STATISTICAL INFERENCE IN NETWORKS: FUNDAMENTAL LIMITS AND EFFICIENT ALGORITHMS

BY

JIAMING XU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Bruce Hajek, Chair
Professor R. Srikant
Assistant Professor Sewoong Oh
Assistant Professor Sujay Sanghavi, University of Texas
Marc Lelarge, École Normal Supérieure

ABSTRACT

Today witnesses an explosion of data coming from various types of networks such as online social networks and biological networks. The goal of this thesis is to understand when and how we can efficiently extract useful information from such network data.

In the first part, we are interested in finding tight-knit communities within a network. Assuming the network is generated according to a planted cluster model, we derive a computationally efficient semidefinite programming relaxation of the maximum likelihood estimation method and obtain a stronger performance guarantee than previously known. If the community sizes are linear in the total number of vertices, the guarantee matches up to a constant factor with the information limit which we also identify, and exactly matches without a constant gap when there is a single community or two equal-sized communities. However, if the community sizes are sublinear in the total number of vertices, the guarantee is far from the information limit. We conjecture that our algorithm achieves the computational limit below which no polynomial-time algorithm can succeed. To provide evidence, we show that finding a community in some regime below the conjectured computational limit but above the information limit is computationally intractable, assuming hardness of the well-known planted clique problem.

The second part studies the problem of inferring the group preference for a set of items based on the partial rankings over different subsets of the items provided by a group of users. A question of particular interest is how to optimally construct the graph used for assigning items to users for ranking. Assuming the partial rankings are generated independently according to the Plackett-Luce model, we analyze computationally efficient estimators based on maximum likelihood and rank-breaking schemes that decompose partial rankings into pairwise comparisons. We provide upper and lower bounds on the estimation error. The lower bound depends on the degree sequence of the assignment graph, while the upper bound depends on the spectral gap of the assignment graph. When the graph is an

expander, the lower and upper bounds match up to a logarithmic factor.

The unifying theme for the two parts of the thesis is the spectral gap of the graph. In both cases, when the graph has a large spectral gap, accurate and efficient inference is possible via maximum likelihood estimation or its convex relaxation. However, when the spectral gap vanishes, accurate inference may be statistically impossible, or it is statistically possible but may be computationally intractable.

To my parents, brother and sister-in-law, and my wife, for their love and support.

ACKNOWLEDGMENTS

This thesis could not have been done without the help from many people. I owe my deepest gratitude to my advisor Professor Hajek. I can say without exaggeration that every piece of my ideas in this thesis originates from the long weekly meetings with him. He is always ready to spend hours and hours tackling the problems together with me. Meanwhile, he also gives me so much freedom that I can develop my own interests sometimes far from the conventional research areas in EE. His way of conducting research and living has changed me in every aspect of my life: I started running, biking, swimming, and enjoying research.

I would like to thank Dr. Laurent Massoulié and Dr. Marc Lelarge, with whom I worked during a summer internship at Technicolor Paris Research Lab. They introduced me to the fascinating topic of community detection, and our joint work planted the initial seeds of the thesis.

I also need to thank my dear friend and wonderful collaborator, Yudong Chen. I learned from him almost everything I know about semidefinite programming. I still remember the night during Allerton 2012 when we had a long discussion on community detection. We suspected there might exist a gap between the information and computational limits. This turns out to be the most important idea of the thesis.

I also thank Professor Yihong Wu. Our collaboration starts from two very similar plots of the phase transition diagram: My plot of the information and computational limits for community detection, and his plot of the same two limits for submatrix detection. It turns out that our collaboration has led to the most fundamentally important result in this thesis: The gap does exist conditional on the well-known planted clique hardness. What I have learned from him is not just a solution to a problem, but the courage to solve a problem no matter how daunting it looks.

I have been fortunate to interact with many professors during my three and a half years at UIUC, and two years at UT-Austin. Professor Srikant taught me game

theory and queueing theory, two of my favorite topics. The fruitful collaboration with Professor Oh led to the second part of the thesis on ranking aggregation. I first learned network science from Professor Shakkottai's class and immediately was attracted by the phase transitions in random graphs. Professor Sanghavi introduced high-dimensional statistical inference to me. I am also grateful to my advisor at UT, Professor Andrews. His understanding and encouragement gave me the courage to make the transition to a new environment. I also acknowledge Professor de Veciana for his support when I was at UT.

