\mathcal{S}_k} n^{-1}\left|\mathcal{E}(\boldsymbol{\Pi})\right| &\le \frac{2(1+2\rho)}{n\ell}\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \\
&= 2(1+2\rho)\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1},
\end{aligned}$$

where the last equality follows from $\|\mathbf{Y}^*\|_1 = n\ell$. Combining with the aforementioned expression $\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = \min_{\boldsymbol{\Pi}\in\mathcal{S}_k} n^{-1}\left|\mathcal{E}(\boldsymbol{\Pi})\right|$ proves Proposition 3.

We prove the the above claims in the sub-sections to follow.

A. Proof of Claim 1

We record a property of the cluster membership matrix $\check{\boldsymbol{\Psi}}$ of the approximate k-medians solution.

Lemma 14: For each cluster pair $a, b \in R_2 \cup R_3$ with $a \neq b$ and each node pair $i \in T_a$, $j \in T_b$, we have $\check{\boldsymbol{\Psi}}_{i\bullet} \neq \check{\boldsymbol{\Psi}}_{j\bullet}$.

Proof: For each pair $a, b \in R_2 \cup R_3$ with $a \neq b$, we have $T_a \neq \emptyset$ and $T_b \neq \emptyset$ by definition. For each pair $i \in T_a$, $j \in T_b$, we have the inequality

$$\begin{aligned}
\|\check{\mathbf{Y}}_{i\bullet} - \check{\mathbf{Y}}_{j\bullet}\|_1 \ge\ & \|\mathbf{Y}^*_{i\bullet} - \mathbf{Y}^*_{j\bullet}\|_1 \\
& - \|\check{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 - \|\check{\mathbf{Y}}_{j\bullet} - \mathbf{Y}^*_{j\bullet}\|_1.
\end{aligned}$$

Since nodes i and j are in two different clusters, each of $\mathbf{Y}^*_{i\bullet}$ and $\mathbf{Y}^*_{j\bullet}$ is a binary vector with exactly ℓ ones, and they have disjoint support; we therefore have $\|\mathbf{Y}^*_{i\bullet} - \mathbf{Y}^*_{j\bullet}\|_1 = 2\ell$. Moreover, note that $\|\check{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 < \ell$ and $\|\check{\mathbf{Y}}_{j\bullet} - \mathbf{Y}^*_{j\bullet}\|_1 < \ell$ by definition of T_a and T_b. It follows that $\|\check{\mathbf{Y}}_{i\bullet} - \check{\mathbf{Y}}_{j\bullet}\|_1 > 2\ell - \ell - \ell = 0$, and thus $\check{\mathbf{Y}}_{i\bullet} \neq \check{\mathbf{Y}}_{j\bullet}$. The latter implies that $\check{\boldsymbol{\Psi}}_{i\bullet} \neq \check{\boldsymbol{\Psi}}_{j\bullet}$; otherwise we would reach a contradiction $\check{\mathbf{Y}}_{i\bullet} = \left(\check{\boldsymbol{\Psi}}_{i\bullet}\right)^\top \check{\mathbf{X}} = \left(\check{\boldsymbol{\Psi}}_{j\bullet}\right)^\top \check{\mathbf{X}} = \check{\mathbf{Y}}_{j\bullet}$. ∎

Proceeding to the proof of Claim 1, we observe that $S, T_1, T_2, \ldots, T_k$ are disjoint and satisfy

$$[n] = S \cup T = S \cup \left(\bigcup_{a\in R_2} T_a\right) \cup \left(\bigcup_{a\in R_3} T_a\right),$$

