Journal of Machine Learning Research 17 (2016) 1-57 Submitted 8/14; Revised 5/15; Published 4/16

Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices*

Yudong Chen YUDONG.CHEN@CORNELL.EDU
Department of Operations Research and Information Engineering
Cornell University
Ithaca, NY 14853, USA

Jiaming Xu JIAMINGX@WHARTON.UPENN.EDU
Department of Statistics, The Wharton School
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Gabor Lugosi

Abstract

We consider two closely related problems: planted clustering and submatrix localization. In the planted clustering problem, a random graph is generated based on an underlying cluster structure of the nodes; the task is to recover these clusters given the graph. The submatrix localization problem concerns locating hidden submatrices with elevated means inside a large real-valued random matrix. Of particular interest is the setting where the number of clusters/submatrices is allowed to grow unbounded with the problem size. These formulations cover several classical models such as planted clique, planted densest subgraph, planted partition, planted coloring, and the stochastic block model, which are widely used for studying community detection, graph clustering and bi-clustering.

For both problems, we show that the space of the model parameters (cluster/submatrix size, edge probabilities and the mean of the submatrices) can be partitioned into four disjoint regions corresponding to decreasing statistical and computational complexities: (1) the *impossible* regime, where all algorithms fail; (2) the *hard* regime, where the computationally expensive Maximum Likelihood Estimator (MLE) succeeds; (3) the *easy* regime, where the polynomial-time convexified MLE succeeds; (4) the *simple* regime, where a local counting/thresholding procedure succeeds. Moreover, we show that each of these algorithms provably fails in the harder regimes.

Our results establish the minimax recovery limits, which are tight up to universal constants and hold even with a growing number of clusters/submatrices, and provide order-wise stronger performance guarantees for polynomial-time algorithms than previously known. Our study demonstrates the tradeoffs between statistical and computational considerations, and suggests that the minimax limits may not be achievable by polynomial-time algorithms.

Keywords: planted partition, planted clique, planted coloring, submatrix localization, graph clustering, bi-clustering, minimax recovery, computational hardness, convex relaxation

∗. Partial preliminary results are presented in the conference paper Chen and Xu (2014).

1. Introduction

In this paper we consider two closely related problems: planted clustering and submatrix localization, both concerning the recovery of hidden structures from a noisy random graph or matrix.

- *Planted Clustering:* Suppose that out of a total of n nodes, rK of them are partitioned into r clusters of size K, and the remaining $n - rK$ nodes do not belong to any clusters; each pair of nodes is connected by an edge with probability p if they are in the same cluster, and with probability q otherwise. Given the adjacency matrix A of the graph, the goal is to recover the underlying clusters (up to a permutation of cluster indices). By varying the values of the model parameters, this formulation covers several classical models including planted clique, planted coloring, planted densest subgraph, planted partition, and stochastic block model (cf. Definition 1 and discussion thereafter).

- *Submatrix Localization:* Suppose $A \in \mathbb{R}^{n_L \times n_R}$ is a random matrix with independent Gaussian entries with unit variance, where there are r submatrices of size $K_L \times K_R$ with disjoint row and column supports, such that the entries inside these submatrices have mean $\mu > 0$, and the entries outside have mean zero. The goal is to identify the locations of these hidden submatrices given A. This formulation generalizes the submatrix detection and bi-clustering models with a single bi-submatrix/cluster that are studied in previous work (cf. Definition 14 and discussion thereafter).

We are particularly interested in the setting where the number r of clusters or submatrices may grow unbounded with the problem dimensions n, n_L, and n_R at an arbitrary rate. We call this the *high-rank* setting because r equals the rank of a matrix representation of the clusters and submatrices (cf. Definitions 1 and 14). The other parameters K, p, q, and μ are also allowed to scale with n or (n_L, n_R).

These two problems have been studied extensively, under various names such as *community detection*, *graph clustering/bi-clustering*, and *reconstruction in stochastic block models*, and have a broad range of applications. They are used as generative models for approximating real-world networks and data arrays with natural cluster/community structures, such as social networks (Fortunato, 2010), gene expressions (Shabalin et al., 2009), and online ratings (Xu et al., 2014). They serve as benchmarks in the evaluation of algorithms for clustering (Mathieu and Schudy, 2010), bi-clustering (Balakrishnan et al., 2011a), community detection (Newman and Girvan, 2004), and other network inference problems. They also provide a venue for studying the average-case behaviors of many graph theoretic problems including max-clique, max-cut, graph partitioning, and coloring (Bollobás and Scott, 2004; Condon and Karp, 2001). The importance of these two problems are well-recognized in areas across computer science, statistics, and physics (Rohe et al., 2011; Arias-Castro and Verzelen, 2014; Nadakuditi and Newman, 2012; Decelle et al., 2011; Mossel et al., 2013; Lelarge et al., 2013; Anandkumar et al., 2014; Bickel and Chen, 2009; Amini et al., 2013).

The planted clustering and submatrix localization problems exhibit an interplay between *statistical* and *computational* considerations. From a statistical point of view, we are interested in identifying the range of the model parameters for which the hidden structures—in this case the clusters and submatrices—can be recovered from the noisy data A. The values of the parameters n, r, K, p, q, μ govern the statistical hardness of the problems: the problems become more difficult with smaller values of $p-q$, μ, K, and larger r, because the observations are noisier and the sought-

after structures are more complicated. A statistically powerful algorithm is one that can recover the hidden structures for a large region of the model parameter space.

From a computational point of view, we are concerned with the running time of different recovery algorithms. An exhaustive search over the solution space (i.e., all possible clusterings or submatrix locations) may make for a statistically powerful algorithm, but is computationally intractable. A simpler algorithm with lower running time is computationally more desirable, but may succeed only in a smaller region of the model parameter space and thus has weaker statistical power.

Therefore, it is important to take a joint statistical-computational view to the planted clustering and submatrix localization problems, and to understand the *tradeoffs* between these two considerations. How do algorithms with different computational complexity achieve different statistical performance? For these two problems, what are the *information limit* (under what conditions on the model parameters does recovery become infeasible for any algorithm) and the *computational limit* (when does it become infeasible for computationally tractable algorithms)?

We take a step in answering questions in this paper. For both problems, our results demonstrate, in a precise quantitative way, the following phenomenon: the parameter space can be partitioned into four disjoint regions, such that each region corresponds to statistically easier instances of the problem than the previous one, and recovery can be achieved by simpler algorithms with lower running time. Significantly, there might exist a large gap between the statistical performance of computationally intractable algorithms and that of computationally efficient algorithms. We elaborate in the next two subsections.

1.1 Planted Clustering: The Four Regimes

For concreteness, we first consider the planted clustering problem in the setting $r \geq 2$, $p > q$ and $p/q = \Theta(1)$. This covers the standard planted bisection/partition/r-disjoint-clique models.

The statistical hardness of cluster recovery is captured by the quantity $\frac{(p-q)^2}{q(1-q)}$, which is essentially a measure of the Signal-to-Noise Ratio (SNR). Our main theorems identify the following four regimes of the problem defined by the value of this quantity. (Here for simplicity, we use the notation $\dot{\gtrsim}$ and $\dot{\lesssim}$, which ignore constant and $\log n$ factors; note that our main theorems do sharply characterize the $\log n$ factors.)

- *The Impossible Regime:* $\frac{(p-q)^2}{q(1-q)} \dot{\lesssim} \frac{1}{K}$. In this regime, there is no algorithm, regardless of its computational complexity, that can recover the clusters with a vanishing probability of error.

- *The Hard Regime:* $\frac{1}{K} \dot{\lesssim} \frac{(p-q)^2}{q(1-q)} \dot{\lesssim} \frac{n}{K^2}$. There exists a computationally expensive algorithm—specifically the Maximum Likelihood Estimator (MLE)—that recovers the clusters with high probability in this regime (as well as in the next two easier regimes; we omit such implications in the sequel). There is no known polynomial-time algorithm that succeeds in this regime.

- *The Easy Regime:* $\frac{n}{K^2} \dot{\lesssim} \frac{(p-q)^2}{q(1-q)} \dot{\lesssim} \frac{\sqrt{n}}{K}$. There exists a polynomial-time algorithm—specifically a convex relaxation of the MLE—that recovers the clusters with high probability in this regime. Moreover, this algorithm provably fails in the hard regime above.

- *The Simple Regime:* $\frac{(p-q)^2}{q(1-q)} \dot{\gtrsim} \frac{\sqrt{n}}{K}$. A simple algorithm based on counting node degrees and common neighbors recovers the clusters with high probability in this regime, and provably fails outside this regime (i.e., in the hard and easy regimes).

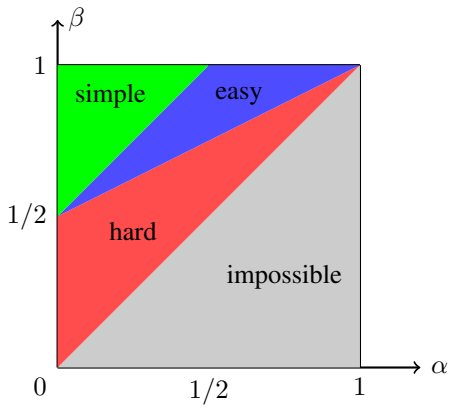

Figure 1: Illustration of the four regimes. The figure applies to the planted clustering problem with $p = 2q = \Theta(n^{-\alpha})$ and $K = \Theta(n^{\beta})$, as well as to the submatrix localization problem with $n_L = n_R = n$, $\mu^2 = \Theta(n^{-\alpha})$ and $K_L = K_R = \Theta(n^{\beta})$.

We illustrate these four regimes in Figure 1 assuming the scaling $p = 2q = \Theta(n^{-\alpha})$ and $K = \Theta(n^{\beta})$ for two constants $\alpha, \beta \in (0, 1)$. Here cluster recovery becomes harder with larger α and smaller β. In this setting, the four regimes correspond to four disjoint and non-empty regions of the parameter space. Therefore, a computationally more expensive algorithm leads to a *significant* (polynomial in n) enhancement in the statistical power. For example, when $\alpha = 1/4$, the simple, polynomial-time, and computationally intractable algorithms succeeds for β larger than 0.75, 0.625, and 0.25, respectively. There is a similar hierarchy for the allowable sparsity of the graph, given by $\alpha < 0.25$, $\alpha < 0.5$, and $\alpha < 0.75$ assuming $\beta = 0.75$.

The results in the impossible and hard regimes together establish the *minimax recovery boundary* of the planted clustering problem, and show that the MLE is statistically order-wise optimal. These two regimes are separated by an "information barrier": in the impossible regime the graph does not carry enough information to distinguish different cluster structures, so recovery is statistically impossible.

Our performance guarantees for the convexified MLE improve upon existing results for polynomial time algorithms in terms of scaling with n, particularly in the setting when the number of clusters are allowed to grow with n.

We *conjecture* that no polynomial-time algorithm can perform significantly better and succeed in the hard regime, i.e., the convexified MLE achieves the *computational limit* order-wise. While we do not prove the conjecture, there are many supporting evidences; cf. Section 2.3. For instance, there is a "spectral barrier", determined by the spectrum of an appropriately defined noise matrix, that prevents the convexified MLE and spectral clustering algorithms from succeeding in the hard regime. In the special setting with a single cluster, the work by Ma and Wu (2015); Hajek et al. (June, 2014) proves that no polynomial-time algorithm can reliably recover the cluster if $\beta < \alpha/4 + 1/2$ conditioned on the planted clique hardness hypothesis.

The simple counting algorithm is an example of an algorithm that uses only "local information": the cluster relation between a pair of nodes is inferred from only their two-hop connection. In

contrast, the convexified MLE crucially uses the global information in the graph spectrum. The counting algorithm fails outside the simple regime due to a "variance barrier" associated with the fluctuations in the node degrees and the numbers of common neighbors, and is statistically order-wise weaker than the global approach of the convexified MLE.

Our main theorems apply beyond the above specific setting and allow for general scalings of p, q, K, and r. The four regimes and the statistical-computational tradeoffs are observed for a broad spectrum of planted problems, including planed partition, planted coloring, planted r-disjoint-clique, and planted densest-subgraph models. Table 1 summarizes the implications of our results for some of these models. More precise and general results are given in Section 2.

1.2 Submatrix Localization: The Four Regimes

Similar results hold for the submatrix localization problem. Consider the setting with $n_L = n_R = n$ and $K_L = K_R = K$. The statistical hardness of submatrix localization is captured by the quantity μ^2, which is again a measure of the SNR. In the high SNR setting with $\mu^2 = \Omega(\log n)$, the submatrices can be trivially identified by element-wise thresholding. In the more interesting low SNR setting with $\mu^2 = O(\log n)$, our main theorems identify the following four regimes, which have the same meanings as before:

- *The Impossible Regime:* $\mu^2 \dot{\lesssim} \frac{1}{K}$. All algorithms fail in this regime.
- *The Hard Regime:* $\frac{1}{K} \dot{\lesssim} \mu^2 \dot{\lesssim} \frac{n}{K^2}$. The computationally expensive MLE succeeds, and it is conjectured that no polynomial-time algorithm succeeds here.
- *The Easy Regime:* $\frac{n}{K^2} \dot{\lesssim} \mu^2 \dot{\lesssim} \frac{\sqrt{n}}{K}$. The polynomial-time convexified MLE succeeds, and provably fails in the hard regime.
- *The Simple Regime:* $\frac{\sqrt{n}}{K} \dot{\lesssim} \mu^2 \dot{\lesssim} 1$. A simple thresholding algorithm succeeds, and provably fails outside this regime.

We illustrate these four regimes in Figure 1 assuming $\mu^2 = \Theta(n^{-\alpha})$ and $K = \Theta(n^{\beta})$. In fact, the results above hold in the more general setting where the entries of A are *sub-Gaussian*.

1.3 Discussions

This paper presents a systematic study of planted clustering and submatrix localization with a growing number of clusters/submatrices. We provide sharp characterizations of the minimax recovery boundary with the lower and upper bounds matching up to constants. We also give improved performance guarantees for convex optimization approaches and the simple counting/thresholding algorithms. In addition, complementary results are given for the *failure conditions* for these algorithms, hence characterizing their performance limits. Our analysis addresses several technical challenges that arise in the high-rank setting. The results in this paper highlight the similarity between planted clustering and submatrix localization, and place several classical problems under a unified framework including planted clique, planted partition, planted coloring, and planted densest subgraph.

The central theme of our investigation is the interaction between the statistical and the computational aspects in the problems, i.e., how to handle more noise and more complicated structures using more computation. Our study parallels a recent line of work that takes a joint statistical and

	Planted r-Disjoint-Clique $1 = p > q \geq 0, r \geq 1$	**Planted Partition/ Stochastic Blockmodel** $1 \geq p > q \geq 0, rK = n$	**Planted Coloring** $0 = p < q \leq 1, rK = n$
Impossible Thm 2, Cor 3	$K \lesssim \left(\frac{q}{1-q} \vee \frac{1}{\log(1/q)}\right) \log n$	$(p-q)^2 \lesssim \frac{p(1-q)\log n}{K}$	$q \lesssim \frac{\log n}{K}$
MLE Thm 4, Cor 5	$K \gtrsim \left(\frac{q}{1-q} \vee \frac{1}{\log(1/q)}\right) \log n$	$(p-q)^2 \gtrsim \frac{p(1-q)\log n}{K}$	$q \gtrsim \frac{\log n}{K}$
Convexified MLE Thm 6	$K \gtrsim \frac{\log n}{1-q} + \sqrt{\frac{qn}{1-q}}$	$(p-q)^2 \gtrsim \frac{p(1-q)\log n}{K} + \frac{q(1-q)n}{K^2}$	$q \gtrsim \frac{\log n}{K} + \frac{(1-q)n}{K^2}$
Simple Counting Thm 10, Rem 11	$K \gtrsim \frac{\log n}{1-q} + \sqrt{\frac{qn\log n}{1-q}}$	$(p-q)^4 \gtrsim \left[\frac{p^2(1-q)}{K} + \frac{nq(1-q)(q \vee p^2)}{K^2}\right] \log n$	$q^2 \gtrsim \frac{(1-q)n\log n}{K^2}$

Table 1: Our results specialized to different planted models. Here the notations $\gtrsim$ and $\lesssim$ ignore constant factors. This table shows the *necessary conditions* for any algorithm to succeed under a mild assumption $K \gtrsim \log(rK)$, as well as the *sufficient* conditions under which the algorithms in this paper succeed, thus corresponding to the four regimes described in Section 1.1. The relevant theorems/corollaries are also listed. The conditions for convexified MLE and simple counting can further be shown to be also *necessary* in a broad range of settings; cf. Theorems 8 and 12. The results in this table are not the strongest possible; see the referenced theorems for more precise statements.

computational view on inference problems (Balakrishnan et al., 2011a; Oymak et al., 2015; Berthet and Rigollet, 2013; Chandrasekaran and Jordan, 2013; Ma and Wu, 2015); several of these works are closely related to special cases of the planted clustering and bi-clustering models. In this sense, we investigate an emblematic and fundamental problem, and therefore expect that the phenomena and principles described in this paper are relevant more generally. Below we provide additional discussion, and comment on connections with the existing work.

1.3.1 HIGH RANK VERSUS RANK ONE

Several recent works investigate the problems of single-submatrix detection/localization (Kolar et al., 2011; Arias-Castro et al., 2011), planted densest subgraph detection (Arias-Castro and Verzelen, 2014), and sparse principal component analysis (PCA) (Amini and Wainwright, 2009) (cf. Section 1.4 for a literature review). Even earlier is the extensive study of the statistical/computational hardness of Planted Clique. The majority of these results focus on the *rank-one* setting with a single clique, cluster, submatrix or principal component. This paper considers the more general *high-rank* setting, where the number r of clusters/submatrices may grow rapidly with the problem size. This setting is important in many real-world networks (see e.g., Leskovec et al., 2008; Rohe et al., 2011), and poses significant challenges to the analysis. Moreover, there are qualitative differences between these two settings. We discuss one such difference in the next paragraph.

1.3.2 THE POWER OF CONVEX RELAXATION

In previous work on the rank-one case of submatrix detection/localization (Ma and Wu, 2015; Balakrishnan et al., 2011a) and sparse PCA (Krauthgamer et al., 2015), it is shown that simple algorithms based on averaging/thresholding have order-wise similar statistical performance as more sophisticated convex relaxation approaches. In contrast, for the problems of finding multiple clusters/submatrices, we show that convex relaxation of MLE is statistically much more powerful than the simple counting/thresholding algorithm. Our analysis reveals that the power of convex relaxation lies in *separating different clusters/submatrices*, but not in identifying a single cluster/submatrix. Our results thus provide one explanation for the (somewhat curious) observation in previous work regarding the lack of benefit of using sophisticated methods, and demonstrate a finer spectrum of computational-statistical tradeoffs.

1.3.3 DETECTION VERSUS ESTIMATION

Several recent works on planted densest subgraph and submatrix detection have focused on the *detection* or *hypothesis testing* version of the problems, i.e., detecting the existence of a dense cluster or an elevated submatrix (cf. Section 1.4 for literature review). In this paper, we study the (support) *estimation* version of the problems, where the goal is to find the precise locations of the clusters/submatrices. In general estimation appears to be harder than detection. For example, if we consider the scalings of μ and K in Figure 1 of this paper, and compare with Figure 1 in Ma and Wu (2015) which studies submatrix detection, we see that the minimax localization boundary is $\beta = \alpha$, whereas the minimax detection boundary is at a lower value $\beta = \min\{\alpha, \alpha/4 + 1/2\}$. For the planted densest subgraph problem, we see a similar gap between the minimax detection and estimation boundaries if we compare our results with results in Arias-Castro and Verzelen (2014); Hajek et al. (June, 2014). In addition, it is shown in (Ma and Wu, 2015; Hajek et al., June, 2014) that if $\beta > \alpha/4 + 1/2$, the planted submatrix or densest subgraph can be detected in linear time; if

$\beta < \alpha/4+1/2$, no polynomial-time test exists assuming the hardness of the planted clique detection problem. For estimation, we prove the sufficient condition $\beta > \alpha/2+1/2$, which is the best known performance guarantee for polynomial-time algorithms—again we see a gap between detection and estimation. For detecting a sparse principal component, see the seminar work Berthet and Rigollet (2013) for proving computational lower bounds conditioned on the hardness of Planted Clique.

1.3.4 EXTENSIONS

It is a simple exercise to extend our results to a variant of the planted clustering model where the graph adjacency matrix has sub-Gaussian entries instead of Bernoulli, corresponding to a weighted graph clustering problem. Similarly, we can also extend the submatrix location problem to the setting with Bernoulli entries, which is the bi-clustering problem on an unweighted graph and covers the *planted bi-clique* problem (Feldman et al., 2013; Ames and Vavasis, 2011) as a special case.

1.4 Related Work

There is a large body of literature, from the physics, computer science and statistics communities, on models and algorithms for graph clustering and bi-clustering, as well as on their various extensions and applications. A complete survey is beyond the scope of this paper. Here we focus on theoretical work on planted clustering/submatrix localization concerning *exact recovery* of the clusters/submatrices. Detailed comparisons of existing results with ours are provided after we present each of our theorems in Sections 2 and 3. We emphasize that our results are *non-asymptotic* for finite values of n, n_L and n_R, whereas some of the results below require $n \to \infty$.

1.4.1 PLANTED CLIQUE, PLANTED DENSEST SUBGRAPH

The planted clique model ($r = 1$, $p = 1$, $q = 1/2$) is the most widely studied planted model. If the clique has size $K = o(\log n)$, recovery is impossible as the random graph $\mathcal{G}(n, 1/2)$ will have a clique with at least the same size; if $K = \Omega(\log n)$, an exhaustive search succeeds (Alon et al., 1998); if $K = \Omega(\sqrt{n})$, various polynomial-time algorithms work (Alon et al., 1998; Dekel et al., 2014; Deshpande and Montanari, 2013); if $K = \Omega(\sqrt{n \log n})$, the nodes in the clique can be easily identified by counting degrees (Kučera, 1995). It is an open problem to find polynomial-time algorithms which succeed in the regime with $K = o(\sqrt{n})$, and it is believed that this cannot be done (Hazan and Krauthgamer, 2011; Juels and Peinado, 2000; Alon et al., 2007; Feldman et al., 2013). The four regimes above can be considered as a special case of our results for the general planted clustering model. The planted densest subgraph model generalizes the planted clique model by allowing general values of p and q. The detection version of this problem is studied in Arias-Castro and Verzelen (2014); Verzelen and Arias-Castro (2013), and conditional computational hardness results are obtained in Hajek et al. (June, 2014).

1.4.2 PLANTED r-DISJOINT-CLIQUES, PARTITION, AND COLORING

Subsequent work considers the setting with $r \geq 1$ planted cliques (McSherry, 2001), as well as the planted partition model (a.k.a. stochastic block model) with general values of $p > q$ (Condon and Karp, 2001; Holland et al., 1983). A subset of these results allow for a growing number r of clusters (e.g., Rohe et al., 2011). Most existing work focuses on the recovery performance of polynomial-time algorithms. The state-of-the-art for planted r-disjoint-clique are given in McSherry

(2001); Chen et al. (2012); Ames and Vavasis (2014), and for planted partition in Chen et al. (2012); Anandkumar et al. (2014); Cai and Li (2015); see Chen et al. (2014b) for a survey. The setting with $p < q$ is sometimes called the *heterophily* case, with the planted coloring model ($p = 0$) as an important special case (Alon and Kahale, 1997; Coja-Oghlan, 2004). Our results on the convexified MLE (cf. Table 1) improve upon the previously known statistical performance of polynomial-time algorithms in the general r setting. The information-theoretic limits (both lower and upper bounds) of cluster recovery were largely unknown when r is growing. Here we identify these limits up to constant factors for general values of p, q, K and r. In particular, our results show that the information limit is achievable by MLE order-wise, while there is a significant gap between the information limit and the performance guarantee of the convexified MLE.

1.4.3 CONVERSE RESULTS FOR PLANTED PROBLEMS

Complementary to the *achievability* results, another line of work focuses on *converse* results, i.e., identifying necessary conditions for recovery, either for any algorithm, or for any algorithm in a specific class. For the planted partition model with $K = \Theta(n)$, necessary conditions for any algorithm to succeed are obtained by information-theoretic arguments in Chaudhuri et al. (2012); Chen et al. (2012); Balakrishnan et al. (2011b); Abbe et al. (2014). For spectral clustering algorithms and convex optimization approaches, more stringent conditions are shown to be needed (Nadakuditi and Newman, 2012; Vinayak et al., 2014). We generalize and improve upon these existing results in the setting with general r and K.