I am equally indebted to my friends and colleagues who made my research and my extracurricular life much richer. In particular, I thank Rui Wu for numerous discussions in the printer room which initiated our projects on recommender systems, Qiaomin Xie for training for and running a full marathon together, and Yuxin Chen for many stimulating discussions.

Finally, I would like to thank my family. My mother, father and brother built for me a home without worries so that I could focus on my studies. Most importantly, I am lucky to have my wife Lili by my side in the final sprint to the finish.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ML	Maximum Likelihood
SDP	Semidefinite Programming
PC	Planted Clique
PC-D	Planted Clique Detection
PDS	Planted Dense Subgraph
PBDS	Planted Binomial-sized Dense Subgraph
PDS-R	Planted Dense Subgraph Recovery
PBDS-D	Planted Binomial-sized Dense Subgraph Detection
SBM	Stochastic Block Model
DCSBM	Degree-corrected Stochastic Block Model
PL	Plackett-Luce Model
BT	Bradley-Terry Model
KL	Kullback-Leibler
CR	Cramér-Rao

CHAPTER 1

INTRODUCTION

1.1 From Data Networks to Network Data

Communication engineers have traditionally been interested in designing efficient information transmission schemes in data networks. For instance, in cellular networks, wireless transmissions can interfere with each other; a key question is how to model, counter, or even exploit the interference to increase the transmission rate. In queueing networks, servers may not be able to serve all arriving packets due to the stochastic nature of the arrival processes and therefore packets may suffer from significant queueing delay; a central question is how to design efficient routing and scheduling algorithms to reduce queueing delay. After decades of effort from both academia and industry, many such questions are now well understood and there has been a wide range of technological advance making data transmission faster and cheaper.

As a result, today we see a surge of online social networks such as Facebook, Twitter and Bitcoin, which generate enormous network data. Also, with the advent of high-throughput measurement methods in biology, a large amount of biological network data has accumulated, such as gene expression data from microarrays and protein-protein interaction networks from mass spectrometry. There are many other sources of network data such as transportation networks, power networks, and sensor networks. This network data contains a wealth of information and it is often desirable to extract useful information from it for various reasons, for instance to predict user preferences, discover disease causes or predict traffic patterns. However, this network data is so noisy and voluminous that the inference is both statistically difficult and computationally challenging. The main goal of this thesis is to show *when* and *how* it is computationally feasible to infer useful information from network data. A secondary goal is to understand how much the theory developed for the transmission of information, notably information theory,

can be applied to solve the inference problems in network data.

1.2 Two Typical Examples of Inference in Networks

Various inference problems can emerge in the context of network data, but most of them fall into one of the following two types. For the first type, we only observe the network structure. The problem of finding subgraphs with special patterns within a network is an example of the first type. For the second type, the observations are associated with nodes or edges. The problem of finding the source of a rumor with knowledge only about who has the rumor and the network topology is an example of the second type [1]. In this thesis, we study the following two questions, which represent the above two types, respectively.

Finding communities within networks Real networks often exhibit community structures with many edges joining the vertices of the same community and relatively few edges joining vertices of different communities. With only knowledge of the network topology, the problem is to identify the underlying tight-knit communities, which is known as the community detection problem. Assuming dense connections within communities and loose connections across different communities, finding communities essentially amounts to partitioning the graph into disjoint parts of given sizes with a minimum number of edges across different parts, a problem known to be NP-hard in the worst case [2]. Nevertheless, some simple algorithms like spectral methods are found to perform well in some networks [3]. Thus, a theory to predict when it is computationally feasible or infeasible to find communities is needed. To quote from the survey of Fortunato [4]:

"What the field lacks the most is a theoretical framework that defines precisely what clustering algorithms are supposed to do."

Inference from partial rankings In many online networks such as rating systems and crowdsourcing systems, a group of users give partial rankings over different subsets of items explicitly or implicitly. The problem is to combine the partial rankings and infer the inherent group preference over all the items. This problem, known as rank aggregation, has received much attention across various disciplines including statistics, psychology, sociology, and computer science.