where the last equality holds since $a \in R_1$ implies $T_a = \emptyset$. To prove the claim, it suffices to show that there exists some $\boldsymbol{\Pi} \in \mathcal{S}_k$ such that $\mathcal{E}(\boldsymbol{\Pi}) \subseteq S \cup \left(\bigcup_{a\in R_3} T_a\right) = [n] \backslash \bigcup_{a\in R_2} T_a$. Indeed, for each $a \in R_2$, any pair $i, j \in T_a$ satisfies $\check{\boldsymbol{\Psi}}_{i\bullet} = \check{\boldsymbol{\Psi}}_{j\bullet}$ by definition. This fact, combined with Lemma 14, implies that there exists some $\boldsymbol{\Pi} \in \mathcal{S}_k$ such that $\left(\check{\boldsymbol{\Psi}}\boldsymbol{\Pi}\right)_{i\bullet} = \boldsymbol{\Psi}^*_{i\bullet}$ for all $i \in \bigcup_{a\in R_2} T_a$. By definition of $\mathcal{E}(\boldsymbol{\Pi})$, we have $\mathcal{E}(\boldsymbol{\Pi}) \cap \left(\bigcup_{a\in R_2} T_a\right) = \emptyset$ and are therefore done.

B. Proof of Claim 2

The claim follows by rearranging the left and right hand sides of the equation

$$|R_2| + 2\,|R_3| \le k = |R_1| + |R_2| + |R_3|,$$

which we now prove. The equality follows from the definition of R_1, R_2 and R_3. For the inequality, note that each element of R_2 contributes at least one distinct row in $\check{\boldsymbol{\Psi}}$ and each element of R_3 contributes at least two distinct rows in $\check{\boldsymbol{\Psi}}$. The indices of these rows are all in T by definition, and Lemma 14 guarantees that these rows are distinct across $R_2 \cup R_3$. The inequality then follows from the fact that $\check{\boldsymbol{\Psi}}$ has k distinct rows.

C. Proof of Claim 3

The first inequality in the claim holds because

$$|S| = \sum_{a\in[k]} |S_a| \ge \sum_{a\in R_1} |S_a| \overset{(i)}{=} \sum_{a\in R_1} \left|C_a^*\right| = |R_1|\,\ell,$$

where step (i) holds since $T_a = \emptyset$ for each $a \in R_1$ and thus $S_a = C_a^*$. On the other hand, we have

$$|S| \overset{(ii)}{\le} \sum_{i\in S} \frac{1}{\ell}\|\check{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 \overset{(iii)}{\le} \frac{1}{\ell}\|\check{\mathbf{Y}} - \mathbf{Y}^*\|_1,$$

where step (ii) holds since $\frac{1}{\ell}\|\check{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 \ge 1$ for each $i \in S$, and step (iii) holds since $S \subseteq [n]$. This proves the second inequality in the claim.

D. Proof of Claim 4

Recall that $\mathbf{Y}^* = \boldsymbol{\Psi}^*\left(\boldsymbol{\Psi}^*\right)^\top$ and $\boldsymbol{\Psi}^* \in \mathbb{M}_{n,k}$. Introducing an extra piece of notation $\mathbf{X}^* := \left(\boldsymbol{\Psi}^*\right)^\top$, we can write $\mathbf{Y}^* = \boldsymbol{\Psi}^*\mathbf{X}^*$. Let $\widetilde{\mathbf{X}} \in \mathbb{R}^{k\times n}$ be the matrix whose a-th row is equal to the element in $\left\{\widehat{\mathbf{Y}}_{i\bullet} : i \in C_a^*\right\}$ that is closest to $\mathbf{X}^*_{a\bullet}$ in ℓ_1 norm (with arbitrary tie-breaking); that is,

$$\widetilde{\mathbf{X}}_{a\bullet} := \operatorname*{arg\,min}_{\mathbf{x}\in\{\widehat{\mathbf{Y}}_{i\bullet} : i\in C_a^*\}} \|\mathbf{x} - \mathbf{X}^*_{a\bullet}\|_1 \quad \text{for each } a \in [k].$$

Finally let $\widetilde{\mathbf{Y}} := \mathbf{\Psi}^* \widetilde{\mathbf{X}}$. We have the inequality

$$\begin{aligned}
\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 &= \sum_{a\in[k]} \sum_{i\in C_a^*} \|\widehat{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 \\
&\overset{(i)}{=} \sum_{a\in[k]} \sum_{i\in C_a^*} \|\widehat{\mathbf{Y}}_{i\bullet} - \mathbf{X}^*_{a\bullet}\|_1 \\
&\geq \sum_{a\in[k]} \sum_{i\in C_a^*} \|\widetilde{\mathbf{X}}_{a\bullet} - \mathbf{X}^*_{a\bullet}\|_1 \\
&\overset{(ii)}{=} \sum_{a\in[k]} \sum_{i\in C_a^*} \|\widetilde{\mathbf{Y}}_{i\bullet} - \mathbf{Y}^*_{i\bullet}\|_1 \\
&= \|\widetilde{\mathbf{Y}} - \mathbf{Y}^*\|_1,
\end{aligned}$$