1.4.4 SHARP THRESHOLDS WITH A BOUNDED NUMBER OF CLUSTERS

Since the conference version of this paper was published (Chen and Xu, 2014), several papers have appeared on the asymptotic information-theoretic limits for exact recovery of planted partition. In the restricted setting with $r = 2$ and $K = n/2$, the recovery thresholds *with sharp constants* are identified and shown to be achievable in polynomial-time in Abbe et al. (2014) for $p, q = O(\log n/n)$, and in Mossel et al. (2015b) for more general scalings of (p, q). Very recently, Abbe and Sandon (2015) established the sharp recovery threshold for the case where $r = O(1)$, $K = \Theta(n)$, and the in-cluster and cross-cluster edge probabilities are heterogeneous and scale as $\log n/n$; the recovery threshold is further shown to be achievable in $o(n^{1+\epsilon})$ time for any $\epsilon > 0$. In this bounded r setting, it is further shown in Hajek et al. (2014); Bandeira (2015); Hajek et al. (2015) that the sharp recovery thresholds are achieved by the semidefinite programming relaxation of the MLE. In comparison, the results in this paper are non-asymptotic and optimal up to universal constant factors, and apply to the general setting with a growing number of clusters/submatrices of size sublinear in n.

1.4.5 APPROXIMATE RECOVERY

While not the focus of this paper, approximate cluster recovery (under various criteria) has also been studied, e.g., for planted partition with $r = O(1)$ clusters in Mossel et al. (2015a, 2013); Massoulié (2014); Yun and Proutiere (2014); Decelle et al. (2011). These results are not directly comparable to ours, but often the approximate recovery conditions differ from the ones for exact recovery by a $\log n$ factor. When constant factors are concerned, the existence of a hard regime was conjectured in Decelle et al. (2011); Mossel et al. (2015a).

1.4.6 SUBMATRIX LOCALIZATION

The statistical-computational tradeoffs in locating a single submatrix (i.e., $r = 1$) are discussed in Balakrishnan et al. (2011a); Kolar et al. (2011); the information limit is shown to be achieved (order-wise) by a computationally intractable algorithm, and the success and failure conditions for various polynomial-time procedures are also derived. The work in Ames (2013) focuses on success conditions for a convex relaxation approach; we improve on their results particularly in the high-rank setting. The single-submatrix *detection* problem is studied in Butucea and Ingster (2013); Shabalin et al. (2009); Sun and Nobel (2013); Arias-Castro et al. (2011); Bhamidi et al. (2012), and conditional hardness results are established in the recent work in Ma and Wu (2015).

1.5 Paper Organization and Notation

The remainder of this paper is organized as follows. In Section 2 we set up the planted clustering model and present our main theorems for the impossible, hard, easy, and simple regimes. In Section 3 we turn to the submatrix localization problem and provide the corresponding theorems for the four regimes. Section 4 provides a brief summary with a discussion of future directions. We prove the main theorems for planted clustering and submatrix localization in Sections 5 and 6, respectively.

The following notation is used in the paper. Let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$, and $[m] = \{1, 2, \ldots, m\}$ for any positive integer m. We use c_1, c_2 etc. to denote absolute numerical constants whose values can be made explicit and are independent of the model parameters. We use the standard big-O notations: for two sequences $\{a_n\}, \{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ to mean $a_n \le c_1 b_n$ for an absolute constant c_1 and all n; similarly, $a_n \gtrsim b_n$ and $a_n = \Omega(b_n)$ mean $a_n \ge c_2 b_n$. Moreover, $a_n \asymp b_n$ and $a_n = \Theta(b_n)$ mean both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold.

2. Main Results for Planted Clustering

The planted clustering problem is defined by five parameters $n, r, K \in \mathbb{N}$ and $p, q \in [0, 1]$ with $n \ge rK$.

Definition 1 (Planted Clustering) *Suppose n nodes (which are identified with $[n]$) are divided into two subsets V_1 and V_2 with $|V_1| = rK$ and $|V_2| = n - rK$. The nodes in V_1 are partitioned into r disjoint clusters $C_1^*, \ldots, C_r^*$ (called* true clusters*), where $|C_m^*| = K$ for each $m \in [r]$ and $\bigcup_{m=1}^r C_m^* = V_1$. Nodes in V_2 do not belong to any of the clusters and are called* isolated nodes*. A random graph is generated based on the cluster structure: for each pair of nodes and independently of all others, we connect them by an edge with probability p (called* in-cluster edge density*) if they are in the same cluster, and otherwise with probability q (called* cross-cluster edge density*).*

We emphasize again that the values of p, q, r, and K are allowed to be functions of n. The goal is to exactly recover the true clusters $\{C_m^*\}_{m=1}^r$ up to a permutation of cluster indices given the random graph.

The model parameters (p, q, r, K) are assumed to be known to the algorithms. This assumption is often not necessary and can be relaxed (Chen et al., 2012; Arias-Castro and Verzelen, 2014). It is also possible to allow for non-uniform cluster sizes (Ailon et al., 2013) as well as heterogeneous edge probabilities and node degrees (Chaudhuri et al., 2012; Chen et al., 2014b; Cai and Li, 2015). These extensions are certainly important in practical applications; we do not delve into them, and point to the referenced papers above and the references therein for work in this direction.

To facilitate subsequent discussion, we introduce a matrix representation of the planted clustering problem. We represent the true clusters $\{C_m^*\}_{m=1}^r$ by a *cluster matrix* $Y^* \in \{0,1\}^{n\times n}$, where $Y_{ii}^* = 1$ for $i \in V_1$, $Y_{ii}^* = 0$ for $i \in V_2$, and $Y_{ij}^* = 1$ if and only if nodes i and j are in the same true cluster. Note that the rank of Y^* equals r, hence the name of the high-rank setting. The adjacency matrix of the graph is denoted as A, with the convention $A_{ii} = 0, \forall i \in [n]$. Under the planted clustering model, we have $\mathbb{P}(A_{ij} = 1) = p$ if $Y_{ij}^* = 1$ and $\mathbb{P}(A_{ij} = 1) = q$ if $Y_{ij}^* = 0$ for all $i \neq j$. The problem reduces to recovering Y^* given A.

The planted clustering model generalizes several classical planted models.

- *Planted r-Disjoint-Clique* (McSherry, 2001). Here $p = 1$ and $0 < q < 1$, so r cliques of size K are planted into an Erdős-Rényi random graph $G(n, q)$. The special case with $r = 1$ is known as the *planted clique* problem (Alon et al., 1998).

- *Planted Densest Subgraph* (Arias-Castro and Verzelen, 2014). Here $0 < q < p < 1$ and $r = 1$, so there is a subgraph of size K and density p planted into a $G(n, q)$ graph.

- *Planted Partition* (Condon and Karp, 2001). Also known as the *stochastic blockmodel* (Holland et al., 1983). Here $n = rK$ and $p, q \in (0, 1)$. The special case with $r = 2$ can be called *planted bisection* (Condon and Karp, 2001). The case with $p < q$ is sometimes called *planted noisy coloring* or *planted r-cut* (Decelle et al., 2011; Bollobás and Scott, 2004).

- *Planted r-Coloring* (Alon and Kahale, 1997). Here $n = rK$ and $0 = p < q < 1$, so each cluster corresponds to a group of disconnected nodes that are assigned with the same color.

For clarity we shall focus on the homophily setting with $p > q$; results for the $p < q$ case are similar. In fact, any achievability or converse result for the $p > q$ case immediately implies a corresponding result for $p < q$. To see this, observe that if the graph A is generated from the planted clustering model with $p < q$, then the flipped graph $A' := J - A - I$ (J is the all-one matrix and I is the identity matrix) can be considered as generated with in/cross-cluster edge densities $p' = 1-p$ and $q' = 1-q$, where $p' > q'$. Therefore, a problem with $p < q$ can be reduced to one with $p' > q'$. Clearly the reduction can also be done in the other direction.

2.1 The Impossible Regime: Minimax Lower Bounds

In this section, we characterize the necessary conditions for cluster recovery. Let $\mathcal{Y}$ be the set of cluster matrices corresponding to r clusters of size K; i.e.,

$$\mathcal{Y} = \big\{Y \in \{0,1\}^{n\times n} \,|\text{there exist disjoint clusters } \{C_m\}_{m=1}^r \text{ such that } |C_m| = K, \forall m \in [r], \\ \text{and } Y \text{ is the corresponding cluster matrix}\big\}.$$

We use $\widehat{Y} \equiv \widehat{Y}(A)$ to denote an estimator which takes as input the graph A and outputs an element of $\mathcal{Y}$ as an estimate of the true Y^*. Our results are stated in terms of the Kullback-Leibler (KL) divergence between two Bernoulli distributions with means u and v, denoted by $D(u\|v) := u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v}$. The following theorem gives a lower bound on the minimax error probability of recovering Y^*.

Theorem 2 (Impossible) *Suppose* $128 \le K \le n/2$. *Under the planted clustering model with* $p > q$, *if one of the following two conditions holds:*

$$K \cdot D(q\|p) \le \frac{1}{192}\left[\log(rK) \wedge K\right], \tag{1}$$

$$K \cdot D(p\|q) \le \frac{1}{192}\log n, \tag{2}$$

then

$$\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \frac{1}{4},$$

where the infimum ranges over all measurable function of the graph.

The theorem shows it is fundamentally impossible to recover the clusters with success probability close to 1 in the regime where (1) or (2) holds, which is thus called the *impossible regime*. This regime arises from an *information/statistical barrier*: The KL divergence on the LHSs of (1) and (2) determines how much information of Y^* is contained in the data A. If the in-cluster and cross-cluster edge distributions are close (measured by the KL divergence) or the cluster size is small, then A does not carry enough information to distinguish different cluster matrices.

It is sometimes more convenient to use the following corollary, derived by upper-bounding the KL divergence in (1) and (2) using its Taylor expansion. This corollary was used when we overviewed our results in Section 1.1. See table 1 for its implications for specific planted models.

Corollary 3 *Suppose* $128 \le K \le n/2$. *Under the planted clustering model with* $p > q$, *if any one of the following three conditions holds:*

$$K(p-q)^2 \le \frac{1}{192} q(1-q)\log n, \tag{3}$$

$$Kp \le \frac{1}{193}\left[\log(rK) \wedge K\right], \tag{4}$$

$$Kp\log\frac{p}{q} \le \frac{1}{192}\log n, \tag{5}$$

then $\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}[\widehat{Y} \neq Y^*] \ge \frac{1}{4}$.

Note the asymmetry between the roles of p and q in the conditions (1) and (2); this is made apparent in Corollary 3. To see why the asymmetry is natural, recall that by a classical result of Grimmett and McDiarmid (1975), the largest clique in a random graph $G(n,q)$ has size $k_q = \Theta(\log n/\log(1/q))$ almost surely. Such a clique cannot be distinguished from a true cluster if $K \lesssim k_q$, even when $p = 1$. This is predicted by the condition (5). When $q = 0$, cluster recovery requires $p \gtrsim \frac{\log(rK)}{K}$ to ensure all true clusters are connected within themselves, matching the condition (4). The term K on the RHS of (1) and (4) is relevant only when $K \le \log(rK)$. Potential improvement on this term is left to future work.

Comparison with previous work: When $r = 1$ and $q = 1/2$, our results recover the $K = \Theta(\log n)$ threshold for the classical planted clique problem. For planted partition with $r = O(1)$ clusters of size $K = \Theta(n)$ and $p/q = \Theta(1)$, the work in Chaudhuri et al. (2012); Chen et al. (2014a) establishes the necessary condition $p - q \lesssim \sqrt{p/n}$; our result is stronger by a logarithmic factor.

The work in Abbe et al. (2014) also considers planted partition with $r = 2$ and focus on the special case with the scaling $p, q = \Theta(\log(n)/n)$; they establish the condition $p+q-2\sqrt{pq} < 2\log(n)/n$, which is consistent with our results up to constants in this regime. Compared to previous work, we handle the more general setting where p, q and r may scale arbitrarily with n.

2.2 The Hard Regime: An Optimal Algorithm

In this subsection, we characterize the sufficient conditions for cluster recovery which match the necessary conditions given in Theorem 2 up to constant factors. We consider the Maximum Likelihood Estimator of Y^* under the planted clustering model, which we now derive. The log-likelihood of observing the graph A given a cluster matrix $Y \in \mathcal{Y}$ is

$$\begin{aligned}\log \mathbb{P}_Y(A) &= \log \prod_{i<j} p^{A_{ij}Y_{ij}} q^{A_{ij}(1-Y_{ij})} (1-p)^{(1-A_{ij})Y_{ij}} (1-q)^{(1-A_{ij})(1-Y_{ij})} \\ &= \log \frac{p(1-q)}{q(1-p)} \sum_{i<j} A_{ij}Y_{ij} + \log \frac{1-p}{1-q} \sum_{i<j} Y_{ij} + \log \frac{q}{1-q} \sum_{i<j} A_{ij} + \sum_{i<j} \log(1-q). \quad (6)\end{aligned}$$

Given A, the MLE maximizes the the log-likelihood over the set $\mathcal{Y}$ of all possible cluster matrices. Note that $\sum_{i<j} Y_{ij} = r\binom{K}{2}$ for all $Y \in \mathcal{Y}$, so the last three terms in (6) are independent of Y. Therefore, the MLE for the $p > q$ case is given as in Algorithm 1.

Algorithm 1 Maximum Likelihood Estimator ($p > q$)

$$\widehat{Y} = \arg\max_{Y} \sum_{i,j} A_{ij}Y_{ij} \quad (7)$$

$$\text{s.t. } Y \in \mathcal{Y}. \quad (8)$$

Algorithm 1 is equivalent to finding r disjoint clusters of size$K$ that maximize the number of edges inside the clusters (similar to Densest K-Subgraph), or minimize the number of edges outside the clusters (similar to Balanced Cut) or the disagreements between A and Y (similar to Correlation Clustering in Bansal et al. 2004). Therefore, while Algorithm 1 is derived from the planted clustering model, it is in fact quite general and not tied to the modeling assumptions. Enumerating over the set $\mathcal{Y}$ is computationally intractable in general since $|\mathcal{Y}| = \Omega(e^{rK})$.

The following theorem provides a success condition for the MLE.

Theorem 4 (Hard) . *Under the planted clustering model with $p > q$, there exists a universal constant c_1 such that for any $\gamma \geq 1$, the optimal solution $\widehat{Y}$ to the problem (7)–(8) is unique and equal to Y^* with probability at least $1 - 16(\gamma rK)^{-1} - 256n^{-1}$ if both of the following hold:*

$$\begin{aligned} K \cdot D(q\|p) &\geq c_1 \log(\gamma rK), \\ K \cdot D(p\|q) &\geq c_1 \log n. \end{aligned} \quad (9)$$

We refer to the regime in which the condition (9) holds but (14) below fails as the *hard regime*, as clustering is statistically possible but conjectured to be computationally hard (cf. Conjecture 9). The conditions (9) above and (1)–(2) in Theorem 2 match up to a constant factor under the mild

assumption $K \geq \log(rK)$. This establishes the minimax recovery boundary for planted clustering and the minimax optimality of the MLE up to constant factors.

By lower bounding the KL divergence, we obtain the following corollary, which is sometimes more convenient to use. See Table 1 for its implications for specific planted models.

Corollary 5 *For planted clustering with $p > q$, there exists a universal constant c_2 such that for any $\gamma \geq 1$, the optimal solution $\widehat{Y}$ to the problem (7)–(8) is unique and equal to Y^* with probability at least $1 - 16(\gamma rK)^{-1} - 256n^{-1}$ provided*

$$K(p-q)^2 \geq c_2 q(1-q)\log n, \quad Kp \geq c_2 \log(\gamma rK) \quad \textit{and} \quad Kp \log \frac{p}{q} \geq c_2 \log n. \tag{10}$$

The condition (10) can be simplified to $K(p-q)^2 \gtrsim q(1-q)\log n$ if $q = \Theta(p)$, and to $Kp \log \frac{p}{q} \gtrsim \log n, Kp \gtrsim \log(rK)$ if $q = o(p)$. These match the converse conditions in Corollary 3 up to constants.

Comparison with previous work: Theorem 4 provides the first minimax results (tight up to constant factors) when the number of clusters is allowed to grow, potentially at a *nearly-linear* rate $r = O(n/\log n)$. Interestingly, for a fixed cluster size, the recovery boundary (9) depends very weakly on the number of clusters r though the logarithmic term. For $r = 1$ and $p = 2q = 1$, we recover the recovery boundary for planted clique $K \asymp \log n$. For the planted densest subgraph model where $p/q = \Theta(1)$, p bounded away from 1 and $Kq \gg 1$, the minimax *detection* boundary is shown in Arias-Castro and Verzelen (2014) to be $\frac{(p-q)^2}{q} \asymp \min\{\frac{1}{K}\log\frac{n}{K}, \frac{n^2}{K^4}\}$; our results show that the minimax *recovery* boundary is $\frac{(p-q)^2}{q} \asymp \frac{\log n}{K}$, which is strictly above the detection boundary because $\frac{n^2}{K^4}$ can be much smaller than $\frac{\log n}{K}$. For the planted bisection model with two equal-sized clusters: if $p, q = \Theta(\log(n)/n)$, the sharp recovery boundary is found in Abbe et al. (2014) and Mossel et al. (2015b) to be $K(\sqrt{p} - \sqrt{q})^2 > \log n$, which is consistent with our results up to constants; if $p, q = O(1/n)$, the correlated recovery limit is shown in Mossel et al. (2015a); Massoulié (2014); Mossel et al. (2013) to be $K(p-q)^2 > p+q$, which is consistent with our results up to a logarithmic factor.

2.3 The Easy Regime: Polynomial-Time Algorithms

In this subsection, we present a polynomial-time algorithm for the planted clustering problem and show that it succeeds in the easy regime described in the introduction.

Our algorithm is based on taking a convex relaxation of the MLE in Algorithm 1. Note that the objective function (7) in the MLE is linear, but the constraint $Y \in \mathcal{Y}$ involves a set $\mathcal{Y}$ that is discrete, non-convex and exponentially large. We replace this non-convex constraint with a trace norm (a.k.a. nuclear norm) constraint and a set of linear constraints. This leads to the convexified MLE given in Algorithm 2. Here the trace norm $\|Y\|_*$ is defined as the sum of the singular values of Y. Note that the true Y^* is feasible to the optimization problem (11)–(13) since $\|Y^*\|_* = \text{trace}(Y^*) = rK$.

Algorithm 2 Convexified Maximum Likelihood Estimator ($p > q$)

$$\widehat{Y} = \arg \max_{Y \in \mathbb{R}^{n \times n}} \sum_{i,j} A_{ij} Y_{ij} \tag{11}$$

$$\text{s.t. } \|Y\|_* \le rK, \tag{12}$$

$$\sum_{i,j} Y_{ij} = rK^2, \quad 0 \le Y_{ij} \le 1, \forall i, j. \tag{13}$$

The optimization problem in Algorithm 2 is a semidefinite program (SDP) and can be solved in polynomial time by standard interior point methods or various fast specialized algorithms such as ADMM; e.g., see Jalali and Srebro (2012); Ames (2013). Similarly to Algorithm 1, this algorithm is not strictly tied to the planted clustering model as it can also be considered as a relaxation of Correlation Clustering or Balanced Cut. In the case where the values of r and K are unknown, one may replace the hard constraints (12) and (13) with an appropriately weighted objective function; cf. Chen et al. (2014b).

The following theorem provides a sufficient condition for the success of the convexified MLE. See Table 1 for its implications for specific planted models.

Theorem 6 (Easy) *Under the planted clustering model with $p > q$, there exists a universal constant c_1 such that with probability at least $1 - n^{-10}$, the optimal solution to the problem (11)–(13) is unique and equal to Y^* provided*

$$K^2(p-q)^2 \ge c_1 \left[p(1-q)K \log n + q(1-q)n \right]. \tag{14}$$

When $r = 1$, we refer to the regime where the condition (14) holds and (17) below fails as the *easy regime*. When $r > 1$, the easy regime is where (14) holds and (17) or (18) below fails.

If $p/q = \Theta(1)$, it is easy to see that the smallest possible cluster size allowed by (14) is $K = \Theta(\sqrt{n})$ and the largest number of clusters is $r = \Theta(\sqrt{n})$, both of which are achieved when $p, q, |p - q| = \Theta(1)$. This generalizes the tractability threshold $K = \Omega(\sqrt{n})$ of the classic planted clique problem. If $q = o(p)$ (we call it the high SNR setting), the condition (14) becomes to $Kp \gtrsim \max\{\log n, \sqrt{qn}\}$. In this case, it is possible to go beyond the $\sqrt{n}$ limit on the cluster size. In particular, when $p = \Theta(1)$, the smallest possible cluster size is $K = \Theta(\log n \vee \sqrt{qn})$, which can be much smaller than $\sqrt{n}$.

Remark 7 *Theorem 6 immediately implies guarantees for other tighter convex relaxations. Define the sets $\mathcal{B} := \{Y \mid Eq.(13) \text{ holds}\}$ and*

$$\mathcal{S}_1 := \{Y \mid \|Y\|_* \le rK\},$$
$$\mathcal{S}_2 := \{Y \mid Y \succeq 0; \mathit{trace}(Y) = rK\}.$$

The constraint in Algorithm 2 corresponds to $Y \in \mathcal{S}_1 \cap \mathcal{B}$, while $Y \in \mathcal{S}_2 \cap \mathcal{B}$ is the constraint in a so-called standard SDP relaxation. Clearly $(\mathcal{S}_1 \cap \mathcal{B}) \supseteq (\mathcal{S}_2 \cap \mathcal{B}) \supseteq \mathcal{Y}$. Therefore, if we replace the constraint (12) with $Y \in \mathcal{S}_2$, we obtain a tighter *relaxation of the MLE, and Theorem 6 guarantees that it also succeeds to recover Y^* under the condition (14). The same is true if we consider other tighter relaxations, such as those involving the triangle inequalities (Mathieu and Schudy, 2010),*

the row-wise constraints $\sum_j Y_{ij} \leq K, \forall i$ *(Ames, 2013), the max norm (Jalali and Srebro, 2012) or the Fantope constraint (Vu et al., 2013). For the purpose of this work, these variants of the convex formulation make no significant difference, and we choose to focus on* (11)–(13) *for generality.*

Comparison with previous work: We refer to Chen et al. (2014b) for a survey of the performance of state-of-the-art polynomial-time algorithms under various planted models. Theorem 6 matches and in many cases improves upon existing results in terms of the scaling. For example, for planted partition with general r, the previous best results are $(p-q)^2 \gtrsim p(K\log^4 n + n)/K^2$ in Chen et al. (2012) and $(p-q)^2 \gtrsim pn\,\mathrm{polylog}\, n/K^2$ in Anandkumar et al. (2014). Theorem 6 removes some extra $\log n$ factors, and is also order-wise better when $q = o(p)$ (the high SNR case) or $1-q = o(1)$. For planted r-disjoint-clique, existing results require $1-q$ to be $\Omega((rn+rK\log n)/K^2)$ (McSherry, 2001), $\Omega(\sqrt{n}/K)$ (Ames and Vavasis, 2014) or $\Omega((n + K\log^4 n)/K^2)$ (Chen et al., 2012). We improve them to $\Omega((n + K\log n)/K^2)$.

2.3.1 CONVERSE FOR THE TRACE NORM RELAXATION APPROACH

We have a partial converse to the achievability result in Theorem 6. The following theorem characterizes the conditions under which the trace norm relaxation (11)–(13) provably fails with high probability; we suspect the standard SDP relaxation with the constraint $Y \in \mathcal{S}_2 \cap \mathcal{B}$ also fails with high probability under the same conditions, but we do not have a proof.

Theorem 8 (Easy, Converse) *Under the planted clustering model with $p > q$, for any constant $1 > \epsilon_0 > 0$, there exist positive universal constants c_1, c_2 for which the following holds. Suppose $c_1 \log n \leq K \leq \frac{n}{2}$, $q \geq c_1 \frac{\log n}{n}$ and $p \leq 1 - \epsilon_0$. If*

$$K^2(p-q)^2 \leq c_2(Kp + qn),$$

then with probability at least $1 - n^{-10}$, Y^* is not an optimal solution of the program (11)–(13).

Theorem 8 proves the failure of our trace norm relaxation that has access to the *exact* number and sizes of the clusters. Consequently, replacing the constraints (12) and (13) with a Lagrangian penalty term in the objective would not help for *any* value of the Lagrangian multipliers. Under the assumptions of Theorems 6 and 8, by ignoring log factors, the *sufficient and necessary* condition for the success of our convexified MLE is

$$\frac{p}{K(p-q)^2} + \frac{qn}{K^2(p-q)^2} \dot{\lesssim} 1. \tag{15}$$

We can compare (15) with the success condition (10) for the MLE, which simplifies to

$$\frac{p}{K(p-q)^2} \dot{\lesssim} 1.$$

We see that the convexified MLE is statistically sub-optimal due to the extra second term in (15). This term is responsible for the $K = \Omega(\sqrt{n})$ threshold on the cluster size for the tractability of planted clique. The term has an interesting interpretation. Let $\widetilde{A} := A - q\mathbf{1}\mathbf{1}^\top + qI$ be the centered adjacency matrix. The matrix $E := (Y - \mathbf{1}\mathbf{1}^\top) \circ (\widetilde{A} - \mathbb{E}\widetilde{A})$, [1] i.e., the deviation $\widetilde{A} - \mathbb{E}\widetilde{A}$ restricted to

1. Here $\circ$ denotes the element-wise product.

the inter-cluster node pairs, can be viewed as the "cross-cluster noise matrix". Note that the squared largest singular value of the matrix $\mathbb{E}\widetilde{A} = (p-q)Y^*$ is $K^2(p-q)^2$, whereas the squared largest singular value of E concentrates around $\Theta(qn)$ (see e.g., Chatterjee 2014). Therefore, the second term $\frac{qn}{K^2(p-q)^2}$ in (15) is the "spectral noise-to-signal ratio" that determines the performance of the convexified MLE. In fact, our proofs for Theorems 6 and 8 build on this intuition.

Comparison with previous work: Our converse result in Theorem 8 is inspired by, and improves upon, the recent work in Vinayak et al. (2014), which focuses on the special case $p > 1/2 > q$ and considers a convex relaxation approach that is equivalent to our relaxation (11)–(13) but without the additional equality constraint in (13). The approach is shown to fail when $K^2(p-\frac{1}{2})^2 \lesssim qn$. Our result is stronger in the sense that it applies to a tighter relaxation and a larger region of the parameter space.

2.3.2 LIMITS OF POLYNOMIAL-TIME ALGORITHMS

We compare the minimax recovery threshold in Theorems 2 and 4 with the performance boundary of the polynomial-time convexified MLE in Theorem 6. In general, there exists a substantial gap between these two boundaries (cf. Figure 1). We conjecture that no polynomial-time algorithm has order-wise better statistical performance than the convexified MLE and succeeds significantly beyond the condition (14) in Theorem 6.

Conjecture 9 *For any constant $\epsilon > 0$, there is no algorithm with running time polynomial in n that, for all n and with probability at least $1/2$, outputs the true Y^* of the planted clustering problem with $p > q$ and*

$$(p-q)^2K^2 \leq n^{-\epsilon}\left(Kp(1-p) + q(1-q)n\right). \tag{16}$$

If the conjecture is true, then in the asymptotic setting $p = 2q = n^{-\alpha}$ and $K = n^{\beta}$, the *computational limit* for the cluster recovery is given by $\beta = \frac{\alpha}{2} + \frac{1}{2}$, i.e., the boundary between the green and red regimes in Figure 1.

A rigorous proof of Conjecture 9 seems difficult with current techniques. There are other possible convex formulations for planted clustering. The space of possible polynomial-time algorithms is even larger. It is impossible for us to study each of them separately and obtain a converse result as in Theorem 8. There are however several evidences that support the conjecture:

- The special case with $p = 2q = 1$ corresponds to the $K = o(\sqrt{n})$ regime for the classical Planted Clique problem, which is conjectured to be computationally hard (Alon et al., 2007; Rossman, 2010; Feldman et al., 2013), and has been used as an assumption for proving other hardness results (Hazan and Krauthgamer, 2011; Juels and Peinado, 2000; Koiran and Zouzias, 2014). Graphically, the Planted Clique hardness corresponds to the division of the $\alpha = 0$ line of the parameter space shown in Figure 1 (with $r = 1$). Conjecture 9 can be considered as *a generalization of the Planted Clique conjecture* to the whole parameter space, that is, to the setting with multiple clusters and general values of p and q. This generalized conjecture may be used to study the hardness of other problems (Chen, 2015).

- A weaker version of such generalization is proved in Hajek et al. (June, 2014): they show that in the setting with a single cluster, no polynomial-time algorithm can reliably recover the planted clusters if $\beta < \alpha/4 + 1/2$ conditioned on the planted clique hardness hypothesis.

- As discussed earlier, if (16) holds, then the graph spectrum is dominated by noise and fails to reveal the underlying cluster structure. The condition (16) therefore represents a "spectral barrier" for clustering. The work in Nadakuditi and Newman (2012) uses this spectral barrier argument to prove the failure of a large class of algorithms that rely on the graph spectrum. The proof of our Theorem 8 reveals that the convexified MLE fails for a similar reason.
- In the sparse graph case with $p, q = O(1/n)$, it is argued in Decelle et al. (2011), using non-rigorous but deep arguments from statistical physics, that it is intractable to achieve the correlated recovery under Condition (16).

We note that in the "high SNR" case with $\frac{p}{q} \gtrsim \frac{n}{K \log n}$ and $\log K \gtrsim \log n$, the minimax limit and performance boundary of the convexified MLE coincide up to constant factors at $K(p-q)^2 \asymp p(1-q)\log n$. This means, up to constants, the convex relaxation is tight and hence a computationally efficient and statistically order-optimal estimator, so the hard regime disappears. [2] Similar phenomenon can be observed in the setting with linear size clusters $K \asymp n$ (which implies $r \lesssim 1$) and $p \asymp q$, as is illustrated by the $\beta = 1$ line in Figure 1.

2.4 The Simple Regime: A Counting Algorithm

In this subsection, we consider a simple recovery procedure in Algorithm 3, which is based on counting node degrees and common neighbors.

Algorithm 3 A Simple Counting Algorithm

1. (Identify isolated nodes) For each node i, compute its degree d_i. Declare i as isolated if $d_i < \frac{(p-q)K}{2} + qn$.
2. (Identify clusters when $r > 1$) For every pair of non-isolated nodes i, j, compute the number of common neighbors $S_{ij} := \sum_{k:k\neq i, k\neq j} A_{ik}A_{jk}$, and assign them into the same cluster if $S_{ij} > \frac{(p-q)^2 K}{3} + 2Kpq + q^2(n-2K)$. Declare error if inconsistency found.

We note that steps 1 and 2 of Algorithm 3 are considered in Kučera (1995) and Dyer and Frieze (1989) respectively for the special cases of recovering a single planted clique or two planted clusters. Let E be the set of edges. It is not hard to see that step 1 runs in time $O(|E|)$ and step 2 runs in time $O(n|E|)$, since each node only needs to look up its local neighborhood up to distance two. It is possible to achieve even smaller expected running time using clever data structures.

The following theorem provides sufficient conditions for the simple counting algorithm to succeed. Compared to the previous work in Kučera (1995); Dyer and Frieze (1989), our results apply to general values of p, q, r, and K. See Table 1 for its implications for specific planted models.

Theorem 10 (Simple) *For planted clustering with $p > q$, there exist universal constants c_1, c_2 such that Algorithm 3 correctly finds the isolated nodes with probability at least $1 - 2n^{-1}$ if*

$$K^2(p-q)^2 \geq c_1[Kp(1-q) + nq(1-q)]\log n, \tag{17}$$

and finds the clusters with probability at least $1 - 4n^{-1}$ if further

$$K^2(p-q)^4 \geq c_2[Kp^2(1-q^2) + nq^2(1-q^2)]\log n. \tag{18}$$

2. The hard regime may still exist if constant factors are concerned; cf. Mossel et al. 2015a; Decelle et al. 2011.

Remark 11 *If $p, q \to 1$ as $n \to \infty$, we can obtain slightly better performance by counting the common* non-*neighbors in Step 2, which succeeds under condition (18) with p and q replaced by $1-p$ and $1-q$, respectively, i.e., the RHS of (18) simplifies to $c_2 n(1-q)^2 \log n$.*

In the case with a single clusters $r = 1$, we refer to the regime where the condition (17) holds as the *simple regime*; in the case with $r > 1$, the simple regime is where both conditions (17) and (18) hold. It is instructive to compare these conditions with the success condition (14) for the convexified MLE. The condition (17) has an additional $\log n$ factor on the RHS. This means when $r = 1$ and the only task is to find the isolated nodes, the counting algorithm performs nearly as well as the sophisticated convexified MLE. On the other hand, when $r > 1$ and one needs to distinguish between different clusters, the convexified MLE order-wise outperforms the counting algorithm whenever $p/q = \Theta(1)$, as the condition (18) is order-wise more restrictive than (14). Nevertheless, when $p, q, p - q = \Theta(1)$, both algorithms can recover $\widetilde{O}(\sqrt{n})$ clusters of size $\widetilde{\Omega}(\sqrt{n})$, making the simple counting algorithm a legitimate candidate in such a setting and a benchmark to which other algorithms can be compared with.

In the high SNR case with $q = o(p)$, the counting algorithm can recover clusters with size much smaller than $\sqrt{n}$; e.g., if $p = \Theta(1)$ and $q = o(1)$, it only requires $K \gtrsim \max\{\log n, \sqrt{qn \log n}\}$.

2.4.1 CONVERSE FOR THE COUNTING ALGORITHM

We have a (nearly-)matching converse to Theorem 10. The following theorem characterizes when the counting algorithm provably fails.

Theorem 12 (Simple, Converse) *Under the planted clustering model with $p > q$, for any constant $0 < \epsilon_0 < 1$, there exist universal constants $c_1, c_2 > 0$ for which the following holds. Suppose $K \le \frac{n}{2}$, $p \le 1 - \epsilon_0$, $q \ge c_1 \log n/n$ and $Kp^2 + nq^2 \ge c_1 \log n$. Algorithm 3 fails to correctly identify all the isolated nodes with probability at least $1/4$ if*

$$K^2(p-q)^2 < c_2 \left[(Kp + nq)\log(rK) + nq \log(n - rK)\right], \tag{19}$$

and fails to correctly recover all the clusters with probability at least $1/4$ if

$$K^2(p-q)^4 < c_2(Kp^2 + nq^2)\log(rK). \tag{20}$$

Remark 13 *Theorem 12 requires a technical condition $Kp^2 + nq^2 \ge c_1 \log n$, which is actually not too restrictive. If $Kp^2 + nq^2 = o(\log n)$, then two nodes from the same cluster will have no common neighbor with probability $(1-p^2)^K (1-q^2)^{n-K} \ge \exp[-\Theta(p^2 K + q^2(n-K))] = \exp[-o(\log n)]$, so Algorithm 3 cannot succeed with the probability specified in Theorem 10.*

Apart from some technical conditions, Theorems 10 and 12 show that the conditions (17) and (18) are both sufficient and necessary. In particular, the counting algorithm cannot succeed outside the simple regime, and is indeed strictly weaker in separating different clusters as compared to the convexified MLE. Our proof reveals that the performance of the counting algorithm is limited by a *variance barrier*: The RHS of (17) and (18) are associated with the variance of the node degrees and common neighbors (i.e., d_i and S_{ij} in Algorithm 3), respectively. There exist nodes whose degrees deviate from their expected value on the order of the standard deviation, and if the condition (17) does not hold, then the deviation will outweigh the difference between the expected degrees of the isolated nodes and those of the non-isolated nodes. A similar argument applies to the number of common neighbors.

3. Main Results for Submatrix Localization

In this section, we turn to the submatrix localization problem, sometimes known as bi-clustering (Balakrishnan et al., 2011a). We consider the following specific setting, which is defined by six parameters $n_L, n_R, K_L, K_R, r \in \mathbb{N}$, and $\mu \in \mathbb{R}_+$ such that $n_L \geq rK_L$ and $n_R \geq rK_R$. We use the shorthand notation $n := n_L \vee n_R$.

Definition 14 (Submatrix Localization) *A random matrix $A \in \mathbb{R}^{n_L \times n_R}$ is generated as follows. Suppose that rK_L rows of A are partitioned into r disjoint subsets $\{C_1^*, \ldots, C_r^*\}$ of equal size K_L, and rK_R columns of A are partitioned into r disjoint subsets $\{D_1^*, \ldots, D_r^*\}$ of equal size K_R. For each (i,j), we have $A_{ij} = \mu + \Delta_{ij}$ if $(i,j) \in C_m^* \times D_m^*$ for some $m \in [r]$ and $A_{ij} = \Delta_{ij}$ otherwise, where $\mu > 0$ is a fixed number and (Δ_{ij}) are i.i.d. zero-mean sub-Gaussian random variables with parameter 1.* [3] *The goal is to recover the locations of the hidden submatrices $\{(C_m^*, D_m^*), m \in [r]\}$ given the matrix A.*

In the language of bi-clustering, the sets $\{C_1^*, \ldots, C_r^*\}$ are called *left clusters* and $\{D_1^*, \ldots, D_r^*\}$ are called *right clusters*. Row (column, resp.) indices which do not belong to any cluster are called *isolated* left (right, resp.) nodes. One can think of A as the bipartite affinity matrix between the n_L left nodes and n_R right nodes, and the goal is to recover the left and right clusters. Similarly as before, we define the *bi-clustering matrix* $Y^* \in \{0,1\}^{n_L \times n_R}$, where $Y_{ij}^* = 1$ if and only if $(i,j) \in C_m^* \times D_m^*$ for some $m \in [r]$. The problem reduces to recovering Y^* given A.

As before, all the parameters μ, K_L, K_R, r are allowed to scale with n_L and n_R, and we assume that their values are known. Note that it is without loss of generality to assume the mean of A_{ij} is zero outside the submatrices and the variance of A_{ij} is one, because otherwise we can shift and rescale A. The above model generalizes the previous submatrix localization/detection models (Ma and Wu, 2015; Butucea and Ingster, 2013; Arias-Castro et al., 2011) and bi-clustering models (Kolar et al., 2011; Balakrishnan et al., 2011a) which consider the special case with a *single* submatrix (i.e., $r = 1$).

In the next four subsections, we shall focus on the low-SNR setting $\mu^2 = O(\log n)$ and present theorems establishing the four regimes. These results parallel those for the planted clustering. In the high SNR setting $\mu^2 = \Omega(\log n)$, the submatrices can be easily identified by naive element-wise thresholding, so we deal with this case separately in the last subsection.

3.1 The Impossible Regime: Minimax Lower Bounds

The following theorem gives conditions on $(n_L, n_R, K_L, K_R, \mu)$ under which the minimax error probability is large and thus it is statistically impossible to reliably locate the submatrices. With slight abuse of notation, we use $\mathcal{Y} \subset \{0,1\}^{n_L \times n_R}$ to denote the set of all possible bi-clustering matrices corresponding to r left (right, resp.) clusters of equal size K_L (K_R, resp.).

Theorem 15 (Impossible) *Under the submatrix localization model, suppose $\{A_{ij}\}$ are Gaussian random variables, $K_L \leq n_L/2$, $K_R \leq n_R/2$, and $n_L, n_R \geq 128$. If*

$$\mu^2 \leq \frac{1}{12} \max\left\{ \frac{\log(n_R - K_R)}{K_L}, \frac{\log(n_L - K_L)}{K_R} \right\}, \tag{21}$$

then $\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \frac{1}{2}$, where the infimum ranges over all measurable functions of A.

3. A random variable X is said to be sub-Gaussian with parameter 1 if $\mathbb{E}[e^{tX}] \leq e^{t^2/2}$ for all $t \in \mathbb{R}$.

The regime where (21) holds is called the *impossible* regime, corresponding to an information barrier that no algorithm can break. We note the similarity between the impossible regimes for submatrix localization and planted clustering. In particular, if we assume the in/cross-cluster edges in planted clustering have comparable variance, i.e., $\frac{p(1-p)}{q(1-q)} = \Theta(1)$, then the conditions (21) and (3) coincide up to constant factors by setting $n_L = n_R = n, K_L = K_R = K$, and $\mu = \frac{p-q}{\sqrt{q(1-q)}}$. Such correspondence also exists in the next three regimes.

Comparison with previous work: Theorem 15 holds in the general high rank setting with arbitrary $r \geq 1$. In $r = 1$ case, our result recovers the minimax lower bound in Kolar et al. (2011).

3.2 The Hard Regime: Optimal Algorithm

Recall that $\mathcal{Y}$ is the set of all valid bi-clustering matrices. We consider the combinatorial optimization problem given in Algorithm 4. In the setting where $\{\Delta_{ij}\}$ are Gaussian random variables, this can be shown to be the MLE of Y^*.

Algorithm 4 Maximum Likelihood Estimator

$$\widehat{Y} = \arg\max_{Y \in \mathcal{Y}} \sum_{i,j} A_{ij} Y_{ij}. \tag{22}$$

Theorem 16 below provides a success condition for Algorithm 4.

Theorem 16 (Hard) *Suppose $K_L, K_R \geq 8$. There exists a constant c_1 such that with probability at least $1 - 512en^{-1}$, the optimal solution to the problem (22) is unique and equals Y^* if*

$$\mu^2 \geq c_1 \frac{\log n}{K_L \wedge K_R}. \tag{23}$$

We refer to the regime where the condition (23) holds and (27) fails as the *hard* regime. Note that the bound (23) matches (21) up to a constant factor, so Algorithm 4 is minimax optimal up to a constant factor. Theorems 15 and 16 together establish the minimax recovery boundary for submatrix localization at $\mu^2 \asymp \frac{\log n}{K_L \wedge K_R}$.

Comparison with previous work: Theorem 16 provides the first minimax-optimal achievability result when the number r of submatrices may grow with n_L and n_R. In particular, r is allowed to grow at a nearly linear rate $r = O(n/\log n)$ assuming $n_L = n_R = n$. In the special case with a single planted submatrix ($r = 1$), Theorem 16 recovers the achievability result in Kolar et al. (2011).

3.3 The Easy Regime: Polynomial-Time Algorithms

As previous, we obtain a convex relaxation of the combinatorial MLE formulation (22) by replacing the constraint $Y \in \mathcal{Y}$ with the trace norm and linear constraints, for which we use the fact that the true Y^* satisfies $\|Y^*\|_* = r\sqrt{K_L K_R}$. This is given as Algorithm 5, which is a semidefinite program (SDP) and can be solved in polynomial time.

Algorithm 5 Convexified Maximum Likelihood Estimator

$$\widehat{Y} = \arg\max_{Y \in \mathbb{R}^{n\times n}} \sum_{i,j} A_{ij} Y_{ij} \tag{24}$$

$$\|Y\|_* \le r\sqrt{K_L K_R}, \tag{25}$$

$$\sum_{i,j} Y_{ij} = rK_L K_R, \quad 0 \le Y_{ij} \le 1, \forall i, j. \tag{26}$$

The following theorem provides a sufficient condition for the success of Algorithm 5.

Theorem 17 (Easy) *There exists a universal constant c_1 such that with probability at least $1-n^{-10}$, the optimal solution to the program (24)–(26) in Algorithm 5 is unique and equals Y^* if*

$$\mu^2 \ge c_1 \left(\frac{\log n}{K_L \wedge K_R} + \frac{n}{K_L K_R} \right). \tag{27}$$

When $r = 1$, the *easy regime* refers to where the condition (27) holds but (29) fails. When $r > 1$, the easy regime is where the condition (27) holds but (30) fails. Suppose $n_L = n_R = n$ and $K_L = K_R = K$; the convexified MLE is guaranteed to succeed when $\mu^2 \gtrsim \frac{K \log n + n}{K^2}$.

The following theorem provides a nearly matching converse to Theorem 17.

Theorem 18 (Easy, Converse) *There exist positive universal constants c_1, c_2 such that the following holds. Under the submatrix localization model, suppose $\mu \le 1/100$, $n_L = n_R = n$, $K_L = K_R = K$, $c_1 \log n \le K \le \frac{n}{2}$, and (Δ_{ij}) are Gaussian random variables. If*

$$\mu^2 \le c_2 \frac{n}{K^2}, \tag{28}$$

then with probability at least $1 - n^{-10}$, Y^* is not an optimal solution of the program (24)–(26).

Theorems 17 and 18 together establish that the recovery boundary for the convexified MLE in Algorithm 5 is $\mu^2 \asymp \frac{n}{K^2}$ ignoring logarithmic factors. There is a substantial gap from the minimax boundary $\mu^2 \asymp \frac{1}{K}$ established in the last two subsections (again ignoring logarithmic factors). Our analysis reveals that the performance of the convexified MLE is determined by a spectral barrier similar to that in planted clustering. In particular, the squared largest singular values of the *signal matrix* Y^* and the *noise matrix* $A - \mathbb{E}A$ are $\Theta(\mu^2 K^2)$ and $\Theta(n)$, respectively, so the condition $\mu^2 \gtrsim \frac{n}{K^2}$ for the convexified MLE can be seen as an spectral SNR condition.

As in the planted clustering model, we conjecture that no polynomial-time algorithm can achieve better statistical performance than the convexified MLE.

Conjecture 19 *For any constant $\epsilon > 0$, there is no algorithm with running time polynomial in n that, for all n and with probability at least $1/2$, outputs the true Y^* for the submatrix localization problem with $\mu \le 1$, $n_L = n_R = n$, $K_L = K_R = K \ge c_1 \log n$ and*

$$\mu^2 \le \frac{n^{1-\epsilon}}{K^2}.$$

Comparison with previous work: The achievability and converse results in Theorems 17 and 18 hold even when r grows with n. In the special case with $r = 1$, the work in Kolar et al. (2011) considers a convex relaxation of sparse singular value decomposition; they focus on the high SNR regime with $\mu^2 \gtrsim \log n$, and show that the performance of their convex relaxation is no better than a simple element-wise thresholding approach (cf. Section 3.5). Our convex program is different from theirs, and succeeds in the low SNR regime provided $\mu^2 \gtrsim \frac{K \log n + n}{K^2}$. The work in Ames (2013) studies the success conditions of a convex formulation similar to Kolar et al. (2011); with the additional assumption of bounded support of the distribution of A_{ij}, they show that their approach succeeds under an order-wise more restricted condition $\mu^2 \gtrsim \frac{n \cdot r}{K^2}$.

3.4 The Simple Regime: A Thresholding Algorithm

We consider a simple thresholding algorithm as given in Algorithm 6. The algorithm computes the column and row sums of A as well as the correlation between the columns and rows. It is similar in spirit to the simple counting Algorithm 3 for the planted clustering problem.

Algorithm 6 A Simple Thresholding Algorithm

1. (Identify isolated nodes) For each left node $i \in [n_L]$, declare it as isolated if the row sum $d_i := \sum_{j=1}^{n_R} A_{ij} \leq \frac{\mu K_R}{2}$. For each right node $j \in [n_R]$, declare it as isolated if the column sum $d'_j := \sum_{i=1}^{n_L} A_{ij} \leq \frac{\mu K_L}{2}$.
2. (Identify clusters when $r > 1$) For each pair of non-isolated left nodes $i, i' \in [n_L]$, assign them to the same cluster if $S_{ii'} := \sum_{j=1}^{n_R} A_{ij} A_{i'j} \geq \frac{\mu^2 K_R}{2}$. Declare error if inconsistency is found. Assign the non-isolated right nodes into clusters in a similar manner. Let $\{C_k\}$ and $\{D_k\}$ be the resulting left and right clusters.
3. (Associate left and right clusters) For each $k \in [r]$ and $l \in [r]$, associate the left cluster C_k with the right cluster D_l if the block sum $B_{kl} := \sum_{i \in C_k, j \in D_l} A_{ij} \geq \mu K_L K_R / 2$.

Steps 1, 2 and 3 of the algorithm run in time $O(n_L n_R)$, $O(n_L^2 n_R + n_R^2 n_L)$ and $O(n_L n_R)$, respectively. We note that Step 1 is previously considered in Kolar et al. (2011) for locating a single submatrix. The following theorem provides success conditions for this simple algorithm.

Theorem 20 (Simple) *There exist universal constants c_1, c_2 such that Algorithm 6 identifies the isolated nodes with probability at least $1 - en_L^{-1} - en_R^{-1}$ if*

$$\mu^2 \geq c_1 \max \left\{ \frac{n_L \log n_R}{K_L^2}, \frac{n_R \log n_L}{K_R^2} \right\}, \tag{29}$$

and exactly recovers Y^* with probability at least $1 - e(rK_L)^{-1} - e(rK_R)^{-1} - en^{-1}$ if further

$$\mu^4 \geq c_2 \max \left\{ \frac{n_L \log(rK_R)}{K_L^2}, \frac{n_R \log(rK_L)}{K_R^2} \right\}. \tag{30}$$

When $r = 1$, we refer to the regime for which the condition (29) holds as the *simple* regime. When $r > 1$, the simple regime is where both conditions (29) and (30) hold.