where step (i) holds since for each $a \in [k]$, the a-th row of $\mathbf{X}^*$ is the distinct row in $\mathbf{Y}^*$ corresponding to the cluster C_a^* and thus $\mathbf{X}^*_{a\bullet} = \mathbf{Y}^*_{i\bullet}$ for all $i \in C_a^*$; step (ii) can be justified by applying the same argument to $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. It follows that

$$\|\widetilde{\mathbf{Y}} - \widehat{\mathbf{Y}}\|_1 \leq \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 + \|\widetilde{\mathbf{Y}} - \mathbf{Y}^*\|_1 \leq 2\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$$

and hence

$$\begin{aligned}
\|\check{\mathbf{Y}} - \widehat{\mathbf{Y}}\|_1 &\overset{(i)}{\leq} \rho\|\overline{\mathbf{Y}} - \widehat{\mathbf{Y}}\|_1 \\
&\overset{(ii)}{\leq} \rho\|\widetilde{\mathbf{Y}} - \widehat{\mathbf{Y}}\|_1 \\
&\leq 2\rho\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1,
\end{aligned}$$

where step (i) holds by the approximation ratio guarantee (26), and step (ii) holds since $(\overline{\mathbf{\Psi}}, \overline{\mathbf{X}})$ is optimal for the k-medians problem (25) (recall that $\overline{\mathbf{Y}} := \overline{\mathbf{\Psi}}\overline{\mathbf{X}}$) while $(\mathbf{\Psi}^*, \widetilde{\mathbf{X}})$ is feasible for the same problem (because $\mathbf{\Psi}^* \in \mathbb{M}_{n,k}$ and the rows of $\widetilde{\mathbf{X}}$ are selected from $\widehat{\mathbf{Y}}$ by construction). Combining pieces, we obtain that

$$\begin{aligned}
\|\check{\mathbf{Y}} - \mathbf{Y}^*\|_1 &\leq \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 + \|\check{\mathbf{Y}} - \widehat{\mathbf{Y}}\|_1 \\
&\leq (1 + 2\rho)\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1.
\end{aligned}$$

REFERENCES

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Soc. Netw.*, vol. 5, pp. 109–137, Jun. 1983.

[2] E. Abbe, "Community detection and stochastic block model: Recent developments," *J. Mach. Learn. Res.*, 2017. [Online]. Available: http://www.princeton.edu/%7Eeabbe/publications/sbm_jmlr_4.pdf

[3] C. Moore, "The computer science and physics of community detection: Landscapes, phase transitions, and hardness," *Bull. Eur. Assoc. Theor. Comput. Sci.*, no. 121, Feb. 2017.

[4] O. Guédon and R. Vershynin, "Community detection in sparse networks via Grothendieck's inequality," *Probab. Theory Rel. Fields*, vol. 165, nos. 3–4, pp. 1025–1049, 2016.

[5] B. Bollobás and A. D. Scott, "Max cut for random graphs with a planted partition," *Combinat., Probab. Comput.*, vol. 13, nos. 4–5, pp. 451–474, 2004.

[6] E. Abbe and C. Sandon, "Recovering communities in the general stochastic block model without knowing the parameters," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 676–684.

[7] Y. Chen and J. Xu, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 882–938, 2016.

[8] E. Mossel, J. Neeman, and A. Sly. (2012). "Stochastic block models and reconstruction." [Online]. Available: https://arxiv.org/abs/1202.1499

[9] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 84, p. 066106, Dec. 2011, doi: 10.1103/PhysRevE.84.066106.

[10] A. A. Amini and E. Levina. (2014). "On semidefinite relaxations for the block model." [Online]. Available: https://arxiv.org/abs/1406.5647

[11] B. P. W. Ames and S. A. Vavasis, "Convex optimization for the planted k-disjoint-clique problem," *Math. Program.*, vol. 143, nos. 1–2, pp. 299–337, 2014.