We provide a converse to Theorem 20. The following theorem shows that the conditions (29) and (30) are also (nearly) necessary for the simple thresholding algorithm to succeed.

Theorem 21 (Simple, Converse) *Under the submatrix localization model where the distributions of $\{A_{ij}\}$ are Gaussian, there exist universal constants c_1, c_2 such that with probability at least $1 - e^{-(rK_L)^{\Omega(1)}} - e^{-(rK_R)^{\Omega(1)}}$, Algorithm 6 fails to correctly identify all the isolated nodes if*

$$\mu^2 \le c_1 \max\left\{\frac{n_L \log n_R}{K_L^2}, \frac{n_R \log n_L}{K_R^2}\right\}, \tag{31}$$

and fails to correctly recover all the clusters if $n_L = \Omega(rK_R)$, $n_R = \Omega(rK_L)$ and

$$\mu^4 \le c_2 \max\left\{\frac{n_L \log(rK_R)}{K_L^2}, \frac{n_R \log(rK_L)}{K_R^2}\right\}. \tag{32}$$

When $n_L = n_R = n$, $K_L = K_R = K$, Theorems 20 and Theorem 21 establish that the recovery boundary for the simple thresholding algorithm is $\mu^2 \asymp \frac{n \log n}{K^2}$ if $r = 1$, and $\mu^2 \asymp \frac{\sqrt{n \log n}}{K}$ if $r > 1$ and $rK = \Theta(n)$. Comparing with the success condition (27) for the convex optimization approach, we see that the simple thresholding algorithm is order-wise less powerful in separating different submatrices. Similar to planted clustering, the performance is determined by the variance barrier associated with the variance of the quantities d_i and $S_{ii'}$ computed in Algorithm 6.

3.5 The High SNR Setting

As mentioned before, the high SNR setting with $\mu^2 = \Omega(\log n)$ can be handled by a simple element-wise thresholding algorithm, which is given in Algorithm 7.

Algorithm 7 Element-wise Thresholding for Submatrix Localization

For each $(i, j) \in [n_L] \times [n_R]$, set $\widehat{Y}_{ij} = 1$ if $A_{ij} \ge \frac{1}{2}\mu$, and $\widehat{Y}_{ij} = 0$ otherwise. Output $\widehat{Y}$.

For the special case with one submatrix ($r = 1$), the success of element-wise thresholding in the high SNR setting is proved in Kolar et al. (2011). Their result can be easily extended to the general case with $r \ge 1$. We record this extension in Theorem 22 below. The theorem also shows that element-wise thresholding fails if $\mu^2 = o(\log n)$, so it is not very useful in the low SNR setting.

Theorem 22 (Element-wise Thresholding) *There exists a universal constant $c_1 > 4$ such that the following holds. Algorithm 7 outputs $\widehat{Y} = Y^*$ with probability at least $1 - n^{-3}$ provided*

$$\mu^2 > c_1 \log n. \tag{33}$$

If the distributions of the A_{ij}'s are Gaussian, then with probability at least $1 - n^{-3}$, the output of Algorithm 7 satisfies $\widehat{Y} \ne Y^*$ provided

$$\mu^2 \le 4 \log n. \tag{34}$$

4. Discussion and Future Work

In this paper, we show that the planted clustering problem and the submatrix localization problem admit successively faster algorithms with weaker statistical performance. We provide sufficient and

necessary conditions for the success of the intractable MLE, the convexified MLE and the simple counting/thresholding algorithm, showing that they work in progressively smaller regions of the parameter space. This thus represents a series of tradeoffs between the statistical and computational performance. Our results hold in the high-rank setting with a growing number of clusters or submatrices. Our results indicate that there may exist a large gap between the information limit and the computational limit, i.e., the information limit might not be achievable via polynomial-time algorithms.

Several future directions are of interest. Immediate goals include removing some of the technical assumptions in our theorems. It is useful in practice to identify a finer spectrum, ideally close to a continuum, of computational-statistical tradeoffs. It is also interesting to extend to the settings with overlapping clusters and submatrices, and to the those where the values of the model parameters are unknown. Proving our conjectures on the computational hardness in the hard regime is also interesting, and this direction is pursued in Ma and Wu (2015); Hajek et al. (June, 2014).

5. Proofs for Planted Clustering

Throughout this section, we consider the planted clustering model with $p > q$. Let $n_1 := rK = |V_1|$ and $n_2 := n - rK = |V_2|$ be the numbers of non-isolated and isolated nodes, respectively.

5.1 Proof of Theorem 2 and Corollary 3

In the sequel we will make use of the following upper and lower bounds on the KL divergence $D(u\|v)$ between two Bernoulli distributions with parameter $u \in [0,1]$ and $v \in [0,1]$. We have

$$D\left(u\|v\right) := u\log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v} \overset{(a)}{\leq} u\frac{u-v}{v} + (1-u)\frac{v-u}{1-v} = \frac{(u-v)^2}{v(1-v)}, \tag{35}$$

where (a) follows from the inequality $\log x \leq x-1, \forall x \geq 0$. Moreover, viewing $D(x\|v)$ as a function of x and using the Taylor's expansion, we can find some $\xi \in [u \wedge v, u \vee v]$ such that

$$D\left(u\|v\right) = D\left(v\|v\right) + (u-v)D'\left(v\|v\right) + \frac{(u-v)^2}{2}D''\left(\xi\|v\right) \overset{(b)}{\geq} \frac{(u-v)^2}{2(u\vee v)[1-(u\wedge v)]}, \tag{36}$$

where (b) follows from $D'\left(v\|v\right) = 0$ and $D''\left(\xi\|v\right) = 1/[\xi(1-\xi)]$.

Theorem 2 is established through the following three lemmas, each of which provides a sufficient condition for having a non-vanishing error probability.

Lemma 23 *Suppose that* $128 \leq K \leq n/2$. *Let* $\delta := \frac{n_1(K-1)}{n(n-1)}$ *and* $\bar{p} := \delta p + (1-\delta)q$. *We have* $\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \frac{1}{2}$ *if*

$$K \cdot D(p\|\bar{p}) + \frac{n^2}{n_1}(1-\delta)\frac{(q-\bar{p})^2}{\bar{p}(1-\bar{p})} \leq \frac{1}{4}\log\frac{n}{K}. \tag{37}$$

Moreover, the condition (37) *is implied by*

$$K(p-q)^2 \leq \frac{1}{4}q(1-q)\log\frac{n}{K}. \tag{38}$$

Proof We use an information theoretical argument via Fano's inequality. Recall that $\mathcal{Y}$ is the set of all cluster matrices corresponding to r clusters of size K. Let $\mathbb{P}_{(Y^*,A)}$ be the joint distribution of (Y^*, A) when Y^* is sampled from $\mathcal{Y}$ uniformly at random and then A is generated according to the planted clustering model with Y^* being the true cluster matrix. Lower-bounding the supremum by the average, we have

$$\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \inf_{\widehat{Y}} \mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right].$$

Therefore it suffices to bound $\mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right]$ from below. Let $H(X)$ denote the entropy of a random variable X and $I(X;Z)$ the mutual information between X and Z. By Fano's inequality, we have for any $\widehat{Y}$,

$$\mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right] \geq 1 - \frac{I(Y^*;A)+1}{\log|\mathcal{Y}|}. \tag{39}$$

We first bound $\log|\mathcal{Y}|$. Simple counting gives that $|\mathcal{Y}| = \binom{n}{n_1}\frac{n_1!}{r!(K!)^r}$. Note that $\binom{n}{n_1} \geq (\frac{n}{n_1})^{n_1}$ and $\sqrt{n}(\frac{n}{e})^n \leq n! \leq e\sqrt{n}(\frac{n}{e})^n$. It follows that

$$|\mathcal{Y}| \geq (n/n_1)^{n_1} \frac{\sqrt{n_1}(n_1/e)^{n_1}}{e\sqrt{r}(r/e)^r e^r K^{r/2}(K/e)^{n_1}} \geq \left(\frac{n}{K}\right)^{n_1} \frac{1}{e(r\sqrt{K})^r}.$$

This implies $\log|\mathcal{Y}| \geq \frac{1}{2}n_1 \log\frac{n}{K}$ under the assumption that $8 \leq K \leq n/2$ and $n \geq 32$.

We next upper bound $I(Y^*;A)$. Note that $H(A) \leq \binom{n}{2}H(A_{12})$ because the A_{ij}'s are identically distributed by symmetry. Furthermore, the A_{ij}'s are mutually independent conditioned on Y^*, so $H(A|Y^*) = \binom{n}{2}H(A_{12}|Y^*_{12})$. It follows that $I(Y^*;A) = H(A) - H(A|Y^*) \leq \binom{n}{2}I(Y^*_{12};A_{12})$. We can bound $I(Y^*_{12};A_{12})$ below. Direct counting gives

$$\mathbb{P}(Y^*_{12}=1) = \frac{\binom{n-2}{K-2}\binom{n-K}{K}\cdots\binom{n-rK+K}{K}\frac{1}{(r-1)!}}{|\mathcal{Y}|} = \frac{n_1(K-1)}{n(n-1)} = \delta,$$

and thus $\mathbb{P}(A_{12}=1) = \bar{p} := \delta p + (1-\delta)q$. Therefore, $I(Y^*_{12};A_{12}) = \delta D\left(p\|\bar{p}\right) + (1-\delta)D\left(q\|\bar{p}\right)$. Using the upper bound (35) on the KL divergence and the condition (37), we obtain

$$I(Y^*_{12};A_{12}) = \delta D(p\|\bar{p}) + (1-\delta)D(q\|\bar{p}) \leq \delta D(p\|\bar{p}) + (1-\delta)\frac{(q-\bar{p})^2}{\bar{p}(1-\bar{p})} \leq \frac{n_1}{4n^2}\log\frac{n}{K}.$$

It follows that $I(Y^*;A) \leq \binom{n}{2}I(Y^*_{12}\|A_{12}) \leq \frac{n_1}{8}\log\frac{n}{K}$. Substituting into (39) gives

$$\mathbb{P}_{(Y^*,A)}\left[Y \neq Y^*\right] \geq 1 - \frac{\frac{n_1}{4}\log\frac{n}{K}+2}{n_1\log\frac{n}{K}} = \frac{3}{4} - \frac{2}{n_1\log\frac{n}{K}} \geq \frac{1}{2},$$

where the last inequality holds because $K \leq n/2$ and $n_1 \geq 32$. This proves the sufficiency of (37).

We turn to the second part of the lemma. Observe that

$$\begin{aligned} K\cdot D(p\|\bar{p}) + \frac{n^2}{n_1}(1-\delta)\frac{(q-\bar{p})^2}{\bar{p}(1-\bar{p})} &\overset{(a)}{\leq} K\frac{(p-\bar{p})^2}{\bar{p}(1-\bar{p})} + \frac{K}{\delta}(1-\delta)\frac{(q-\bar{p})^2}{\bar{p}(1-\bar{p})} \\ &= K\frac{\delta(1-\delta)(p-q)^2}{\bar{p}(1-\bar{p})} \overset{(b)}{\leq} \frac{K(p-q)^2}{q(1-q)}, \end{aligned}$$

where (a) holds due to $\delta \le \frac{n_1 K}{n^2}$ and (35), and (b) holds because $\bar{p}(1-\bar{p}) \ge \delta p(1-p) + (1-\delta)q(1-q) \ge (1-\delta)q(1-q)$ thanks to the concavity of $x(1-x)$. Combining the last display equation with (38) gives (37). ■

Lemma 24 *Suppose* $128 \le K \le n/2$. *We have* $\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \frac{1}{2}$ *if*

$$K \max\left\{D(p\|q), D(q\|p)\right\} \le \frac{1}{24}\log(n-K). \tag{40}$$

Proof Let $\bar{M} := n - K$, and $\bar{\mathcal{Y}} := \{Y_0, Y_1, \ldots, Y_{\bar{M}}\}$ be a subset of $\mathcal{Y}$ with cardinality $\bar{M}+1$ to be specified later. Let $\bar{\mathbb{P}}_{(Y^*,A)}$ denote the joint distribution of (Y^*, A) when we first sample Y^* from $\bar{\mathcal{Y}}$ uniformly at random, and then sample A according to the planted clustering model with Y^* being the true cluster matrix. By Fano's inequality, we have

$$\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \inf_{\widehat{Y}} \bar{\mathbb{P}}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right] \ge 1 - \frac{I(Y^*; A) + 1}{\log|\bar{\mathcal{Y}}|}. \tag{41}$$

We construct $\bar{\mathcal{Y}}$ as follows. Let Y_0 be the cluster matrix such that the clusters $\{C_l\}_{l=1}^r$ are given by $C_l = \{(l-1)K+1, \ldots, lK\}$. Informally, each Y_i with $i \ge 1$ is obtained from Y_0 by swapping the cluster memberships of the nodes K and $K+i$. Formally, for each $i \in [\bar{M}]$: (1) if the node $(K+i)$ belongs to cluster C_l for some l, then Y_i is the cluster matrix for which the first cluster consists of the nodes $\{1, 2, \ldots, K-1, K+i\}$, the l-th cluster is given by $C_l \setminus \{K+i\} \cup \{K\}$, and all the other clusters are identical to those given by Y_0; (2) if the node $(K+i)$ is an isolated node in Y_0 (i.e., it does not belong to any cluster), then Y_i is the cluster matrix for which the first cluster consists of the nodes $\{1, 2, \ldots, K-1, K+i\}$, the node K is an isolated node, and all the other clusters identical to those given by Y_0.

Let $\mathbb{P}_i$ be the distribution of the graph A conditioned on $Y^* = Y_i$. Note that each $\mathbb{P}_i$ is the product of $\frac{1}{2}n(n-1)$ Bernoulli distributions. We have the following chain of inequalities:

$$I(Y^*; A) \overset{(a)}{\le} \frac{1}{(\bar{M}+1)^2}\sum_{i,i'=0}^{\bar{M}} D\left(\mathbb{P}_i \| \mathbb{P}_{i'}\right) \overset{(b)}{\le} 3K \cdot D\left(p\|q\right) + 3K \cdot D\left(q\|p\right),$$

where (a) follows from the convexity of KL divergence, and (b) follows by our construction of $\{Y_i\}$. If the condition (40) in the lemma holds, then $I(Y^*; A) \le \frac{1}{4}\log(n-K) = \frac{1}{4}\log\left|\bar{\mathcal{Y}}\right|$. Since $\log(n-K) \ge \log(n/2) \ge 4$ when $n \ge 128$, it follows from (41) that the minimax error probability is at least $1/2$. ■

Lemma 25 *Suppose* $128 \le K \le n/2$. *We have* $\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \frac{1}{4}$ *if*

$$Kp \le \frac{1}{8}\min\{\log(rK/2), K\}, \tag{42}$$

$$\textbf{\textit{or}} \quad K(1-q) \le \frac{1}{4}\log K. \tag{43}$$

Proof We first prove the sufficiency of the condition (42). We call a node a *disconnected node* if it is not connected to any other node in its own cluster. Let E be the event that there exist two disconnected nodes from two different clusters, and set $\rho := \mathbb{P}[E|Y^*]$, which is in fact independent of what value Y^* takes in $\mathcal{Y}$. Suppose Y^* is uniformly distributed over $\mathcal{Y}$; we claim that $\mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \rho/2$ for all $\widehat{Y}$. To see this, consider the maximum likelihood estimator (MLE) of Y^*, which is given by $\widehat{Y}_{\mathrm{ML}}(a) := \arg\max_y \mathbb{P}[A = a|Y^* = y]$ with tie broken uniformly at random. It is a standard fact that the MLE minimizes the error probability under the uniform prior, so for all $\widehat{Y}$ we have

$$\mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sum_{a \in \{0,1\}^{n \times n}} \mathbb{P}\left[\widehat{Y}_{\mathrm{ML}}(a) \neq y\right] \mathbb{P}\left[A = a | Y^* = y\right]. \tag{44}$$

Let $\mathcal{A}_y \subseteq \{0,1\}^{n \times n}$ denote the set of adjacency matrices with at least two disconnected nodes in two different clusters with respect to the clusters defined by $y \in \mathcal{Y}$. For each $a \in \mathcal{A}_y$, let $y'(a)$ denote the cluster matrix obtained by swapping the two rows and columns of y corresponding to the first two disconnected nodes in a. It is easy to check that for each $a \in \mathcal{A}_y$, the likelihood satisfies $\mathbb{P}[A = a|Y^* = y] \leq \mathbb{P}[A = a|Y^* = y'(a)]$ and therefore $\mathbb{P}[\widehat{Y}_{\mathrm{ML}}(a) \neq y] \geq 1/2$. It follows from (44) that

$$\mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \frac{1}{|\mathcal{Y}|} \sum_{y} \sum_{a \in \mathcal{A}_y} \frac{1}{2} \cdot \mathbb{P}[A = a|Y^* = y] = \frac{1}{2}\rho,$$

where the last equality holds because $\mathbb{P}[\mathcal{A}_y|Y^* = y] \equiv \rho$ independently of y. This proves our claim.

Since the maximum error probability is lower bounded by the average error probability, to establish the lemma it suffices to show $\rho \geq 1/2$. Without loss of generality, suppose r is even and we fix the clusters Y^* to be such that the first $rK/2$ nodes $\{1, \ldots, rK/2\}$ form $r/2$ clusters, and the next $rk/2$ nodes $\{rK/2 + 1, \ldots, rK\}$ form another $r/2$ clusters. For each $i \in [rK/2]$, let ξ_i be the indicator random variable for node i being a disconnected node. Then $\rho_1 := \mathbb{P}\left[\sum_{i=1}^{rK/2} \xi_i \geq 1\right]$ is the probability that there exits at least one disconnected node among the first $rK/2$ nodes. We use a second moment argument (Durrett, 2007) to lower-bound ρ_1. Observe that $\xi_1, \ldots, \xi_{rK/2}$ are (dependent) Bernoulli variables with mean $\mu := (1-p)^{K-1}$. For $i \neq j$, we have

$$\mathbb{E}\left[\xi_i \xi_j\right] = \mathbb{P}\left[\xi_i = 1, \xi_j = 1\right] \leq (1-p)^{2K-3} = \frac{1}{1-p}\mu^2.$$

Therefore, we obtain

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{rK/2} \xi_i\right] &\leq \frac{1}{2} rK\mu(1-\mu) + \frac{1}{2} rK(rK/2-1)\left(\frac{1}{1-p} - 1\right)\mu^2 \\
&\leq \frac{1}{2} rK\mu + \frac{1}{4} r^2 K^2 \mu^2 \frac{p}{1-p}.
\end{aligned}$$

Under the condition (42) we have $p \leq 1/8$ and

$$\mu = (1-p)^{K-1} \overset{(a)}{\geq} e^{-2(K-1)p} \geq (rK/2)^{-1/4}, \tag{45}$$

where (a) uses the inequality $1 - x \geq e^{-2x}, \forall x \in [0, \frac{1}{2}]$. Applying the Chebyshev's inequality, we get

$$\mathbb{P}\left[\left|\sum_{i=1}^{rK/2} \xi_i - rK\mu/2\right| \geq rK\mu/2\right] \leq \frac{\frac{1}{2}rK\mu + \frac{1}{4}(rK\mu)^2\frac{p}{1-p}}{r^2K^2\mu^2/4} \leq \frac{2}{rK\mu} + \frac{p}{1-p} \leq \frac{1}{4},$$

where the last inequality holds due to (45) and $p \leq 1/8$. It follows that $\rho_1 \geq \frac{3}{4}$. If we let ρ_2 denote the probability that there exits a disconnected node among the next $rK/2$ nodes $\{rK/2 + 1, \ldots, rK\}$, then by symmetry $\rho_2 \geq \frac{3}{4}$. Therefore $\rho \geq \rho_1\rho_2 \geq 1/2$, proving the sufficiency of (42).

We next prove the sufficiency of the condition (43) by bounding the error probability using a similar strategy. For $k = 1, 2$, we call a node in cluster k a *betrayed node* if it is connected to all nodes in cluster $3 - k$. Let E' be the event of having a betrayed node in each of the clusters 1 and 2, and let $\rho' := \mathbb{P}[E']$. Suppose Y^* is uniformly distributed over $\mathcal{Y}$; we can use a similar argument as above to show that $\mathbb{P}[\widehat{Y} \neq Y^*] \geq \rho'/2$ for any $\widehat{Y}$. Fix Y^* to be such that the clusters 1 and 2 are given by the nodes $[K]$ and $\{K+1, \ldots, 2K\}$, respectively. For each $i \in [K]$, let ξ_i' be the indicator for node i being is a betrayed node. Then $\rho_1' := \mathbb{P}\left[\sum_{i=1}^K \xi_i' > 0\right]$ is the probability of having a betrayed node in cluster 1. Note that the condition (43) implies $1 - q \leq 1/2$. We have

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^K \xi_i' = 0\right] = \left(1 - q^K\right)^K \leq \exp\left(-Kq^K\right) &\overset{(b)}{\leq} \exp\left[-K\exp\left(-2(1-q)K\right)\right] \\ &\overset{(c)}{\leq} \exp(-K^{1/2}) \leq 1/4,\end{aligned}$$

where (b) follows from the inequality $q^K = (1 - (1-q))^K \geq \exp\left(-2(1-q)K\right)$ for $1 - q \leq 1/2$, and (c) holds under the condition (43). We therefore obtain $\rho_1' \geq \frac{3}{4}$. Let ρ_2' be the probability of having a betrayed node in cluster 2; by symmetry $\rho_2' \geq 3/4$. By the union bound we get $\rho' \geq 1 - (1 - \rho_1') - (1 - \rho_2') \geq 1/2$, proving the sufficiency of (43). ■

We can now prove Theorem 2 by combining the above three lemmas.

Proof [of Theorem 2] Since $256 \leq 2K \leq n$, we have the following relations for the logarithmic terms:

$$\log(n - K) \geq \log(n/2) \geq \frac{1}{2}\log n, \quad \text{and} \quad \log(rK/2) \geq \frac{1}{2}\log(rK). \tag{46}$$

Our goal is to show that if the condition (1) or (2) holds, then we can draw the conclusion that the minimax error probability is large.

First assume (1) holds. By (36), we know the condition (1) implies

$$K(p-q)^2 \leq \frac{1}{96}p(1-q)\left(\log(rK) \wedge K\right). \tag{47}$$

We distinguish between two cases. (i) If $p \leq 2q$, then (47) implies $K(p-q)^2 \leq \frac{1}{48}q(1-q)\log(rK)$; it follows from (35) and (46) that $KD(p\|q) \leq \frac{1}{48}\log(rK) \leq \frac{1}{24}\log(n-K)$, and thus Lemma 24 proves the conclusion. (ii) If $p > 2q$, then (47) implies $Kp \leq \frac{1}{24}\log(rK) \wedge K \leq \min\{\frac{1}{24}K, \frac{1}{12}\log(\frac{rK}{2})\}$, and Lemma 25 proves the conclusion.

Next assume the condition (2) holds. Using the lower-bound (36) on the KL divergence, we know that (2) implies

$$K(p-q)^2 \le \frac{1}{96}p(1-q)\log n. \tag{48}$$

We distinguish between two cases. (i) If $1-q \le 2(1-p)$, then (48) implies $K(p-q)^2 \le \frac{1}{48}p(1-p)\log n$; it follows from (35) and (46) that $KD(q\|p) \le \frac{1}{48}\log n \le \frac{1}{24}\log(n-K)$, and thus Lemma 24 implies the conclusion. (ii) If $1-q > 2(1-p)$ and $K \ge \log n$, then (48) implies

$$K(1-q) \le \frac{1}{24}\log n \le \frac{1}{12}\max\left\{\log\frac{n}{K}, \log K\right\}. \tag{49}$$

We further divide the analysis into two sub-cases.

Case $(ii.1)$: $K \ge \log n$. It follows from (49) that $1-q \le \frac{1}{24}$, i.e., $q \ge \frac{23}{24}$, and thus $(p-q)^2 \le 2q(1-q)^2$. Therefore, the inequality (49) implies either the condition (38) in Lemma 23 or the condition (43) in Lemma 25, which proves the conclusion.

Case $(ii.2)$: $K < \log n$. It follows that $\delta := \frac{n_1(K-1)}{n(n-1)} \le \frac{1}{10}$ and $\log\frac{n}{K} \ge \frac{1}{2}\log n$. Note that $\bar{p} := \delta p + (1-\delta)q \ge \max\{\delta p, q\}$ and $1-\bar{p} \ge \frac{9}{10}(1-q)$. Therefore, we have

$$\frac{n^2(q-\bar{p})^2}{n_1\bar{p}(1-\bar{p})} = \frac{n^2\delta^2(p-q)^2}{n_1\bar{p}(1-\bar{p})} \le \frac{2n^2\delta(p-q)^2}{n_1 p(1-q)} \overset{(a)}{\le} 4KD(p\|q) \overset{(b)}{\le} \frac{1}{24}\log\frac{n}{K}, \tag{50}$$

where we use (36) in (a) and (2) in (b). On the other hand, we have

$$\begin{aligned} D(p\|\bar{p}) &= p\log\frac{p}{\bar{p}} + (1-p)\log\frac{1-p}{1-\bar{p}} \le p\log\frac{p}{q} + (1-p)\log\frac{10(1-p)}{9(1-q)} \\ &\le D(p\|q) + (1-q)\log\frac{10}{9} \le \frac{1}{6K}\log\frac{n}{K}, \end{aligned} \tag{51}$$

where the last inequality follows from (2) and (49). Equations (50) and (51) imply the assumption (37) in Lemma 23, and therefore the conclusion follows. ■

5.1.1 PROOF OF COROLLARY 3

The corollary is derived from Theorem 2 using the upper bound (35) on the KL divergence. In particular, condition (3) implies condition (2) in Theorem 2 in view of (35). Similarly, condition (4) implies condition (1) because $D(q\|p) \le \frac{p}{1-p}$ in view of (35) and $p \le \frac{1}{193}$. Finally, condition (5) implies condition (2) because $D(p\|q) \le p\log\frac{p}{q}$ by definition of the KL divergence.

5.2 Proof of Theorem 4 and Corollary 5

Let $\langle X, Y\rangle := \mathrm{Tr}(X^\top Y)$ denote the trace inner product between two matrices. For each feasible solution $Y \in \mathcal{Y}$ of the optimization problem (7), we define $\Delta(Y) := \langle A, Y^* - Y\rangle$ and $d(Y) := \langle Y^*, Y^* - Y\rangle$. To prove the theorem, it suffices to show that $\Delta(Y) > 0$ for all feasible Y with $Y \ne Y^*$. For simplicity, in this proof we use a different convention that $Y^*_{ii} = 0$ and $Y_{ii} = 0$ for all

$i \in V$. Note that $\mathbb{E}[A] = qJ + (p-q)Y^* - qI$, where J is the $n \times n$ all-one matrix and I is the $n \times n$ identity matrix. We may decompose $\Delta(Y)$ into an expectation term and a fluctuation term:

$$\Delta(Y) = \langle \mathbb{E}[A], Y^* - Y \rangle + \langle A - \mathbb{E}[A], Y^* - Y \rangle = (p-q)d(Y) + \langle A - \mathbb{E}[A], Y^* - Y \rangle, \quad (52)$$

where the second equality follows from $\sum_{i,j} Y_{ij} = \sum_{i,j} Y^*_{ij}$ by feasibility of Y. For the second fluctuation term above, observe that

$$\langle A - \mathbb{E}[A], Y^* - Y \rangle = 2 \underbrace{\sum_{\substack{(i<j): Y^*_{ij}=1 \\ Y_{ij}=0}} (A_{ij} - p)}_{T_1(Y)} - 2 \underbrace{\sum_{\substack{(i<j): Y^*_{ij}=0 \\ Y_{ij}=1}} (A_{ij} - q)}_{T_2(Y)}.$$

Here each of $T_1(Y)$ and $T_2(Y)$ is the sum of $\frac{1}{2}d(Y)$ i.i.d. centered Bernoulli random variables with parameter p and q, respectively.

Using the Chernoff bound, we can bound the fluctuation for each fixed $Y \in \mathcal{Y}$:

$$\mathbb{P}\left\{T_1(Y) \le -\frac{p-q}{4}d(Y)\right\} \le \exp\left[-\frac{1}{2}d(Y)D\left(\frac{p+q}{2}\middle\|p\right)\right], \quad (53)$$

$$\mathbb{P}\left\{T_2(Y) \ge \frac{p-q}{4}d(Y)\right\} \le \exp\left[-\frac{1}{2}d(Y)D\left(\frac{p+q}{2}\middle\|q\right)\right]. \quad (54)$$

We need to control the fluctuation uniformly over $Y \in \mathcal{Y}$. Define the equivalence class $[Y] := \{Y' \in \mathcal{Y} : Y'_{ij} = Y_{ij}, \forall (i,j) \in \text{support}(Y^*)\}$, where $\text{support}(Y^*) := \{(i,j) : Y^*_{ij} = 1\}$. Obverse that all cluster matrices in the equivalence class $[Y]$ have the same value of $T_1(Y)$. The following combinatorial lemma upper bounds the number of Y's and $[Y]$'s such that $d(Y) = t$. Note that $2(K-1) \le d(Y) \le rK^2$ for all feasible $Y \ne Y^*$.

Lemma 26 *For each integer $t \in [K, rK^2]$, we have*

$$\left|\{Y \in \mathcal{Y} : d(Y) = t\}\right| \le \left(\frac{16t^2}{K^2}\right)^2 n^{32t/K},$$

$$\left|\{[Y] : d(Y) = t\}\right| \le \frac{16t^2}{K^2}(rK)^{16t/K}.$$

We prove this lemma in Appendix A. We also need the following lemma, which lower-bounds $D\left(\frac{p+q}{2}\|q\right)$ and $D\left(\frac{p+q}{2}\|p\right)$ using $D(p\|q)$ and $D(q\|p)$, respectively. The proof is given in Appendix B.

Lemma 27 *For any $0 \le p \le q \le 1$, we have*

$$D\left(\frac{p+q}{2}\middle\|q\right) \ge \frac{1}{36}D(p\|q), \quad (55)$$

$$D\left(\frac{p+q}{2}\middle\|p\right) \ge \frac{1}{36}D(q\|p). \quad (56)$$

Using the union bound, (53), Lemma 26 and Lemma 27, we obtain

$$\begin{aligned}
&\mathbb{P}\left\{\exists[Y] : Y \neq Y^*, T_1(Y) \leq -\frac{p-q}{4}d(Y)\right\} \\
&\leq \sum_{t=K}^{rK^2} \mathbb{P}\left\{\exists[Y] : d(Y) = t, T_1(Y) \leq -\frac{p-q}{4}t\right\} \\
&\leq \sum_{t=K}^{rK^2} \left|\{\exists[Y] : d(Y) = t\}\right| \cdot \mathbb{P}\left\{T_1(Y) \leq -\frac{p-q}{4}t\right\} \\
&\leq \sum_{t=K}^{rK^2} \frac{16t^2}{K^2}(rK)^{16t/K} \exp\left(-\frac{1}{72}tD(q\|p)\right) \\
&\overset{(a)}{\leq} 16 \sum_{t=K}^{rK^2} (rK)^2(\gamma rK)^{-5t/K} \leq 16(\gamma rK)^{-1},
\end{aligned}$$

where (a) follows from the theorem assumption that $D(q\|p) \geq c_1 \log(\gamma rK)/K$ for a sufficiently large constant c_1. Similarly, using (54) we have

$$\begin{aligned}
&\mathbb{P}\left\{\exists Y \in \mathcal{Y} : Y \neq Y^*, T_2(Y) \geq \frac{p-q}{4}d(Y)\right\} \\
&\leq \sum_{t=K}^{rK^2} \mathbb{P}\left\{\exists Y \in \mathcal{Y} : d(Y) = t, T_2(Y) \geq \frac{p-q}{4}t\right\} \\
&\leq \sum_{t=K}^{rK^2} \left|\{Y \in \mathcal{Y} : d(Y) = t\}\right| \cdot \mathbb{P}\left\{T_2(Y) \geq \frac{p-q}{4}t\right\} \\
&\leq \sum_{t=K}^{rK^2} \frac{256t^4}{K^4} n^{32t/K} \cdot \exp\left(-\frac{1}{72}tD(p\|q)\right) \overset{(b)}{\leq} 256n^{-1},
\end{aligned}$$

where (b) follows from the theorem assumption that $D(p\|q) \geq c_1 \log n/K$ for a sufficiently large constant c_1. Combining the above two bounds with (52), we obtain

$$\mathbb{P}\left\{\exists Y \in \mathcal{Y} : \Delta(Y) \leq 0\right\} \leq 16(\gamma rK)^{-1} + 256n^{-1}.$$

Therefore Y^* is the unique optimal solution with the same probability. This proves the theorem.

5.2.1 PROOF OF COROLLARY 5

The corollary is derived from Theorem 4 using the lower bound (36) on the KL divergence. First assume $e^2q \geq p$. Then $K(p-q)^2 \gtrsim q(1-q)\log n$ implies condition (9) in view of (36). Next assume $e^2q < p$. It follows that $\log \frac{p}{q} \leq 2\log \frac{p}{eq}$. By definition, $D(p\|q) \geq p\log \frac{p}{q} + (1-p)\log(1-p) \geq p\log \frac{p}{eq}$. Hence, $Kp\log \frac{p}{q} \gtrsim \log n$ implies $KD(p\|q) \gtrsim \log n$. Furthermore, $D(q\|p) \geq \frac{1}{2}(1-1/e^2)p$ in view of (36) and $p > e^2q$. Therefore, $Kp \gtrsim \log(rK)$ implies $KD(q\|p) \gtrsim \log(rK)$.

5.3 Proof of Theorem 6

We prove Theorem 6 and Theorem 17 (for submatrix localization) together in this section. Our proof relies only on two standard concentration results for the adjacency matrix A (Proposition 28 below).

We need some unified notation for the two models. For both models we use n_L and n_R to denote the problem dimensions, with the understanding that $n_L = n_R = n$ for planted clustering. Similarly, for planted clustering the left and right clusters are identical and $K_L = K_R = K$. Let $U \in \mathbb{R}^{n_L \times r}$ and $V \in \mathbb{R}^{n_R \times r}$ be the normalized characteristic matrices of the left and right clusters, respectively:

$$U_{ik} = \begin{cases} \frac{1}{\sqrt{K_L}}, & \text{if the left node } i \text{ is in the } k\text{-th left cluster,} \\ 0, & \text{otherwise,} \end{cases}$$

and similarly for V. Here $U = V$ for planted clustering. The true cluster matrix Y^* has the rank-r Singular Value Decomposition given by $Y^* = \sqrt{K_L K_R} U V^\top$. Define the projections $\mathcal{P}_T(M) = UU^\top M + MVV^\top - UU^\top MVV^\top$ and $\mathcal{P}_{T^\perp}(M) = M - \mathcal{P}_T(M)$. Several matrix norms will be used: the spectral norm $\|X\|$ (the largest singular value of X), the nuclear norm $\|X\|_*$ (the sum of the singular values), the ℓ_1 norm $\|X\|_1 = \sum_{i,j} |X_{ij}|$ and the ℓ_∞ norm $\|X\|_\infty = \max_{i,j} |X_{ij}|$.

We define a quantity $\nu > 0$ and a matrix $\bar{A} \in \mathbb{R}^{n_L \times n_R}$, which roughly correspond to the signal strength and the mean of A. For planted clustering, let $\nu := p - q$ and $\bar{A} := qJ + (p-q)Y^*$, where J is the all-one matrix. For submatrix localization, let $\nu := \mu$ and $\bar{A} := \mu Y^*$. The proof hinges on the following probabilistic property of the random matrix $A - \bar{A}$.

Proposition 28 *Under the condition* (14) *for planted clustering,* or *the condition* (27) *for submatrix localization, the following holds with probability at least* $1 - n^{-10}$*:*

$$\|A - \bar{A}\| \le \frac{1}{8}\nu\sqrt{K_L K_R}, \tag{57}$$

$$\|\mathcal{P}_T(A - \bar{A})\|_\infty \le \frac{1}{8}\nu. \tag{58}$$

We prove the proposition in Section 5.3.1 to follow. In the rest of the proof we assume the event that (57) and (58) hold. To establish the theorems, it suffices to show that $\langle Y^* - Y, A\rangle > 0$ for all feasible solution Y of the convex program with $Y \neq Y^*$. For any feasible Y, we may write

$$\begin{aligned} \langle Y^* - Y, A\rangle &= \langle \bar{A}, Y^* - Y\rangle + \langle A - \bar{A}, Y^* - Y\rangle \\ &= \nu\langle Y^*, Y^* - Y\rangle + \langle A - \bar{A}, Y^* - Y\rangle = \frac{\nu}{2}\|Y^* - Y\|_1 + \langle A - \bar{A}, Y^* - Y\rangle, \end{aligned} \tag{59}$$

where the second equality follows from the definition of $\bar{A}$, and the third equality holds because Y obeys the linear constraints $\sum_{i,j} Y_{ij} = \sum_{i,j} Y^*_{ij}$ and $Y_{ij} \in [0,1], \forall i, j$.

On the other hand, we have $\|Y^*\|_* \ge \|Y\|_*$ thanks to the constraint (12) or (25). Let $W := \frac{8(A-\bar{A})}{\nu\sqrt{K_L K_R}}$. By (57) we have $\|\mathcal{P}_{T^\perp}(W)\| \le \|W\| \le 1$, so $UV^\top + \mathcal{P}_{T^\perp}(W)$ is a subgradient of $f(X) := \|X\|_*$ at $X = Y^*$ (cf. Recht et al. 2010 for characterization of the subgradient of the nuclear norm). It follows that

$$0 \ge \|Y\|_* - \|Y^*\|_* \ge \left\langle UV^\top + \mathcal{P}_{T^\perp}(W), Y - Y^*\right\rangle = \langle W, Y - Y^*\rangle + \left\langle UV^\top - \mathcal{P}_T(W), Y - Y^*\right\rangle.$$

Rearranging terms and using the definition of W gives

$$\left\langle A-\bar{A}, Y^*-Y\right\rangle = \frac{\nu\sqrt{K_LK_R}}{8}\left\langle W, Y^*-Y\right\rangle \geq \frac{\nu\sqrt{K_LK_R}}{8}\left\langle -UV^\top + \mathcal{P}_T(W), Y^*-Y\right\rangle. \quad (60)$$

Assembling (59) and (60), we obtain that for any feasible Y,

$$\begin{aligned}\langle Y^*-Y, A\rangle &\geq \frac{\nu}{2}\|Y^*-Y\|_1 + \frac{\nu\sqrt{K_LK_R}}{8}\left\langle -UV^\top + \mathcal{P}_T(W), Y^*-Y\right\rangle \\ &\geq \left(\frac{\nu}{2} - \frac{\nu\sqrt{K_LK_R}}{8}\|UV^\top\|_\infty - \|\mathcal{P}_T(A-\bar{A})\|_\infty\right)\|Y^*-Y\|_1,\end{aligned}$$

where the last inequality follows from the duality between the ℓ_1 and ℓ_∞ norms. Using (58) and the fact that $\|UV^\top\|_\infty = 1/\sqrt{K_LK_R}$, we get

$$\langle Y^*-Y, A\rangle \geq \left(\frac{\nu}{2} - \frac{\nu}{8} - \frac{\nu}{8}\right)\|Y^*-Y\|_1 = \frac{\nu}{4}\|Y^*-Y\|_1,$$

where the R.H.S. is strictly positive for all $Y \neq Y^*$. This completes the proof of Theorems 6 and 17.

5.3.1 PROOF OF PROPOSITION 28

We first prove (58). By definition of $\mathcal{P}_T$, we have

$$\begin{aligned}\|\mathcal{P}_T(A-\bar{A})\|_\infty \leq& \|UU^\top(A-\bar{A})\|_\infty + \|(A-\bar{A})VV^\top\|_\infty + \|UU^\top(A-\bar{A})VV^\top\|_\infty \\ \leq& 3\max\left(\|UU^\top(A-\bar{A})\|_\infty, \|(A-\bar{A})VV^\top\|_\infty\right). \quad (61)\end{aligned}$$

Suppose the left node i belongs to the left cluster k. Then

$$(UU^\top(A-\bar{A}))_{ij} = \frac{1}{K_L}\sum_{l\in C_k^*}(A-\bar{A})_{lj} = \frac{1}{K_L}\sum_{l\in C_k^*}(A-\mathbb{E}A)_{lj} + \frac{1}{K_L}\sum_{l\in C_k^*}(\mathbb{E}A-\bar{A})_{lj}. \quad (62)$$

To proceed, we consider the two models separately.

Planted clustering: The entries of the matrix $A-\mathbb{E}A$ are centered Bernoulli random variables with variance bounded by $p(1-q)$ and mutually independent up to symmetry with respect to the diagonal. The first term of (62) is the average of K_L such random variables; by Bernstein's inequality (stated as Theorem 32 in Appendix C), we have with probability at least $1-n^{-13}$ and for some universal constant c_2,

$$\left|\textstyle\sum_{l\in C_k^*}(A-\mathbb{E}A)_{lj}\right| \leq \sqrt{26p(1-q)K\log n} + 9\log n \leq c_2\sqrt{p(1-q)K\log n},$$

where the last inequality follows because $Kp(1-q) > c_1\log n$ in view of the condition (14). By definition of $\bar{A}$, $\mathbb{E}[A]-\bar{A}$ is a diagonal matrix with each diagonal entry equal to $-p$ or $-q$, so the second term of (62) has magnitude at most $1/K$. By the union bound over all (i,j) and substituting back to (61), we have with probability at least $1-2n^{-11}$,

$$\|\mathcal{P}_T(A-\bar{A})\|_\infty \leq 3c_2\sqrt{p(1-q)\log n/K} + 3/K \leq (p-q)/8 = \nu/8,$$

where the last inequality follows from the condition (14). This proves (58) in the proposition.

Submatrix localization: We have $\bar{A} = \mathbb{E}A$ by definition, so the second term of (62) is zero. The first term is the average of K_L independent centered random variables with unit sub-Gaussian norm. By a standard sub-Gaussian concentration inequality (e.g., Proposition 5.10 in Vershynin 2012), we have for some universal constant c_3 and with probability at least $1 - n^{-13}$,

$$\left|\sum_{l \in \mathcal{C}_k} (A - \mathbb{E}A)_{lj}\right| \leq c_3 \sqrt{K_L \log n}.$$

So $\|UU^\top (A - \mathbb{E}A)\|_\infty \leq c_3 \sqrt{\log n / K_L}$ with probability at least $1 - n^{-11}$ by the union bound. Similarly, $\|(A - \mathbb{E}A)VV^\top\|_\infty \leq c_2 \sqrt{\log n / K_R}$ with the same probability. Combining with (61) gives

$$\|\mathcal{P}_T(A - \bar{A})\|_\infty \leq \sqrt{\log n / \min\{K_L, K_R\}} \leq \nu/8 = \mu/8,$$

where the last inequality holds under the condition (27). This proves (58) in the proposition.

We now turn to (57) in the proposition, and again consider the two models separately.

- *Planted clustering:* Note that $\|A - \bar{A}\| \leq \|A - \mathbb{E}[A]\| + \|\bar{A} - \mathbb{E}[A]\| \leq \|A - \mathbb{E}[A]\| + 1$. Under the condition (14), $Kp(1-q) \geq c_1 \log n$. We bound the spectral norm term in the lemma below.

  Lemma 29 *If $Kp(1-q) \geq c_1 \log n$, then there exists some universal constant c_4 such that $\|A - \mathbb{E}[A]\| \leq c_4 \sqrt{p(1-q)K \log n + q(1-q)n}$ with probability at least $1 - n^{-10}$.*

  We prove the lemma in Section 5.3.2 to follow. Applying the lemma, we obtain

  $$\|A - \bar{A}\| \leq c_4 \sqrt{p(1-q)K \log n + q(1-q)n} + 1 \leq \frac{K(p-q)}{8} = \frac{K\nu}{8},$$

  where the second inequality holds under the condition (14).

- *Submatrix localization:* The matrix $A - \bar{A} = A - \mathbb{E}A$ has i.i.d. sub-Gaussian entries. Using a standard concentration bound for the spectral norm of such a matrix (e.g., Theorem 5.39 in Vershynin 2012), we get that for a universal constant c_5 and with probability at least $1 - n^{-10}$,

  $$\|A - \mathbb{E}A\| \leq c_5 \sqrt{n} \leq \frac{\mu}{8} \sqrt{K_L K_R} = \frac{\nu}{8} \sqrt{K_L K_R},$$

  where the second inequality holds under the condition (27).

5.3.2 PROOF OF LEMMA 29

Let $R := \text{support}(Y^*)$ and $\mathcal{P}_R(\cdot) : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ be the operator that sets the entries outside R to zero. Set $B_1 := \mathcal{P}_R(A - \mathbb{E}[A])$ and $B_2 := A - \mathbb{E}[A] - B_1$. Then B_1 is a block-diagonal symmetric matrix with r blocks of size $K \times K$ and its upper-triangular entries are independent with zero mean and variance bounded by $p(1-q)$. Applying the matrix Bernstein inequality in Tropp (2012) and using the assumption that $Kp(1-q) \geq c_1 \log n$ in the lemma, we get that there exits some universal constant c_6 such that $\|B_1\| \leq c_6 \sqrt{p(1-q)K \log n}$ with probability at least $1 - n^{-11}$.

On the other hand, B_2 is symmetric and its upper-triangular entries are independent centered Bernoulli random variables with variance bounded by $\sigma^2 := \max\{q(1-q), c_7 \log n / n\}$ for any

universal constant c_7. If $\sigma^2 \geq \frac{\log^7 n}{n}$, then Theorem 8.4 in Chatterjee (2014) implies that $\|B_2\| \leq 3\sigma\sqrt{n}$ with probability at least $1-n^{-11}$. If $c_7\frac{\log n}{n} \leq \sigma^2 \leq \frac{\log^7 n}{n}$ for a sufficiently large constant c_7, then Lemma 2 in Massoulié and Tomozei (2014) implies that $\|B_2\| \leq c_8\sigma\sqrt{n}$ with probability at least $1-n^{-11}$ for some universal constant c_8. (See Lemma 8 in Vu 2014 for a similar derivation.) Putting together, we conclude that with probability at least $1-2n^{-11}$,

$$\begin{aligned}\|A-\mathbb{E}[A]\| \leq \|B_1\| + \|B_2\| &\leq c_6\sqrt{p(1-q)K\log n} + c_8\max\{\sqrt{q(1-q)n}, \sqrt{\log n}\} \\ &\leq c_4\sqrt{p(1-q)K\log n + q(1-q)n},\end{aligned}$$

where the last inequality holds because $Kp(1-q) \geq c_1 \log n$ by assumption. This proves the lemma.

5.4 Proof of Theorem 8

We first claim that $K(p-q) \leq c_2\sqrt{Kp+qn}$ implies $K(p-q) \leq c_2\sqrt{2qn}$ under the assumption that $K \leq n/2$ and $qn \geq c_1 \log n$. In fact, if $Kp \leq qn$, then the claim trivially holds. If $Kp > qn$, then $q < Kp/n \leq p/2$. It follows that

$$Kp/2 < K(p-q) \leq c_2\sqrt{Kp+qn} \leq c_2\sqrt{2Kp}.$$

Therefore, we have $Kp < 8c_2^2$, which contradicts the assumption that $Kp > qn \geq c_1 \log n$. So $Kp > qn$ cannot hold. Consequently, it suffices to show that if $K(p-q) \leq c_2\sqrt{2qn}$, then Y^* is not an optimal solution. We do this by deriving a contradiction assuming the optimality of Y^*.