[12] Y. Chen, S. Sanghavi, and H. Xu, "Clustering sparse graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2204–2212.

[13] S. Oymak and B. Hassibi. (2011). "Finding dense clusters via 'low rank + sparse' decomposition." [Online]. Available: https://arxiv.org/abs/1104.5186

[14] M. Krivelevich and D. Vilenchik, "Semirandom models as benchmarks for coloring algorithms," in *Proc. 3rd Workshop Anal. Algorithmics Combinat. (ANALCO)*, 2006, pp. 211–221.

[15] Y. Chen, X. Li, and J. Xu. (Dec. 2015). "Convexified modularity maximization for degree-corrected stochastic block models." [Online]. Available: https://arxiv.org/abs/1512.08425

[16] Y. Chen, S. Sanghavi, and H. Xu, "Improved graph clustering," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6440–6455, Oct. 2014.

[17] A. Montanari and S. Sen, "Semidefinite programs on sparse random graphs and their application to community detection," in *Proc. 48th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, Cambridge, MA, USA, Jun. 2016, pp. 814–827, doi: 10.1145/2897518.2897548.

[18] L. Massoulié, "Community detection thresholds and the weak Ramanujan property," in *Proc. 46th Annu. ACM Symp. Theory Comput.*, 2014, pp. 694–703.

[19] E. Abbe and C. Sandon. (2015). "Detection in the stochastic block model with multiple clusters: Proof of the achievability conjectures, acyclic BP, and the information-computation gap." [Online]. Available: https://arxiv.org/abs/1512.09080

[20] E. Mossel, J. Neeman, and A. Sly. (2013). "A proof of the block model threshold conjecture." [Online]. Available: https://arxiv.org/abs/1311.4115

[21] E. Mossel, J. Neeman, and A. Sly, "Reconstruction and estimation in the planted partition model," *Probab. Theory Rel. Fields*, vol. 162, nos. 3–4, pp. 431–461, 2015.

[22] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for the planted bisection model," *Electron. J. Probab.*, vol. 21, no. 21, pp. 1–24, 2016, doi: 10.1214/16-EJP4185.

[23] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *Proc. IEEE 56th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2015, pp. 670–688.

[24] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. (2015). "Achieving optimal misclassification proportion in stochastic block model." [Online]. Available: https://arxiv.org/abs/1505.03772

[25] S.-Y. Yun and A. Proutiere. (2014). "Accurate community detection in the stochastic block model via spectral algorithms." [Online]. Available: https://arxiv.org/abs/1412.7335

[26] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 471–487, Jan. 2016.

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

[28] W. Perry and A. S. Wein. (2015). "A semidefinite program for unbalanced multisection in the stochastic block model." [Online]. Available: https://arxiv.org/abs/1507.05605

[29] A. S. Bandeira, "Random Laplacian matrices and convex relaxations," *Found. Comput. Math.*, vol. 18, pp. 345–379, Apr. 2018.

[30] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. (2015). "Multisection in the stochastic block model using semidefinite programming." [Online]. Available: https://arxiv.org/abs/1507.02323

[31] P. Chin, A. Rao, and V. Vu, "Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery," in *Proc. 28th Conf. Learn. Theory (COLT)*, Paris, France, Jul. 2015, pp. 391–423. [Online]. Available: http://jmlr.org/proceedings/papers/v40/Chin15.html

[32] S.-Y. Yun and A. Proutiere, "Optimal cluster recovery in the labeled stochastic block model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 965–973.

[33] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan, "Learning communities in the presence of errors," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 1258–1291.

[34] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery," *IEEE Trans. Netw. Sci. Eng.*, vol. 1, no. 1, pp. 10–22, Jan. 2014.
[35] B. Hajek, Y. Wu, and J. Xu, "Exact recovery threshold in the binary censored block model," in *Proc. IEEE Inf. Theory Workshop-Fall (ITW)*, Oct. 2015, pp. 99–103.
[36] S. Heimlicher, M. Lelarge, and L. Massoulié. (2012). "Community detection in the labelled stochastic block model." [Online]. Available: https://arxiv.org/abs/1209.2910
[37] M. Lelarge, L. Massoulié, and J. Xu, "Reconstruction in the labelled stochastic block model," *IEEE Trans. Netw. Sci. Eng.*, vol. 2, no. 4, pp. 152–163, Oct./Dec. 2015.
[38] A. Saade, M. Lelarge, F. Krzakala, and L. Zdeborová, "Spectral detection in the censored block model," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1184–1188.
[39] T. T. Cai and X. Li, "Robust and computationally feasible community detection in the presence of arbitrary outlier nodes," *Ann. Statist.*, vol. 43, no. 3, pp. 1027–1059, 2015. [Online]. Available: http://arxiv.org/abs/1404.6000
[40] A. Frieze and C. McDiarmid, "Algorithmic theory of random graphs," *Random Struct. Algorithms*, vol. 10, nos. 1–2, pp. 5–42, 1997.
[41] U. Feige and J. Kilian, "Heuristics for semirandom graph problems," *J. Comput. Syst. Sci.*, vol. 63, no. 4, pp. 639–671, 2001.
[42] F. Ricci-Tersenghi, A. Javanmard, and A. Montanari, "Performance of a community detection algorithm based on semidefinite programming," *J. Phys., Conf. Ser.*, vol. 699, no. 1, p. 012015, 2016. [Online]. Available: http://stacks.iop.org/1742-6596/699/i=1/a=012015
[43] A. Coja-Oghlan, "Coloring semirandom graphs optimally," in *Automata, Languages and Programming*, 2004, pp. 383–395.
[44] A. Moitra, W. Perry, and A. S. Wein, "How robust are reconstruction thresholds for community detection?" in *Proc. 48th Annu. ACM SIGACT Symp. Theory Comput.*, 2016, pp. 828–841.
[45] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, "Phase transitions in semidefinite relaxations," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 16, pp. E2218–E2223, 2016.
[46] Y. Lu and H. H. Zhou. (2016). "Statistical and computational guarantees of Lloyd's algorithm and its variants." [Online]. Available: https://arxiv.org/abs/1612.02099
[47] M. Charikar, S. Guha, É. Tardos, and D. B. Shmoys, "A constant-factor approximation algorithm for the *k*-median problem," in *Proc. 31st Annu. ACM Symp. Theory Comput.*, 1999, pp. 1–10.
[48] A. Y. Zhang and H. H. Zhou, "Minimax rates of community detection in stochastic block models," *Ann. Statist.*, vol. 44, no. 5, pp. 2252–2280, 2016.
[49] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. (2016). "Community detection in degree-corrected block models." [Online]. Available: https://arxiv.org/abs/1607.06993
[50] R. H. Keshavan, S. Oh, and A. Montanari, "Matrix completion from a few entries," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2009, pp. 324–328.
[51] B. P. W. Ames, "Guaranteed clustering and biclustering via semidefinite programming," *Math. Program.*, vol. 147, no. 1, pp. 429–465, 2013.
[52] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 210–268.
[53] P. Rigollet, *18.S997: High Dimensional Statistics* (Lecture Notes). Cambridge, MA, USA: MIT OpenCourseWare, 2015.
[54] A. S. Bandeira and R. van Handel, "Sharp nonasymptotic bounds on the norm of random matrices with independent entries," *Ann. Probab.*, vol. 44, no. 4, pp. 2479–2506, 2016, doi: 10.1214/15-AOP1025.
[55] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.

Yingjie Fei is a Ph.D. candidate at the School of Operations Research and Information Engineering at Cornell University. He obtained his B.A. in Mathematics from New York University in 2015. His research interests include machine learning and statistical learning.

Yudong Chen is an Assistant Professor with the School of Operations Research and Information Engineering at Cornell University. He obtained his Ph.D. degree in Electrical and Computer Engineering in 2013 from The University of Texas at Austin, and M.S. and B.S. degrees in Control Science and Engineering from Tsinghua University. He was a postdoc in the Electrical Engineering and Computer Sciences department at the University of California, Berkeley. He has served on the senior program committees of AAAI and AISTATS. His research work lies in machine learning, high-dimensional statistics, and optimization, with applications in network scheduling, wireless communication, and financial systems.