Let J be the $n \times n$ all-ones matrix. Let $\mathcal{R} := \text{support}(Y^*)$ and $\mathcal{A} := \text{support}(A)$. Recall the cluster characteristic matrix U and the projection $\mathcal{P}_T(M) = UU^\top M + MUU^\top - UU^\top MUU^\top$ defined in Section 5.3, and that $Y^* = KUU^\top$ is the SVD of Y^*. Consider the Lagrangian

$$L(Y;\lambda,\mu,F,G) := -\langle A, Y\rangle + \lambda\left(\|Y\|_* - \|Y^*\|_*\right) + \eta\left(\langle J, Y\rangle - rK^2\right) - \langle F, Y\rangle + \langle G, Y-J\rangle,$$

where $\lambda, \eta \in \mathbb{R}$ and $F, G \in \mathbb{R}^{n\times n}$ are the Lagrangian multipliers. Since $Y = \frac{rK^2}{n^2}J$ is strictly feasible, strong duality holds by Slater's condition. Therefore, if Y^* is an optimal solution, then there must exist some $F, G \in \mathbb{R}^{n\times n}$ and λ for which the KKT conditions hold:

$$\left.\begin{aligned} 0 \in \left.\frac{\partial L(Y;\lambda,\mu,F,G)}{\partial Y}\right|_{Y=Y^*}, \end{aligned}\right\} \text{Stationary condition}$$

$$\left.\begin{aligned} F_{ij} \geq 0, G_{ij} \geq 0, \forall (i,j), \\ \lambda \geq 0, \end{aligned}\right\} \text{Dual feasibility}$$

$$\left.\begin{aligned} F_{ij} = 0, \forall (i,j) \in \mathcal{R}, \\ G_{ij} = 0, \forall (i,j) \in \mathcal{R}^c. \end{aligned}\right\} \text{Complementary slackness}$$

Recall that $M \in \mathbb{R}^{n\times n}$ is a sub-gradient of $\|X\|_*$ at $X = Y^*$ if and only if $\mathcal{P}_T(M) = UU^\top$ and $\|M - \mathcal{P}_T(M)\| \leq 1$. Set $H = F - G$; then the KKT conditions imply that there exist some numbers $\lambda \geq 0$, $\eta \in \mathbb{R}$ and matrices W, H obeying

$$A - \lambda\left(UU^\top + W\right) - \eta J + H = 0; \tag{63}$$

$$\mathcal{P}_T W = 0; \qquad \|W\| \leq 1; \tag{64}$$

$$H_{ij} \leq 0, \ \forall (i,j) \in \mathcal{R}; \qquad H_{ij} \geq 0, \ \forall (i,j) \in \mathcal{R}^c. \tag{65}$$

Now observe that $UU^\top WUU^\top = 0$ by (64). We left and right multiply (63) by $UU^\top$ to obtain

$$\breve{A} - \lambda UU^\top - \eta J + \breve{H} = 0,$$

where for any $X \in \mathbb{R}^{n\times n}$, $\breve{X} := UU^\top XUU^\top$ is the matrix obtained by averaging each $K \times K$ block of X. Consider the last display equation on the entries in $\mathcal{R}$ and $\mathcal{R}^c$ respectively. Applying the Bernstein inequality (Theorem 32) on each entry $\breve{A}_{ij}$, we have with probability at least $1 - 2n^{-11}$,

$$p - \frac{\lambda}{K} - \eta + \breve{H}_{ij} \geq -\frac{c_3\sqrt{p(1-p)\log n}}{K} - \frac{c_4 \log n}{2K^2} \overset{(a)}{\geq} -\frac{\epsilon_0}{8}, \quad \forall (i,j) \in \mathcal{R}, \tag{66}$$

$$q - \eta + \breve{H}_{ij} \leq \frac{c_3\sqrt{q(1-q)\log n}}{K} + \frac{c_4 \log n}{2K^2} \overset{(b)}{\leq} \frac{\epsilon_0}{8}, \quad \forall (i,j) \in \mathcal{R}^c \tag{67}$$

for some universal constants $c_3, c_4 > 0$, where (a) and (b) follow from the assumption $K \geq c_1 \log n$ with a sufficiently large universal constant c_1. In the rest of the proof, we assume (66) and (67) hold. Using (65), we get that

$$\begin{aligned} \eta &\geq q - \frac{c_3\sqrt{q(1-q)\log n}}{K} - \frac{c_4 \log n}{2K^2} \geq q - \frac{\epsilon_0}{8}, \\ \eta &\leq p + \frac{c_3\sqrt{p(1-p)\log n}}{K} + \frac{c_4 \log n}{2K^2} - \frac{\lambda}{K} \leq p + \frac{\epsilon_0}{8} - \frac{\lambda}{K}. \end{aligned} \tag{68}$$

It follows that

$$\begin{aligned} \lambda &\leq K(p-q) + c_3(\sqrt{p(1-p)\log n} + \sqrt{q(1-q)\log n}) + \frac{c_4 \log n}{K} \\ &\leq 4\max\left\{K(p-q), c_3\sqrt{p(1-p)\log n}, c_3\sqrt{q(1-q)\log n}, \frac{c_4}{c_1}\right\}. \end{aligned} \tag{69}$$

On the other hand, the conditions (64) and (63) imply

$$\begin{aligned} \lambda^2 &= \left\|\lambda(UU^\top + W)\right\|^2 \geq \frac{1}{n}\left\|\lambda(UU^\top + W)\right\|_F^2 \\ &= \frac{1}{n}\left\|A - \eta J + H\right\|_F^2 \geq \frac{1}{n}\left\|A_{\mathcal{R}^c} - \eta J_{\mathcal{R}^c} + H_{\mathcal{R}^c}\right\|_F^2 \geq \frac{1}{n}\sum_{(i,j)\in\mathcal{R}^c}(1-\eta)^2 A_{ij}, \end{aligned}$$

where $X_{\mathcal{R}^c}$ denotes that matrix obtained from X by setting the entries outside $\mathcal{R}^c$ to zero. Using (68), $\lambda \geq 0$ and the assumption $p \leq 1 - \epsilon_0$, we obtain $\eta \leq 1 - \frac{7}{8}\epsilon_0$, and therefore

$$\lambda^2 \geq \frac{49}{64n}\epsilon_0^2 \sum_{(i,j)\in\mathcal{R}^c} A_{ij}. \tag{70}$$

Note that $\sum_{(i,j)\in\mathcal{R}^c} A_{ij}$ equals twice the sum of $\binom{n}{2} - r\binom{K}{2}$ i.i.d. Bernoulli random variables with parameter q. By the Chernoff bound of Binomial distributions and the assumption that $qn \geq c_1 \log n$, we have with probability at least $1 - n^{-11}$, $\sum_{(i<j)\in\mathcal{R}^c} A_{ij} \geq c_5 qn^2$ for some universal constant c_5. It follows from (70) that $\lambda^2 \geq \frac{1}{2}\epsilon_0^2 c_5 qn$. Combining with (69) and the assumption that $qn \geq c_1 \log n$, we conclude that with probability at least $1 - 3n^{-11}$, $K^2(p-q)^2 \geq \frac{1}{32}\epsilon^2 c_5 qn$. Choosing c_2 in the theorem assumption to be sufficiently small such that $2c_2^2 < \frac{1}{32}\epsilon^2 c_5$, we obtain $K(p-q) > c_2\sqrt{2qn}$, which leads to a contradiction. This completes the proof of the theorem.

5.5 Proof of Theorem 10

Define event

$$\mathcal{E}_1 = \left\{ \min_{i \in V_1} d_i > \frac{(p-q)K}{2} + qn \right\} \cap \left\{ \max_{i \in V_2} d_i < \frac{(p-q)K}{2} + qn \right\}.$$

Define $\mathcal{E}_2$ to be the event that $S_{ij} > \frac{(p-q)^2 K}{3} + 2Kpq + q^2 n$ for all nodes i, j from the same true cluster and $S_{ij} < \frac{(p-q)^2 K}{3} + 2Kpq + q^2 n$ for all pairs of nodes (i, j) from two different true clusters. On event $\mathcal{E}_1$, all nodes in V_1 are correctly declared to be non-isolated, and all nodes in V_2 are correctly declared to be isolated. One event $\mathcal{E}_1 \cap \mathcal{E}_2$, the counting algorithm correctly identifies all the true clusters. Hence, to prove the theorem, it suffices to show $\mathbb{P}\{\mathcal{E}_1^c\} \le 2n^{-1}$ and $\mathbb{P}\{\mathcal{E}_2^c\} \le 2n^{-1}$.

Let $\mathrm{Bin}(n, \alpha)$ denote the binomial distribution with n trials and success probability α. For each non-isolated node $i \in V_1$, its degree d_i is the sum of two independent binomial random variables distributed respectively as $\mathrm{Bin}(K-1, p)$ and $\mathrm{Bin}(n-K, q)$. For each isolated node $i \in V_2$, its degree d_i is distributed as $\mathrm{Bin}(n-1, q)$. It follows that $\mathbb{E}[d_i] = (n-1)q + (K-1)(p-q)$ if $i \in V_1$ and $\mathbb{E}[d_i] = (n-1)q$ if $i \in V_2$. Define $\sigma_1^2 := Kp(1-q) + nq(1-q)$, then $\mathrm{Var}[d_i] \le \sigma_1^2$ for all i. Set $t_1 := \frac{1}{2}(K-1)(p-q)$. Since $p - q \le p(1-q)$, it follows that $t_1 \le \sigma_1^2$. Hence, Bernstein inequality (Theorem 32) gives

$$\mathbb{P}\{|d_i - \mathbb{E}[d_i]| \ge t_1\} \le 2\exp\left(-\frac{t_1^2}{2\sigma_1^2 + 2t_1/3}\right) \le 2\exp\left(-\frac{(K-1)^2(p-q)^2}{12\sigma_1^2}\right) \le 2n^{-2},$$

where the last inequality follows from the assumption (17). By the union bound, we get that $\mathbb{P}\{\mathcal{E}_1^c\} \le 2n^{-1}$.

For two nodes i and j in the same true cluster, the number of their common neighbors S_{ij} is the sum of two independent binomial random variables distributed respectively as $\mathrm{Bin}(K-2, p^2)$ and $\mathrm{Bin}(n-K, q^2)$. Similarly, for two nodes i, j in two different true clusters, S_{ij} is the sum of two independent binomial variables $\mathrm{Bin}(2(K-1), pq)$ and $\mathrm{Bin}(n-2K, q^2)$. Hence, $\mathbb{E}[S_{ij}]$ equals $(K-2)p^2 + (n-K)q^2$ if i and j are in the same true cluster and $2(K-1)pq + (n-2K)q^2$ if they are in two different true clusters. The difference of the expectations in two cases equals $K(p-q)^2 - 2p(p-q)$. Let $\sigma_2^2 := 2Kp^2(1-q^2) + nq^2(1-q^2)$, then $\mathrm{Var}[S_{ij}] \le \sigma_2^2$. Set $t_2 := K(p-q)^2/3$. Since $p - q \le p(1-q)$, it follows that $t_2 \le \sigma_2^2$. Applying the Bernstein inequality (Theorem 32), we obtain that

$$\mathbb{P}\{|S_{ij} - \mathbb{E}[S_{ij}]| \ge t_2\} \le 2\exp\left(-\frac{t_2^2}{2\sigma_2^2 + 2t_2/3}\right) \le 2\exp\left(-\frac{K^2(p-q)^4}{24\sigma_2^2}\right) \le 2n^{-3},$$

where the last inequality follows from the assumption (18). By the union bound, we get that $\mathbb{P}\{\mathcal{E}_2^c\} \le 2n^{-1}$.

5.6 Proof of Theorem 12

For simplicity we assume K and n_2 are even numbers; the case where K or n_2 is odd can be proved similarly. We partition the non-isolated nodes V_1 into two equal-sized subsets V_{1+} and V_{1-} such that half of the nodes in each cluster are in V_{1+}. Similarly, the isolated nodes V_2 are partitioned into two equal-sized subsets V_{2+} and V_{2-}. The idea is to use the following large-deviation *lower* bound to the quantities d_i and S_{ij}.

Theorem 30 (Theorem 7.3.1 in Matoušek and Vondrák (2008)) *Let $X_1, \ldots, X_N$ be independent random variables such that $0 \le X_i \le 1$ for all i. Suppose $X = \sum_{i=1}^N X_i$ and $\sigma^2 := \sum_{i=1}^N Var[X_i] \ge 200$. Then for all $0 \le \tau \le \sigma^2/100$ and some universal constant $c_3 > 0$, we have*

$$\mathbb{P}\left[X \ge \mathbb{E}[X] + \tau\right] \ge c_3 e^{-\tau^2/(3\sigma^2)}.$$

The main hurdle is that the entries of the graph adjacency matrix A are not completely independent due to the symmetry of A, so we need to take into account the dependence among d_i's and the dependence among S_{ij}'s when applying Theorem 30.

We first argue that under the assumption (19), the simple counting algorithm fails to identify the isolated nodes with probability at least $1/4$. For each node i in $V_{1+} \cup V_{2+}$, let d_{i+} and d_{i-} be the numbers of its neighbors in $V_{1+} \cup V_{2+}$ and $V_{1-} \cup V_{2-}$, respectively, so its total degree is $d_i = d_{i+} + d_{i-}$. Let $\mathrm{Bin}(N, \alpha)$ denote the binomial distribution with N trials and success probability α. We consider the following two cases.

Case 1: $(Kp + (n-K)q)\log n_1 \ge nq \log n_2$. Recall that $n_1 = rK$ and $n_2 = n - n_1$. In this case, it follows from (19) that

$$(K-1)^2(p-q)^2 \le 2c_2(Kp + nq)\log n_1. \tag{71}$$

For each node $i \in V_{1+}$, the quantity d_{i-} is the sum of two independent Binomial random variables distributed as $\mathrm{Bin}(K/2, p)$ and $\mathrm{Bin}((n-K)/2, q)$, respectively. Define

$$\begin{aligned} t &:= (K-1)(p-q) + 2, \\ \gamma_d^- &:= \mathbb{E}[d_{i-}] - t = \frac{1}{2}nq + \frac{1}{2}K(p-q) - t, \\ \sigma_d^2 &:= \mathrm{Var}[d_{i-}] = \frac{1}{2}Kp(1-p) + \frac{1}{2}(n-K)q(1-q). \end{aligned}$$

By assumption of the theorem, we have $K \le n/2$, $q \le p \le 1 - c_0$ and thus $\sigma_d^2 \ge \frac{c_0}{4}(Kp + nq)$. Since $Kp^2 + nq^2 \ge c_1 \log n$ by assumption, it follows that $\sigma_d^2 \ge \frac{1}{4}c_0 c_1 \log n \ge 400$ by choosing the constant c_1 in the assumption sufficiently large. Furthermore, it follows from (71) that by choosing c_1 sufficiently large and c_2 sufficiently small, we have

$$\sigma_d^4 \ge \frac{1}{4}c_0(Kp+nq)\sigma_d^2 \ge \frac{c_0}{8c_2}\frac{(K-1)^2(p-q)^2}{\log n_1} \times \frac{1}{4}c_0 c_1 \log n \ge 200^2(K-1)^2(p-q)^2.$$

Moreover, we have shown that $\sigma_d^2 \ge 400$. Hence, we get that $\sigma_d^2 \ge 100t$. We can now apply Theorem 30 with (71) to get that

$$\mathbb{P}\left[d_{i-} \le \gamma_d^-\right] \ge c_3 \exp\left(-\frac{t^2}{3\sigma_d^2}\right) \ge c_3 n_1^{-c_4}$$

for some universal constant $c_4 > 0$ that can be made arbitrarily small by choosing c_2 in the assumption sufficiently small; the last inequality holds due to $\sigma_d^2 \ge \frac{c_0}{4}(Kp+nq)$, $\sigma_d^2 \ge 400$, and (71) . Let $i^* := \arg\min_{i \in V_{1+}} d_{i-}$. Since the random variables $\{d_{i-} : i \in V_{1+}\}$ are mutually independent, we have

$$\mathbb{P}\left[d_{i^*-} > \gamma_d^-\right] = \prod_{i \in V_{1+}} \mathbb{P}\left[d_{i-} > \gamma_d^-\right] \le (1 - c_3 n_1^{-c_4})^{n_1/2} \le \exp\left(-c_3 n_1^{1-c_4}/2\right) \le 1/4,$$

where the last equality follows from letting c_4 sufficiently small and n_1 sufficiently large. Furthermore, for each $i \in V_{1+}$, the quantity d_{i+} is the sum of two independent Binomial random variables distributed as $\text{Bin}(K/2-1, p)$ and $\text{Bin}((n-K)/2, q)$, respectively. Since the median of $\text{Bin}(N, \alpha)$ is at most $N\alpha + 1$, we know that with probability at least $1/2$, we have $d_{i+} \le \gamma_d^+ := nq/2 + K(p-q)/2 - p + 2$. Now observe that the two sets of random variables $\{d_{i+}, i \in V_{1+}\}$ and $\{d_{i-}, i \in V_{1+}\}$ are independent of each other, so d_{i+} is independent of i^* for each $i \in V_{1+}$. It follows that

$$\mathbb{P}\left[d_{i^*+} \le \gamma_d^+\right] = \sum_{i \in V_{1+}} \mathbb{P}\left[d_{i+} \le \gamma_d^+ | i^* = i\right] \mathbb{P}\left[i^* = i\right] = \sum_{i \in V_{1+}} \mathbb{P}\left[d_{i+} \le \gamma_d^+\right] \mathbb{P}\left[i^* = i\right] \ge \frac{1}{2}.$$

Combining the two display equations above with the union bound, we obtain that with probability at least $1/4$,

$$d_{i^*} = d_{i^*-} + d_{i^*+} \le \gamma_d^- + \gamma_d^+ = (n-1)q.$$

On this event the node i^* will be incorrectly declared as an isolated node.

Case 2: $(Kp + nq)\log n_1 \le nq \log n_2$. In this case we have $(K-1)^2(p-q)^2 \le 2c_2 nq \log n_2$ in view of (19). Set $i^* := \arg\max_{i \in V_{2+}} d_{i-}$. Following the same argument as in Case 1 and using the assumption $nq \ge c_1 \log n$, we can show that $d_{i^*} \ge nq + K(p-q)$ with probability at least $1/4$, and on this event node i^* will incorrectly be declared as a non-isolated node.

We next show that under the assumption (20), the simple algorithm fails to recover the clusters with probability at least $1/4$. For two nodes i and j in V_1, let S_{ij+} be the number of their common neighbors in $V_{1+} \cup V_{2+}$ and S_{ij-} the number of their common neighbors in $V_{1-} \cup V_{2-}$, so the total number of their common neighbors is $S_{ij} = S_{ij+} + S_{ij-}$.

For each pair of nodes (i, j) in V_{1+} from the same cluster, S_{ij-} is the sum of two independent Binomial random variables distributed as $\text{Bin}(K/2, p^2)$ and $\text{Bin}((n-K)/2, q^2)$, respectively. Define

$$\begin{aligned} t' &:= K(p-q)^2 + 4, \\ \gamma_S^- &:= \mathbb{E}[S_{ij-}] - t' = nq^2/2 + K(p^2 - q^2)/2 - t', \\ \sigma_S^2 &:= \text{Var}[S_{ij-}] = \frac{1}{2}Kp^2(1-p^2) + \frac{1}{2}(n-K)q^2(1-q^2). \end{aligned}$$

By assumption, we have $K \le n/2$, $q \le p \le 1 - c_0$ and hence $\sigma_S^2 \ge \frac{c_0}{4}(Kp^2 + nq^2)$. Since by assumption $Kp^2 + nq^2 \ge c_1 \log n$, it follows that $\sigma_S^2 \ge 200$ by choosing c_1 sufficiently large. Moreover, recall that condition (20) reads:

$$K^2(p-q)^4 < c_2(Kp^2 + nq^2)\log n_1.$$

Hence, by choosing the constant c_2 sufficiently small and c_1 sufficiently large, we have that $\sigma_S^2 \ge 100t'$. Theorem 30 with (20) then implies that

$$\mathbb{P}\left[S_{ij-} \le \gamma_S^-\right] \ge c_3 \exp\left(-\frac{(t')^2}{3\sigma_S^2}\right) \ge c_3 n_1^{-c_5},$$

where the universal constant $c_5 > 0$ can be made sufficiently small by choosing c_2 sufficiently small in (18); the last inequality holds due to $\sigma_S^2 \ge 200$, $\sigma_S^2 \ge \frac{c_0}{4}(Kp^2 + nq^2)$, and (20).

Without loss of generality, we may re-label the nodes such that $V_{1+} = \{1, 2, \ldots, n_1/2\}$ and for each $k = 1, \ldots, n_1/4$, the nodes $2k-1$ and $2k$ are in the same cluster. Note that the random variables $\{S_{(2k-1)(2k)-} : k = 1, 2, \ldots, n_1/4\}$ are mutually independent. Let $i^* := -1 + 2 \arg\min_{k=1,2,\ldots,n_1/4} S_{(2k-1)(2k)-}$ and $j^* := i^* + 1$. It follows that

$$\mathbb{P}\left[S_{i^*j^*-} \geq \gamma_S^-\right] \leq (1 - c_3 n_1^{-c_5})^{n_1/4} \leq \exp(-c_3 n_1^{1-c_5}/4) \leq 1/4.$$

Furthermore, notice that S_{ij+} is the sum of two independent Binomial random variables $\mathrm{Bin}(K/2-2, p^2)$ and $\mathrm{Bin}((n-K)/2, q^2)$. We use a median argument introduced in the first part of the proof to conclude that for all i, j, $S_{ij+} \leq \gamma_S^+ := nq^2/2 + K(p^2 - q^2)/2 - 2p^2 + 2$ with probability at least $1/2$. Since $\{S_{ij+}, i, j \in V_{1+}\}$ only depends on the edges between V_{1+} and $V_{1+} \cup V_{2+}$, and (i^*, j^*) only depends on the edges between V_{1+} and $V_{1-} \cup V_{2-}$, it follows that $\{S_{ij+}, i, j \in V_{1+}\}$ and (i^*, j^*) are independent of each other. Therefore, $S_{i^*j^*+} \leq \gamma_S^+$ with probability at least $1/2$. Applying the union bound, we get that with probability at least $1/4$,

$$S_{i^*j^*} = S_{i^*j^*-} + S_{i^*j^*+} \leq \gamma_S^- + \gamma_S^+ = 2(K-1)pq + (n-2K)q^2.$$

On this event the nodes i^*, j^* will be incorrectly assigned to two different clusters.

6. Proofs for Submatrix Localization

In this section we prove the theoretical results in Section 3 for submatrix localization. Recall that $n := \max\{n_L, n_R\}$.

6.1 Proof of Theorem 15

We prove the theorem using Fano's inequality. Our arguments extend those used in Kolar et al. (2011). Recall that $\mathcal{Y}$ is the set of all valid bi-clustering matrices. Let $M = n_R - K_R$ and $\bar{\mathcal{Y}} = \{Y_0, Y_1, \ldots, Y_M\}$ be a subset of $\mathcal{Y}$ with cardinality $M+1$, which is specified later. Let $\mathbb{P}_{(Y^*,A)}$ denote the joint distribution of (Y^*, A) when Y^* is sampled from $\bar{\mathcal{Y}}$ uniformly at random and then A is generated according to the submatrix localization model with the true cluster matrix being Y^*. The minimax error probability can be bounded using the average error probability and Fano's inequality:

$$\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \inf_{\widehat{Y}} \mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right] \geq 1 - \frac{I(Y^*; A) + 1}{\log |\bar{\mathcal{Y}}|}, \tag{72}$$

where the mutual information is defined under the distribution $\mathbb{P}_{(Y^*,A)}$.

We construct $\bar{\mathcal{Y}}$ as follows. Let Y_0 be the bi-clustering matrix such that the left clusters $\{C_k\}_{k=1}^r$ are $C_k = \{(k-1)K_L + 1, \ldots, kK_L\}$ and the right clusters $\{D_l\}_{l=1}^r$ are $D_l = \{(l-1)K_R + 1, \ldots, lK_R\}$. Informally, each Y_i with $i \geq 1$ is obtained from Y_0 by keeping the left clusters and swapping two right nodes in two different right clusters. More specifically, for each $i \in [M]$: (1) Y_i has the same left clusters as Y_0; (2) if the right node $K_R + i \in D_l$ for Y_0, then Y_i has the same right clusters as Y_0 except that the first right cluster is $\{1, 2, \ldots, K_R - 1, K_R + i\}$ and the l-th right cluster is $D_l \setminus \{K_R + i\} \cup \{K_R\}$; (3) if the right node $K_R + i$ does not belong to any D_l for Y_0, then Y_i has the same right clusters as Y_0 except that the first right cluster is $\{1, 2, \ldots, K_R - 1, K_R + i\}$ instead.

Let $\mathbb{P}_i$ be the distribution of A conditioned on $Y^* = Y_i$, and $D(\mathbb{P}_i \| \mathbb{P}_{i'})$ the KL divergence between $\mathbb{P}_i$ and $\mathbb{P}_{i'}$. Since each $\mathbb{P}_i$ is a product of $n_L \times n_R$ Gaussian distributions, we have

$$I(Y^*; A) \le \frac{1}{(M+1)^2} \sum_{i,i'=0}^{M} D(\mathbb{P}_i \| \mathbb{P}_{i'})$$
$$\le 3K_L \left[D\left(\mathcal{N}(\mu_1, \sigma^2) \| \mathcal{N}(\mu_2, \sigma^2)\right) + D\left(\mathcal{N}(\mu_2, \sigma^2) \| \mathcal{N}(\mu_1, \sigma^2)\right) \right] = 3K_L \frac{(\mu_1 - \mu_2)^2}{\sigma^2},$$

where we use the convexity of KL divergence in the first inequality, the definition of Y_i in the second inequality, and the expression for the KL divergence between two Gaussian distributions in the equality. If $(\mu_1 - \mu_2)^2 \le \frac{\sigma^2 \log(n_R - K_R)}{12K_L}$, then $I(Y; A) \le \frac{1}{2}\log(n_R - K_R) = \frac{1}{2}\log|\bar{\mathcal{Y}}|$. Since $\log(n_R - K_R) \ge \log(n_R/2) \ge 4$ if $n_R \ge 128$, it follows from (72) that the minimax error probability is at least $1/2$.

Alternatively, we can construct $Y_i, i \ge 1$ from Y_0 by keeping the right clusters and swapping two left nodes in two different left clusters. A similar argument shows that if $(\mu_1 - \mu_2)^2 \le \frac{\sigma^2 \log(n_L - K_L)}{12K_R}$, the minimax error probability is at least $1/2$.

6.2 Proof of Theorem 16

Recall that $\langle X, Y \rangle := \mathrm{Tr}(X^\top Y)$ is the inner product between two matrices. For any feasible solution $Y \in \mathcal{Y}$ of (22), we define $\Delta(Y) := \langle A, Y^* - Y \rangle$ and $d(Y) := \langle Y^*, Y^* - Y \rangle$. To prove the theorem, it suffices to show that $\Delta(Y) > 0$ for all feasible Y with $Y \ne Y^*$. We may write

$$\Delta(Y) = \langle \mathbb{E}[A], Y^* - Y \rangle + \langle A - \mathbb{E}[A], Y^* - Y \rangle = \mu d(Y) + \langle A - \mathbb{E}[A], Y^* - Y \rangle \tag{73}$$

since $\mathbb{E}[A] = \mu Y^*$. The second term above can be written as

$$\langle A - \mathbb{E}[A], Y^* - Y \rangle = \underbrace{\sum_{(i,j): Y^*_{ij}=1, Y_{ij}=0} (A_{ij} - \mu)}_{T_1(Y)} + \underbrace{\sum_{(i,j): Y^*_{ij}=0, Y_{ij}=1} (-A_{ij})}_{T_2(Y)}.$$

Here each of $T_1(Y)$ and $T_2(Y)$ is the sum of $d(Y)$ i.i.d. centered sub-Gaussian random variables with parameter 1. By the sub-Gaussian concentration inequality given in Proposition 5.10 in Vershynin (2012), we obtained that for each $i = 1, 2$ and each fixed $Y \in \mathcal{Y}$,

$$\mathbb{P}\left\{ T_i(Y) \le -\frac{\mu}{2} d(Y) \right\} \le e \exp\left(-C\mu^2 d(Y)\right),$$

where $C > 0$ is an absolute constant. Combining with the union bound and (73), we get

$$\mathbb{P}\left\{\Delta(Y) \le 0\right\} \le 2e \exp\left(-C\mu^2 d(Y)\right), \quad \text{for each } Y \in \mathcal{Y}. \tag{74}$$

Define the equivalence class $[Y] = \{Y' \in \mathcal{Y} : Y'_{ij} = Y_{ij}, \forall (i,j) \in \mathrm{support}(Y^*)$. The following combinatorial lemma (proved in Appendix A) upper-bounds the number of Y's and $[Y]$'s with a fixed value of $d(Y)$. Note that $K_L \wedge K_R \le d(Y) \le rK_LK_R$ for any feasible $Y \ne Y^*$.

Lemma 31 *For each integer $t \in [K_L \wedge K_R, rK_LK_R]$, we have*

$$\left|\{Y \in \mathcal{Y} : d(Y) = t\}\right| \leq \left(\frac{16t^2}{K_LK_R}\right)^2 n_L^{16t/K_R} n_R^{16t/K_L}, \tag{75}$$

$$\left|\{[Y] : d(Y) = t\}\right| \leq \frac{16t^2}{K_LK_R}(rK_L)^{8t/K_R}(rK_R)^{8t/K_L}. \tag{76}$$

Combining Lemma 31 with (74) and the union bound, we obtain

$$\begin{aligned}
&\mathbb{P}\left\{\exists Y \in \mathcal{Y} : Y \neq Y^*, \Delta(Y) \leq 0\right\} \\
\leq &\sum_{t=K_L\wedge K_R}^{rK_LK_R} \mathbb{P}\left\{\exists Y \in \mathcal{Y} : d(Y) = t, \Delta(Y) \leq 0\right\} \\
\leq &2e \sum_{t=K_L\wedge K_R}^{rK_LK_R} \left|\{Y \in \mathcal{Y} : d(Y) = t\}\right| \cdot \mathbb{P}\left\{d(Y) = t, \Delta(Y) \leq 0\right\} \\
\leq &2e \sum_{t=K_L\wedge K_R}^{rK_LK_R} \left(\frac{16t^2}{K_LK_R}\right)^2 n_L^{16t/K_R} n_R^{16t/K_L} \cdot \exp\left(-C\mu^2 t\right) \\
\overset{(a)}{\leq} &2e \sum_{t=K_L\wedge K_R}^{rK_LK_R} 256n^4 n^{-7t/(K_L\wedge K_R)} \leq 512eK_LK_Rrn^{-3} \leq 512en^{-1},
\end{aligned}$$

where (a) follows from the assumption that $\mu^2(K_L \wedge K_R) \geq C'\sigma^2 \log n$ for a sufficiently large constant C'. This means Y^* is the unique optimal solution to (22) with high probability.

6.3 Proof of Theorem 17

We have proved the theorem in Section 5.3.

6.4 Proof of Theorem 18

Note that the theorem assumes $n = n_L = n_R$ and $K = K_L = K_R$. Recall that J is the $n \times n$ all-one matrix, $\mathcal{R} := \text{support}(Y^*)$ and $\mathcal{A} := \text{support}(A)$, and $U, V \in \mathbb{R}^{n\times r}$ are the cluster characteristic matrices defined in Section 5.3, and $Y^* = KUV^\top$ is the SVD of Y^*. By relabeling the left nodes and right nodes, we can always make $U = V$ and thus we assume $U = V$ in the following proof.

Suppose Y^* is an optimal solution to the program. Then by the same argument used in the proof of Theorem 8, there must exist some $\lambda \geq 0$, η, W and H obeying the KKT conditions (63)–(65). Since $UU^\top WUU^\top = 0$ by (64), we can left and right multiply (63) by $UU^\top$ to obtain

$$\breve{A} - \lambda UU^\top - \eta J + \bar{H} = 0,$$

where for any matrix $X \in \mathbb{R}^{n\times n}$, we define the block-averaged matrix $\breve{X} := UU^\top XUU^\top$. Consider the last display equation on each entries in $\mathcal{R}$ and $\mathcal{R}^c$. By the Gaussian probability tail bound, there exists a universal constant $c_3 > 0$ such that with probability at least $1 - 2n^{-11}$,

$$\mu - \frac{\lambda}{K} - \eta + \bar{H}_{ij} \geq -\frac{c_3\sqrt{\log n}}{K}, \forall (i,j) \in \mathcal{R}, \tag{77}$$

$$-\eta + \bar{H}_{ij} \leq \frac{c_3\sqrt{\log n}}{K}, \forall (i,j) \in \mathcal{R}^c. \tag{78}$$

Combining the last two display equations with (65), we get that

$$-\frac{c_3\sqrt{\log n}}{K} \le \eta \le \mu + \frac{c_3\sqrt{\log n}}{K} - \frac{\lambda}{K}.$$

It follows that

$$\lambda \le K\mu + 2c_3\sqrt{\log n} \le 4\max\left\{K\mu, c_3\sqrt{\log n}\right\}. \tag{79}$$

Furthermore, due to (77), (78) and $\lambda \ge 0$, we have

$$\bar{H}_{ij} \le \mu + \frac{2c_3\sqrt{\log n}}{K} \le \mu + \frac{1}{40}, \forall (i,j) \in \mathcal{R}^c, \tag{80}$$

where the last inequality holds when $K \ge c_1 \log n$.

On the other hand, the conditions (64) and (63) imply that

$$\begin{aligned}\lambda^2 &= \left\|\lambda(UU^\top + W)\right\|^2 \ge \frac{1}{n}\left\|\lambda(UU^\top + W)\right\|_F^2 = \frac{1}{n}\|A - \eta J + H\|_F^2 \\ &= \frac{1}{n}\left(\|A_{\mathcal{R}} - \eta J_{\mathcal{R}} + H_{\mathcal{R}}\|_F^2 + \|A_{\mathcal{R}^c} - \eta J_{\mathcal{R}^c} + H_{\mathcal{R}^c}\|_F^2\right).\end{aligned} \tag{81}$$

We now lower bound the RHS of (81). For each (i,j), define the Bernoulli random variables $\bar{b}_{ij} = \mathbf{1}(A_{ij} - \mathbb{E}A_{ij} \ge 1)$ and $\underline{b}_{ij} = \mathbf{1}(A_{ij} - \mathbb{E}A_{ij} \le -1)$, where $\mathbf{1}(\cdot)$ is the indicator function. By tail bounds of the standard Gaussian distribution, we have

$$\mathbb{P}\left(\bar{b}_{ij} = 1\right) = \mathbb{P}\left(\underline{b}_{ij} = 1\right) \ge \rho := \frac{1}{2\sqrt{2\pi}}e^{-1/2}.$$

Note that $\rho \ge \frac{1}{12}$. By Hoeffding's inequality, we know that with probability at least $1 - 2n^{-11}$,

$$\sum_{i,j\in\mathcal{R}^c} \bar{b}_{ij} \ge \frac{1}{2}\rho|\mathcal{R}^c| \quad \text{and} \quad \sum_{i,j\in\mathcal{R}^c} \underline{b}_{ij} \ge \frac{1}{2}\rho|\mathcal{R}^c|. \tag{82}$$

We consider two cases below.

- Case 1: $\eta \ge 40\mu$. By (80) and the Markov inequality, there is at most a fraction of $\frac{1}{30}$ of pairs (i,j) in $\mathcal{R}^c$ that satisfy $H_{ij} > 30\left(\mu + \frac{1}{40}\right)$. Let $\mathcal{D}$ denote the set of pairs (i,j) satisfying both $H_{ij} \le 30\left(\mu + \frac{1}{40}\right)$ and $A_{ij} \le -1$. In view of the second inequality in (82), we have $|\mathcal{D}|/|\mathcal{R}^c| \ge \rho/2 - 1/30 \ge 1/150$. Therefore, for $(i,j) \in \mathcal{D}$, we get that $-\eta + H_{\mathcal{R}^c} \le -10\mu + \frac{3}{4}$, and thus

$$\begin{aligned}\|A_{\mathcal{R}^c} - \mu J_{\mathcal{R}^c} + H_{R^c}\|_F^2 &\ge \sum_{(i,j)\in\mathcal{D}} \|A_{\mathcal{R}^c} - \eta J_{\mathcal{R}^c} + H_{\mathcal{R}^c}\|_F^2 \\ &\ge \sum_{(i,j)\in\mathcal{D}} \left(-1 - 10\mu + \frac{3}{4}\right)^2 \ge \frac{1}{150}|\mathcal{R}^c| \cdot \frac{1}{16}.\end{aligned}$$

- Case 2: $\eta \le 40\mu$. Since $\mu \le \frac{1}{100}$ by assumption, we have $\eta \le 1/2$. Therefore,

$$\begin{aligned}\|A_{\mathcal{R}^c} - \eta J_{\mathcal{R}^c} + H_{R^c}\|_F^2 &\ge \sum_{(i,j)\in\mathcal{R}^c, \bar{b}_{ij}=1} \|A_{\mathcal{R}^c} - \eta J_{\mathcal{R}^c} + H_{\mathcal{R}^c}\|_F^2 \\ &\ge \sum_{(i,j)\in\mathcal{R}^c:\bar{b}_{ij}=1} (1-\eta)^2 \ge \frac{1}{2}\rho|\mathcal{R}^c| \cdot \frac{1}{4}.\end{aligned}$$

Combining the two cases and substituting into (81), we obtain $\lambda^2 \ge c_4|\mathcal{R}^c|/n$ for some constant $c_4 > 0$. Since $|\mathcal{R}^c| = n^2 - rK^2 \ge n(n-K) \ge n^2/2$, we have $\lambda^2 \ge c_4 n/2$. It follows from (79) that

$$\max\left\{K\mu, c_3\sqrt{\log n}\right\} \ge \frac{\sqrt{c_4}}{4\sqrt{2}}\sqrt{n}.$$

Since $n \ge K \ge c_1 \log n$ with a sufficiently large constant c_1, we must have $K\mu \ge \frac{\sqrt{c_4}}{4\sqrt{2}}\sqrt{n}$. This violates the condition (28) in the theorem statement if we choose the universal constant c_2 sufficiently small. We conclude from this contradiction that Y^* is not an optimal solution of the convex program.

6.5 Proof of Theorem 20

We prove that with high probability, each of the three steps of the simple thresholding algorithm succeeds and thus Y^* is exactly recovered.

We first show that the simple thresholding algorithm correctly identifies all isolated nodes. Recall that $d_i = \sum_{j=1}^{n_R} A_{ij}$ is the row sum corresponding to the left node i. Observe that $d_i - \mathbb{E}[d_i]$ is the sum of n_R independent centered sub-Gaussian random variables with parameter 1. Moreover, $\mathbb{E}[d_i] = K_R\mu$ if node i is non-isolated; otherwise, $\mathbb{E}[d_i] = 0$. By Proposition 5.10 in Vershynin (2012), there exists a universal constant $c_3 > 0$ such that

$$\mathbb{P}\{|d_i - \mathbb{E}[d_i]| \ge K_R\mu/2\} \le e\exp\left(-\frac{c_3 K_R^2\mu^2}{n_R}\right) \le en_L^{-2},$$

where the last inequality follows from the assumption (29) by choosing the universal constant c_1 sufficiently large. By the union bound, we have with probability at least $1 - en_L^{-1}$, $d_i > \mu K_R/2$ for all non-isolated left nodes i and $d_i < \mu K_R/2$ for all isolated left nodes i. On this event, all isolated left nodes are correctly identified in Step 1 of the algorithm. A similar argument shows that all isolated right nodes are correctly identified with probability at least $1 - en_R^{-1}$. We use E_1 to denote the event that all the left and right isolated nodes are identified by the algorithm.

We first show that the simple thresholding algorithm correctly identifies all the left and right clusters. Recall that $S_{ii'} = \sum_{j=1}^{n_R} A_{ij}A_{i'j}$ is the inner product of two rows of A corresponding to the left nodes i and i'. If the two left nodes i, i' are in the same cluster, then $\mathbb{E}[S_{ii'}] = K_R\mu^2$; otherwise $\mathbb{E}[S_{ii'}] = 0$. Moreover, $A_{ij}A_{i'j}$ is the product of two independent sub-Gaussian random variables. We use $\|X\|_{\psi_2}$ and $\|X\|_{\psi_1}$ to denote the sub-Gaussian norm and sub-exponential norm[4]

4. The sub-exponential norm and sub-Gaussian norm of a random variable X are defined as $\|X\|_{\psi_i} = \sup_{p\ge 1} p^{-1/i}\left(\mathbb{E}|X|^p\right)^{1/p}$ for $i = 1, 2$, respectively (Vershynin, 2012).

It follows that

$$\begin{aligned}
&\left\|A_{ij}A_{i'j} - \mathbb{E}[A_{ij}]\mathbb{E}[A_{i'j}]\right\|_{\psi_1} \\
&\overset{(a)}{\leq} \left\|(A_{ij} - \mathbb{E}[A_{ij}])(A_{i'j} - \mathbb{E}[A_{i'j}])\right\|_{\psi_1} + \left\|(A_{ij} - \mathbb{E}[A_{ij}])\mathbb{E}[A_{i'j}]\right\|_{\psi_1} + \left\|(A_{ij} - \mathbb{E}[A_{ij}])\mathbb{E}[A_{i'j}]\right\|_{\psi_1} \\
&\overset{(b)}{\leq} 2\left\|A_{ij} - \mathbb{E}[A_{ij}]\right\|_{\psi_2}\left\|A_{i'j} - \mathbb{E}[A_{i'j}]\right\|_{\psi_2} + 2\mu\left\|A_{ij} - \mathbb{E}[A_{ij}]\right\|_{\psi_2} + 2\mu\left\|A_{i'j} - \mathbb{E}[A_{i'j}]\right\|_{\psi_2} \\
&\overset{(c)}{\leq} c'(4\mu + 2),
\end{aligned}$$

where (a) and (b) follow from $\|X + Y\|_{\psi_1} \leq \|X\|_{\psi_1} + \|Y\|_{\psi_1}$ and $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$ for any random variables X, Y, and (c) holds for some universal constant $c' > 0$ because for each (i, j) $A_{ij} - \mathbb{E}[A_{ij}]$ is sub-Gaussian with parameter 1. By the Bernstein inequality for sub-exponential random variables given in Proposition 5.16 in Vershynin (2012), there exists some universal constant $c_4 > 0$ such that

$$\mathbb{P}\left\{|S_{ii'} - \mathbb{E}[S_{ii'}]| \geq K_R\mu^2/2\right\} \leq e\exp\left[-c_4\min\left(\frac{K_R^2\mu^4}{n_R c'^2(4\mu+2)^2}, \frac{K_R\mu^2}{c'(4\mu+2)}\right)\right] \leq e(rK_L)^{-3},$$

where the last inequality follows from the conditions (30) and (29). By the union bound, we get that with probability at least $1 - e(rK_L)^{-1}$, $S_{ii'} > \frac{\mu^2 K_R}{2}$ for all left nodes i, i' from the same left cluster and $S_{ii'} < \frac{\mu^2 K_R}{2}$ for all left nodes i, i' from two different left clusters. On the intersection of this event and the event E_1 defined above, Step 2 of the algorithm identifies the true left clusters. A similar argument shows that the algorithm also identifies the true right clusters with probability at least $1 - e(rK_R)^{-1}$. We use E_2 to denote that event that all the true left and right clusters are identified by the algorithm.

Finally, we show that the simple thresholding algorithm correctly associates all the left and right clusters. Let $B^*_{kl} := \sum_{i\in C^*_k, j\in D^*_l} A_{ij}$ be the block sum of A corresponding to the true let and right clusters C^*_k and D^*_l. By model assumptions, $B^*_{kl} - \mathbb{E}[B^*_{kl}]$ is a sum of $K_L K_R$ independent centered sub-Gaussian random variables with parameter 1. Moreover, $\mathbb{E}[B^*_{kl}] = \mu K_L K_R$ if $k = l$, and $\mathbb{E}[B^*_{kl}] = 0$ otherwise. By the standard sub-Gaussian concentration inequality given in Proposition 5.10 in Vershynin (2012), there exists some universal constant $c_5 > 0$ such that

$$\mathbb{P}\{|B^*_{kl} - \mathbb{E}[B^*_{kl}]| \geq \mu K_L K_R/2\} \leq e\exp\left(-\frac{c_5\mu^2 K_L^2 K_R^2}{K_L K_R}\right) \leq en^{-3},$$

where the last inequality holds because $\mu^2 K_L K_R \geq c_1 \log n$ in view of (29). By the union bound, we get that with probability at least $1 - en^{-1}$, $B^*_{kl} < \mu K_L K_R/2$ for all $k = l$ and $B^*_{kl} > \mu K_L K_R/2$ for all $k \neq l$. On the intersection of this event and the event E_2defined above, the quantities $\{B_{kl}\}$ used in Step 3 of the algorithm satisfy $B_{kl} = B^*_{kl}$, and the algorithm correctly associates the left and right clusters.

6.6 Proof of Theorem 21

We focus on identifying left isolated nodes and left clusters. The proof for the right nodes is identical. The main idea is to show that some of the d_i and $S_{ii'}$'s will have large deviation from their expectations.

Assume $rK_L \geq n_L/2$ first. We will show that if $K_R^2\mu^2 \leq c_1 n_R \log n_L$ for a sufficiently small universal constant c_1, then with high probability there exists a non-isolated left node i^* that is incorrectly declared as isolated. Recall that $d_i = \sum_{j=1}^{n_R} A_{ij}$ is the row sum corresponding to the left node i. If the left node i is non-isolated, then d_i is Gaussian with mean $K_R\mu$ and variance n_R. For a standard Gaussian random variable Z, its tail probability is lower bounded as $Q(t) := \mathbb{P}[Z \geq t] \geq \frac{1}{\sqrt{2\pi}}\frac{t}{t^2+1}\exp(-t^2/2)$. It follows that for a non-isolated left node i, there exist two positive universal constants c_3, c_4 such that

$$\mathbb{P}\left[d_i - \mathbb{E}[d_i] \leq K_R\mu/2\right] \geq c_3 \exp\left(-\frac{c_4 K_R^2\mu^2}{n_R}\right) \geq c_3 n_L^{-c_1c_4}.$$

Let i^* be the non-isolated left node with the minimum d_i. Since $\{d_i\}_{i=1}^{n_L}$ are mutually independent, we have

$$\mathbb{P}\left[d_{i^*} > \frac{K_R\mu}{2}\right] \leq \left(1 - c_3 n_L^{-c_1c_4}\right)^{rK_L} \leq \exp\left(-\frac{1}{2}c_3 n_L^{1-c_1c_4}\right),$$

where the last inequality holds because $rK_L \geq n_L/2$. By choosing c_1 sufficiently small, we conclude that with high probability the non-isolated left node i^* will be incorrectly declared as an isolated node.

If $rK_L \leq n_L/2$, then we can similarly show that if $K_R^2\mu^2 \leq c_1 n_R \log n_L$ for a sufficiently small c_1, then with high probability there exists an isolated left node i^{**} incorrectly declared as non-isolated.

We next show that if

$$K_R^2\mu^4 \leq c_2 n_R \log(rK_L) \tag{83}$$

for a sufficiently small constant c_2, then there exist two left nodes i_1, i_2 in two different clusters that will be incorrectly assigned to the same cluster.

By the assumption that $n_R = \Omega(rK_L)$, the inequality (83) implies $K_R\mu^3 \leq c_2^{3/4} n_R$. Recall that $S_{ii'} = \sum_{j=1}^{n_R} A_{ij}A_{i'j}$. For two left nodes i, i' from two different clusters, we have

$$\mathbb{E}[S_{ii'}] = 0, \quad \text{Var}[S_{ii'}] = 2K_R\mu^2 + n_R,$$
$$\sum_{j=1}^{n_R} \mathbb{E}[|A_{ij}A_{i'j}|^3] = \sum_{j=1}^{n_R} \mathbb{E}[|A_{ij}|^3]\mathbb{E}[|A_{i'j}|^3] \leq c_5(K_R\mu^3 + n_R) \leq c_5(c_2^{3/4}+1)n_R,$$

where c_5 is some universal positive constant. By the Berry-Esseen theorem, there exists a positive universal constant c_6 such that

$$\begin{aligned}
\mathbb{P}\left[S_{ii'} \geq \frac{\mu^2 K_R}{2}\right] &\geq Q\left(\frac{\mu^2 K_R}{2\sqrt{2K_R\mu^2 + n_R}}\right) - \frac{c_6(K_R\mu^3 + n_R)}{(2K_R\mu^2 + n_R)^{3/2}} \\
&\overset{(a)}{\geq} Q\left(\frac{\mu^2 K_R}{\sqrt{n_R}}\right) - \frac{c_6 c_5(c_2^{3/4}+1)}{\sqrt{n_R}} \\
&\overset{(b)}{\geq} Q\left(\sqrt{c_2 \log(rK_L)}\right) - \frac{c_6 c_5(c_2^{3/4}+1)}{\sqrt{rK_L}} \\
&\overset{(c)}{\geq} c_3(rK_L)^{-c_4c_2} - c_6c_5(c_2^{3/4}+1)(rK_L)^{-1/2} \overset{(d)}{\geq} c_7(rK_L)^{-c_4c_2},
\end{aligned}$$

where (a) holds because $Q(t)$ is non-increasing in t, (b) holds in view of (83) and the assumption that $n_R \geq rK_L$, (c) follows from $Q(t) \geq c_3 \exp(-c_4 t^2)$, and (d) holds for some universal constant $c_7 > 0$ by choosing c_2 sufficiently small.

Set $(i_1, i_2) := \arg\max_{(i,i') \in W} S_{ii'}$, where W is a maximal set of node pairs (i, i') satisfying 1) i and i' are from two different clusters, and 2) for any $(i, i'), (j, j') \in W$, the nodes i, i', j, j' are all distinct. Then $|W| \geq rK_L/4$ and $\{S_{ii'} : (i, i') \in W\}$ are mutually independent. It follows that

$$\mathbb{P}\left[S_{i_1 i_2} < \frac{\mu^2 K_R}{2}\right] \leq \left(1 - c_7 (rK_L)^{-c_4 c_2}\right)^{rK_L/4} \leq \exp\left(-\frac{1}{4} c_7 (rK_L)^{1-c_4 c_2}\right).$$

Therefore, with probability at least $1 - \exp\left(-\frac{1}{4} c_7 (rK_L)^{1-c_4 c_2}\right)$, we have $S_{i_1 i_2} \geq \frac{\mu^2 K_R}{2}$. On this event, (i_1, i_2) will be incorrectly assigned to the same cluster.

6.7 Proof of Theorem 22

We prove the first part of the theorem. Since A_{ij} are sub-Gaussian, there exists a universal constant $c_1 > 0$ such that $\mathbb{P}\left[|A_{ij} - \mathbb{E}A_{ij}| \leq \frac{1}{2}\sqrt{c_1 \log n}\right] \geq 1 - n^{-12}$ for each (i, j). Recall that $\mathcal{R} :=$ support(Y^*). By the union bound over all (i, j), we obtain that with probability at least $1 - n^{-3}$,

$$\min_{(i,j) \in \mathcal{R}} A_{ij} > \mu - \frac{1}{2}\sqrt{c_1 \log n} \overset{(a)}{>} \frac{1}{2}\mu \qquad \text{and} \qquad \max_{(i,j) \in \mathcal{R}^c} A_{ij} < \frac{1}{2}\sqrt{c_1 \log n} \overset{(b)}{<} \frac{1}{2}\mu,$$

where (a) and (b) holds in view of the assumption (33). Therefore, the algorithm sets $\widehat{Y}_{ij} = 1$ for $(i, j) \in \mathcal{R}$ and $\widehat{Y}_{ij} = 0$ for $(i, j) \in \mathcal{R}^c$, which implies $\widehat{Y} = Y^*$.

For the second part of the theorem, note that $\{A_{ij}\}$ are Gaussian variables and thus

$$\mathbb{P}\left[A_{ij} \geq \mathbb{E}A_{ij} + \sqrt{2 \log n}\right] = Q\left(\sqrt{2 \log n}\right) \geq \frac{\sqrt{2 \log n}}{\sqrt{2\pi}(2 \log n + 1)n},$$
$$\mathbb{P}\left[A_{ij} \leq \mathbb{E}A_{ij} - \sqrt{2 \log n}\right] = Q\left(\sqrt{2 \log n}\right) \geq \frac{\sqrt{2 \log n}}{\sqrt{2\pi}(2 \log n + 1)n},$$

where the last inequality holds due to $Q(t) \geq \frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} \exp(-t^2/2)$. By independence of the entries of A, we obtain

$$\mathbb{P}\left[\max_{i,j \in \mathcal{R}^c} A_{ij} < \sqrt{2 \log n}\right] \leq \left(1 - \frac{\sqrt{2 \log n}}{\sqrt{2\pi}(2 \log n + 1)n}\right)^{|\mathcal{R}^c|} \leq \exp\left(-\frac{\sqrt{2 \log n}|\mathcal{R}^c|}{\sqrt{2\pi}(2 \log n + 1)n}\right),$$
$$\mathbb{P}\left[\min_{i,j \in \mathcal{R}} A_{ij} > \mu - \sqrt{2 \log n}\right] \leq \left(1 - \frac{\sqrt{2 \log n}}{\sqrt{2\pi}(2 \log n + 1)n}\right)^{|\mathcal{R}|} \leq \exp\left(-\frac{\sqrt{2 \log n}|\mathcal{R}|}{\sqrt{2\pi}(2 \log n + 1)n}\right),$$

Since $|\mathcal{R}| + |\mathcal{R}^c| = n^2$, we must have

$$\min\left\{\mathbb{P}\left[\max_{i,j \in \mathcal{R}^c} A_{ij} < \sqrt{2 \log n}\right], \mathbb{P}\left[\min_{i,j \in \mathcal{R}} A_{ij} > \mu - \sqrt{2 \log n}\right]\right\} \leq e^{-\Omega(n/\sqrt{\log n})}.$$

This inequality, together with the assumption (34), implies that either with probability at least $1 - e^{-\Omega(n/\sqrt{\log n})}$, at least one of the following must occur: $\max_{(i,j) \in \mathcal{R}^c} A_{ij} \geq \sqrt{2 \log n} \geq \frac{3}{5}\mu$ and

$\min_{(i,j)\in R} A_{ij} \le \mu - \sqrt{2\log n} \le \frac{2}{5}\mu$. On this event, the output of the element-wise thresholding algorithm satisfies $\widehat{Y} \neq Y^*$.

Acknowledgments

The authors would like to thank Sivaraman Balakrishnan, Bruce Hajek and Martin J. Wainwright for inspiring discussions. Y. Chen was supported in part by NSF grant CIF-31712-23800, ONR MURI grant N00014-11-1-0688, and a start-up fund from the School of Operations Research and Information Engineering at Cornell University. J. Xu was supported in part by the National Science Foundation under Grant ECCS 10-28464, IIS-1447879, and CCF-1423088, and Strategic Research Initiative on Big-Data Analytics of the College of Engineering at the University of Illinois, and DOD ONR Grant N00014-14-1-0823, and Grant 328025 from the Simons Foundation.

Appendix A. Proof of Lemmas 26 and 31

Notice that Lemma 26 is a special case of Lemma 31 with $n_L = n_R$, $K_L = K_R$ and the left clusters identical to the right clusters. Hence we only need to prove Lemma 31.

Recall that $C_1^*, \ldots, C_r^*$ ($D_1^*, \ldots, D_r^*$, resp.) denote the true left (right, resp.) clusters associated with Y^*. The nodes in $V_L \setminus (\cup_{k=1}^r C_k^*)$ do not belong to any left clusters and are called isolated left nodes. Isolated right nodes are similarly defined.

Fix a $Y \in \mathcal{Y}$ with $d(Y) := \langle Y^*, Y - Y^* \rangle = t$. Based on Y, we construct a new ordered partition $(C_1, \ldots, C_{r+1})$ of V_L and a new ordered partition $(D_1, \ldots, D_{r+1})$ of V_R as follows.

1. Let $C_{r+1} := \{i : Y_{ij} = 0, \forall j\}$ and $D_{r+1} := \{j : Y_{ij} = 0, \forall i\}$.

2. The left nodes in $V_L \setminus C_{r+1}$ are further partitioned into r new left clusters of size K_L, such that the left nodes i and i' are in the same cluster if and only if the i-th and i'-th rows of Y are identical. Similarly, the right nodes in $V_R \setminus D_{r+1}$ are partitioned into r new right clusters of size K_R according to the columns of Y. We now define an ordering $C_1, \ldots, C_r$ of these r new left clusters and an ordering $D_1, \ldots, D_r$ for the new right clusters using the following procedure.

   (a) For each new left cluster C, if there exists a $k \in [r]$ such that $|C \cap C_k^*| > K_L/2$, then we label this new left cluster as C_k; this label is unique because the left cluster size is K_L. We associate C_k with the right cluster $\{j : Y_{ij} = 1, \forall i \in C_k\}$, which is labeled as D_k.

   (b) For each remaining unlabeled right cluster D, if there exists a $k \in [r]$ such that $|D \cap D_k^*| > K_R/2$, then we label this new right cluster as D_k; again this label is unique. We associate D_k with the left cluster $\{i : Y_{ij} = 1, \forall j \in D_k\}$, which is labeled as C_k.

   (c) The remaining unlabeled left clusters are labeled arbitrarily. For each remaining unlabeled right cluster, we label it according to $D_k := \{j : Y_{ij} = 1, \forall i \in C_k\}$.

For each $(k, k') \in [r] \times [r+1]$, we use $\alpha_{kk'} := |C_k^* \cap C_{k'}|$ and $\beta_{kk'} := |D_k^* \cap D_{k'}|$ to denote the sizes of intersections of the true and new clusters. We observe that the new clusters $(C_1, \ldots, C_{r+1}, D_1, \ldots, D_{r+1})$ have the following three properties:

(A0) $(C_1, \ldots, C_r, C_{r+1})$ is a partition of V_L with $|C_k| = K_L$ for each $k \in [r]$; $(D_1, \ldots, D_r, D_{r+1})$ is a partition of V_R with $|D_k| = K_R$ for each $k \in [r]$.

(A1) For each $k \in [r]$, exactly one of the following is true: (1) $\alpha_{kk} > K_L/2$; (2) $\alpha_{kk'} \le K_L/2$ for all $k' \in [r]$ and $\beta_{kk} > K_R/2$; (3) $\alpha_{kk'} \le K_L/2$ and $\beta_{kk'} \le K_R/2$ for all $k' \in [r]$.

(A2) We have

$$\sum_{k=1}^{r} \left(\alpha_{k(r+1)}\beta_{k(r+1)} + \sum_{k',k'':k' \neq k''} \alpha_{kk'}\beta_{kk''} \right) = t;$$

here and henceforth, all summations involving k' or k'' (as the indices of the new clusters) are over the range $[r+1]$ unless defined otherwise.

Here, Property (A0) holds due to $Y \in \mathcal{Y}$, Property (A1) is direct consequence of how we label the new clusters, and Property (A2) follows from the following:

$$\begin{aligned} t = d(Y) &= \sum_{k=1}^{r} |\{(i,j) : (i,j) \in C_k^* \times D_k^*, Y_{ij} = 0\}| \\ &= \sum_{k=1}^{r} |\{(i,j) : (i,j) \in C_k^* \times D_k^*, (i,j) \in C_{r+1} \times D_{r+1}\}| \\ &\quad + \sum_{k=1}^{r} \sum_{(k',k''):k' \neq k''} |\{(i,j) : (i,j) \in C_k^* \times D_k^*, (i,j) \in C_{k'} \times D_{k''}\}|. \end{aligned}$$

Since a different Y corresponds to a different ordered partition, and the ordered partition for any given Y with $d(Y) = t$ must satisfy the above three properties, we obtain the following bound on the cardinality of the set of interest:

$$|\{Y \in \mathcal{Y} : d(Y) = t\}| \le |\{(C_1, \ldots, C_{r+1}, D_1, \ldots, D_{r+1}) : \text{it satisfies (A0)–(A2)}\}|. \tag{84}$$

It remains to upper-bound the right hand side of (84).

Fix any ordered partition $(C_1, \ldots, C_r, C_{r+1}, D_1, \ldots, D_r, D_{r+1})$ with Properties (A0)–(A2). Consider the first true left cluster C_1^*. Define $m_1^{(L)} := \sum_{k':k' \neq 1} \alpha_{1k'}$, which can be considered as the number of nodes in C_1^* that are misclassified by Y. Analogously define $m_1^{(R)} := \sum_{k'':k'' \neq 1} \beta_{1k''}$. We consider the following two cases for the values of α_{11}.

- If $\alpha_{11} > K_L/4$, then

$$\sum_{(k',k''):k' \neq k''} \alpha_{1k'}\beta_{1k''} \ge \alpha_{11} \sum_{k'':k'' \neq 1} \beta_{1k''} > \frac{1}{4} m_1^{(R)} K_L.$$

- If $\alpha_{11} \le K_L/4$, then $m_1^{(L)} \ge 3K_L/4$, and we must also have $\alpha_{1k'} \le K_L/2$ for all $1 \le k' \le r$ by Property (A1). Hence,

$$\begin{aligned}
&\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\beta_{1k''} + \alpha_{1(r+1)}\beta_{1(r+1)} \\
\ge &\sum_{(k',k''):k'\neq k''} \mathbf{1}\left\{k' \neq 1\right\}\mathbf{1}\left\{k'' \neq 1\right\}\alpha_{1k'}\beta_{1k''} + \alpha_{1(r+1)}\beta_{1(r+1)} \\
= &m_1^{(L)}m_1^{(R)} - \sum_{2\le k'\le r} \alpha_{1k'}\beta_{1k'} \ge m_1^{(L)}m_1^{(R)} - \frac{1}{2}K_L m_1^{(R)} \ge \frac{1}{4}m_1^{(R)}K_L.
\end{aligned}$$

  Similarly, we consider the following three cases for the values of β_{11}.

- If $\beta_{11} > K_R/4$, then

$$\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\beta_{1k''} \ge \beta_{11}\sum_{k':k'\neq 1}\alpha_{1k'} > \frac{1}{4}m_1^{(L)}K_R.$$

- If $\beta_{11} \le K_R/4$ and $\beta_{1k''} \le K_L/2$ for all $1 < k'' \le r$, then, similarly to the second case for α_{11} above, we have

$$\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\beta_{1k''} + \alpha_{1(r+1)}\beta_{1(r+1)} \ge \frac{1}{4}m_1^{(L)}K_R.$$

- If $\beta_{11} \le K_R/4$ and $\beta_{1k_0} > K_R/2$ for some $1 < k_0 \le r$, then by Property (A1) we must have $\alpha_{11} > K_L/2$. It follows that $m_1^{(L)} < K_L/2$ and

$$\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\beta_{1k''} \ge \alpha_{11}\beta_{1k_0} > K_L K_R/4 \ge \frac{1}{2}m_1^{(L)}K_R.$$

Combining the above five cases, we conclude that the following is always true:

$$\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\beta_{1k''} + \alpha_{1(r+1)}\beta_{1(r+1)} \ge \frac{1}{4}\left(m_1^{(L)}K_R \vee m_1^{(R)}K_L\right).$$

This inequality continues to hold if we replace $\alpha_{1k'}$, $\beta_{1k''}$, $m_1^{(L)}$ and $m_1^{(R)}$ respectively by $\alpha_{kk'}$, $\beta_{kk''}$, $m_k^{(L)}$ and $m_k^{(R)}$ (defined in a similar manner) for each $k \in [r]$. Summing these inequalities over $k \in [r]$ and using Property (A2), we obtain

$$t = \sum_{k=1}^{r}\left\{\alpha_{k(r+1)}\beta_{k(r+1)} + \sum_{(k',k''):k'\neq k''}\alpha_{kk'}\beta_{kk''}\right\} \ge \left(\frac{K_L}{4}\sum_{k=1}^{r}m_k^{(R)}\right) \vee \left(\frac{K_R}{4}\sum_{k=1}^{r}m_k^{(L)}\right).$$

Consequently, we have $\sum_{k\in[r]} m_k^{(L)} \le 4t/K_R$ and $\sum_{k\in[r]} m_k^{(R)} \le 4t/K_L$; i.e., the total number of misclassified non-isolated left (right, resp.) nodes is upper bounded by $4t/K_R$ ($4t/K_L$, resp.). This means that the total number of misclassified isolated left (right, resp.) nodes is also upper bounded

by $4t/K_R$ ($4t/K_L$, resp.), because by the cluster size constraint in Property (A0), one misclassified isolated node must produce one misclassified non-isolated node.

We can now upper-bound the right hand side of (84) using the above relation between the value of t and the misclassified nodes. For each Y with $d(Y) = t$, the pair

$$(\text{\#misclassified isolated left nodes}, \quad \text{\#misclassified non-isolated left nodes})$$

can take at most $(4t/K_R)^2$ different values. Similarly for the right nodes we have the bound $(4t/K_L)^2$. Given these numbers of misclassified nodes, there are at most $n_L^{8t/K_R} n_R^{8t/K_L}$ different ways to choose the identity of these misclassified nodes. Each misclassified non-isolated left node can then be assigned to one of $r - 1 \le n_L$ different left clusters or left isolated, and each misclassified isolated left node can be assigned to one of $r \le n_L$ different left clusters; an analogous statement holds for the right nodes. Hence, the right hand side of (84) is upper bounded by $\left(\frac{16t^2}{K_L K_R}\right)^2 n_L^{16t/K_R} n_R^{16t/K_L}$. This proves the first part of the lemma.

To count the number of possible equivalence classes $[Y]$, we use a similar argument but only need to consider the misclassified *non-isolated* nodes. The number of misclassified non-isolated left (right, resp.) nodes can take at most $4t/K_R$ ($4t/K_L$, resp.) different values. Given these numbers, there are at most $(rK_L)^{4t/K_R}(rK_R)^{4t/K_L}$ different ways to choose the identity of the misclassified non-isolated nodes. Each misclassified non-isolated left (right, resp.) node can then be assigned to one of $r - 1$ different left (right, resp.) clusters or left isolated. Therefore, the number of possible equivalence classes $[Y]$ with $d(Y) = t$ is upper bounded by $\frac{16t^2}{K_L K_R}(rK_L)^{8t/K_R}(rK_R)^{8t/K_L}$.

Appendix B. Proof of Lemma 27

The inequality (56) follows from (55) by replacing p with $1 - q'$ and q with $1 - p'$, so it suffices to prove (55). Note that for $1 \ge u \ge v \ge 0$, we have

$$D(u\|v) = u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v} \le u \log \frac{u}{v}, \tag{85}$$

$$D(u\|v) \ge u \log \frac{u}{v} + (1-u) \log(1-u) \overset{(a)}{\ge} u \log \frac{u}{ev}, \tag{86}$$

where (a) follows from the inequality $x \log x \ge x - 1, \forall x \in [0,1]$. We consider two cases:

Case 1: $p \le 8q$. In view of (35) and (36), we have $D(p\|q) \le \frac{(p-q)^2}{q(1-q)}$ and $D\left(\frac{p+q}{2}\|q\right) \ge \frac{(p-q)^2}{4(p+q)(1-q)}$. Since $p \le 8q$, it follows that $D\left(\frac{p+q}{2}\|q\right) \ge \frac{(p-q)^2}{36q(1-q)} \ge \frac{1}{36} D(p\|q)$.

Case 2: $p > 8q$. In view of (85) and (86), we have $D(p\|q) \le p \log \frac{p}{q}$ and $D\left(\frac{p+q}{2}\|q\right) \ge \frac{p+q}{2} \log \frac{p+q}{2eq}$. Since $p > 8q$ and $8 > (2e)^{(6/5)}$, we have $\log \frac{p}{q} > \frac{6}{5} \log(2e)$ and thus $\log \frac{p+q}{2eq} > \log \frac{p}{2eq} = \log \frac{p}{q} - \log(2e) > \frac{1}{6} \log \frac{p}{q}$. It follows that $D\left(\frac{p+q}{2}\|q\right) \ge \frac{p}{2} \cdot \frac{1}{6} \log \frac{p}{q} \ge \frac{1}{12} D(p\|q)$.

Appendix C. The Bernstein Inequality

Theorem 32 (Bernstein) *Let $X_1, \ldots, X_N$ be independent random variables such that $|X_i| \le M$ almost surely. Let $\sigma^2 = \sum_{i=1}^N Var(X_i)$. Then for any $t \ge 0$,*

$$\mathbb{P}\left[\sum_{i=1}^N X_i \ge t\right] \le \exp\left(\frac{-t^2}{2\sigma^2 + \frac{2}{3}Mt}\right).$$

A consequent of the above inequality is $\mathbb{P}\left[\sum_{i=1}^{N} X_i \geq \sqrt{2\sigma^2 u} + \frac{2Mu}{3}\right] \leq e^{-u}$ *for any* $u > 0$.

References

E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv:1503.00609*, 2015.

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv:1405.3267*, 2014.

N. Ailon, Y. Chen, and H. Xu. Breaking the small cluster barrier of graph clustering. In *Proceedings of the 30th International Conference on Machine Learning*, pages 995–1003, 2013.

N. Alon and N. Kahale. A spectral technique for coloring random 3-colorable graphs. *SIAM Journal on Computing*, 26(6):1733–1748, 1997.

N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.

N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505. ACM, 2007.

B. P. W. Ames. Guaranteed clustering and biclustering via semidefinite programming. *Mathematical Programming*, pages 1–37, 2013.

B. P. W. Ames and S. A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011.

B. P. W. Ames and S.A. Vavasis. Convex optimization for the planted k-disjoint-clique problem. *Mathematical Programming*, 143(1–2):299–337, 2014.

A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5):2877–2921, 2009.

A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor spectral approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15:2239–2312, June 2014.

E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 06 2014.

E. Arias-Castro, E. J. Candès, and A. Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011a.

S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems 25*, 2011b.

A. S. Bandeira. Random Laplacian matrices and convex relaxations. *arXiv:1504.03987*, 2015.

N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1):89–113, 2004.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 30: 1046–1066, 2013.

S. Bhamidi, P. S. Dey, and A. B. Nobel. Energy landscape for large average submatrix detection problems in Gaussian random matrices. *arXiv:1211.2284*, 2012.

P. J. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

B. Bollobás and A. D. Scott. Max cut for random graphs with a planted partition. *Combinatorics, Probability and Computing*, 13(4-5):451–474, 2004.

C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

T. T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 06 2015.

V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.

K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 35.1–35.23, 2012.

Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61 (5):2909–2923, 2015.

Y. Chen and J. Xu. Statistical-computational phase transitions in planted models: The high-dimensional setting. In *Proceedings of the 31st International Conference on Machine Learning*, pages 244–252, 2014.

Y. Chen, S. Sanghavi, and H. Xu. Clustering sparse graphs. In *Proceedings of the Neural Information Processing Systems Conferece*, pages 2204–2212, 2012.

Y. Chen, A. Jalali, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, 15:2213–2238, June 2014a.

Y. Chen, S. Sanghavi, and H. Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014b.

A. Coja-Oghlan. Coloring semirandom graphs optimally. *Automata, Languages and Programming*, pages 383–395, 2004.

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

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

Y. Dekel, O. Gurel-Gurevich, and Y. Peres. Finding hidden cliques in linear time with high probability. *Combinatorics, Probability and Computing*, 23(01):29–49, 2014.

Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, pages 1–60, September 2013.

R. Durrett. *Random Graph Dynamics*. Cambridge University Press, New York, NY, 2007.

M. E. Dyer and A. M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451 – 489, 1989.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pages 655–664. ACM, 2013.

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

G. R. Grimmett and C. J. H. McDiarmid. On colouring random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 77(2):313–324, 1975.

B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv:1412.6156*, 2014.

B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv:1502.07738*, 2015.

B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. *arXiv:1406.6625. The conference version appeared in Proceedings of COLT 2015*, June, 2014.

E. Hazan and R. Krauthgamer. How hard is it to approximate the best Nash equilibrium? *SIAM Journal on Computing*, 40(1):79–91, 2011.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

A. Jalali and N. Srebro. Clustering using max-norm constrained optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 481–488, 2012.

A. Juels and M. Peinado. Hiding cliques for cryptographic security. *Designs, Codes and Cryptography*, 20(3):269–280, 2000.

P. Koiran and A. Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE Transactions on Information Theory*, 60(8):4999–5006, 2014.

M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, 2011.

R. Krauthgamer, B. Nadler, and D. Vilenchik. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 06 2015.

L. Kučera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57 (2–3):193–212, Feb. 1995.

M. Lelarge, L. Massoulié, and J. Xu. Reconstruction in the labeled stochastic block model. In *IEEE Information Theory Workshop (ITW)*, pages 1–5, 2013.

J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

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

L. Massoulié and D. Tomozei. Distributed user profiling via spectral methods. *Stochastic Systems*, 4(1):1–43, 2014.

C. Mathieu and W. Schudy. Correlation clustering with noisy input. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 712–728. SIAM, 2010.

J. Matoušek and J. Vondrák. The probabilistic method, lecture notes. *Available at http://kam.mff.cuni.cz/~matousek/prob-ln-2pp.ps.gz*, 2008.

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

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arxiv:1311.4115*, 2013.

E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015a. ISSN 0178-8051.

E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 69–75, New York, NY, USA, 2015b. ACM.

R. R. Nadakuditi and M. E. J. Newman. Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108(18):188–701, 2012.

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

S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5): 2886–2908, 2015.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(471), 2010.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

B. Rossman. *Average-case complexity of detecting cliques*. PhD thesis, Massachusetts Institute of Technology, 2010.

A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, 3(3):985–1012, 2009.

X. Sun and A. B. Nobel. On the maximal size of large-average and ANOVA-fit submatrices in a gaussian random matrix. *Bernoulli*, 19(1):275–294, 2013.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012.

N. Verzelen and E. Arias-Castro. Community detection in sparse random networks. *arXiv:1308.2955*, 2013.

R. K. Vinayak, S. Oymak, and B. Hassibi. Sharp performance bounds for graph clustering via convex optimization. In *38th International Conference on Acoustics, Speech, and Signal Processing*, 2014.

V. H. Vu. A simple SVD algorithm for finding hidden partitions. *arXiv:1404.3918*, 2014.

V. Q. Vu, J. Cho, J. Lei, and K. Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*, pages 2670–2678, 2013.

J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying. Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs. In *SIGMETRICS*, pages 29–41, 2014.

S. Yun and A. Proutiere. Community detection via random and adaptive sampling. In *Proceedings of the 27th Conference on Learning Theory*, 2014.