While consistency of various rank aggregation algorithms has been studied when a growing number of sampled partial preferences is observed over a fixed number of items [5, 6], little is known in the high-dimensional setting where the number of items and number of observed partial rankings scale simultaneously, which arises in many modern datasets. Inference becomes even more challenging when each individual provides limited information. For example, in the well-known Netflix challenge dataset, 480,189 users submitted ratings on 17,770 movies, but on average a user rated only 209 movies. Intuitively, inference becomes harder when few partial rankings are available, or the sizes of the partial rankings (i.e., the number of items to be ranked) are small, meaning fewer observations. Here a central question is: *For a given graph used for assigning items to users for ranking, when and how can we efficiently predict the group preference, and what are the key graph characteristics that determine one's ability to infer the inherent preference?*

1.3 Planted Models and Fundamental Limits

This thesis takes the model-based approach, assuming an observation is generated from some planted model:

$$y = f(\theta^*; \text{noise}),$$

where θ^* is the planted parameter, y denotes the observation, and f represents some parametric model. The goal is to infer θ^* with the knowledge of y and the parametric model. In particular, we will focus on the following two planted models.

- **Planted Cluster Model**

  θ^*: Community membership of nodes.

  y: A graph.

  f: Each pair of two nodes is connected by an edge independently at random with probability p if they are from the same community; with probability q otherwise.

- **Plackett-Luce Model**

  θ^*: Group preference over all items.

y: Partial rankings.

f: For a given subset S of items, independently assign each item $i \in S$ an unobserved value X_i, exponentially distributed with mean $e^{-\theta_i^*}$, and then output the ascending order of $\{X_i : i \in S\}$.

Our planted cluster model encompasses several classical planted random graph models including planted clique [7], planted coloring [8], planted dense subgraph [9], planted partition [10], and the stochastic block model [11], which are widely used for studying the problem of finding communities [4, 12, 13, 9, 14, 15, 16, 17, 18, 19, 20, 21]. They also provide a venue for studying the average-case behaviors of many NP-hard graph theoretic problems including max-clique, max-cut, graph partitioning and coloring [22, 10]. The Plackett-Luce (PL) model [23] is widely used for studying the rank aggregation problem [24, 25, 26, 27, 28, 29]. In the special case with pairwise comparisons, the PL model reduces to the popular Bradley-Terry (BT) model [30].

This thesis focus on the following two fundamental limits under the two planted models:

- **Information limit**. The fundamental limit above which the accurate inference is possible and below which it is impossible for any algorithm regardless of its computational complexity.

- **Computational limit**. The fundamental limit above which the accurate inference is achievable in polynomial-time and below which it is computationally intractable.

Our study of the two fundamental limits is motivated by the following observations at the intersection of information theory, statistics, theoretical computer science, applied probability and random graphs, optimization, machine learning and social network analysis.

- From the perspective of information theory, the inference problem can be viewed as a communication problem, where the planted parameter is the information to be transmitted, the parametric model acts like an encoder and a noisy channel which encodes and distorts the information, and the inference task is nothing but the decoding procedure which aims to recover the information. This suggests it may be possible to find the fundamental limits of the inference problems using information-theoretic tools.

- From the perspective of theoretical statistics, the existing minimax bounds on estimation errors do not take computational complexity into account, and in the high-dimensional statistical inference problems, the inference procedures may not be computationally feasible. Thus it is important to understand the minimax bounds under computational complexity constraint.

- From the perspective of theoretical computer science, the existing worst-case hardness results are sometimes pessimistic: The inputs eliciting the worst-case behavior may rarely occur in practice. The average-case complexity may be a more relevant measure of an algorithm's performance.

- From the perspective of applied probability and random graphs, the existing phase transition results mostly focus on the threshold for the emergence of a special network structure under a simple random graph model, such as the emergence of the giant component in Erdős-Rényi random graph. It is of interest to establish similar phase transition results for inference problems in a network.

- From the perspective of optimization, the problem of finding communities can be cast as a combinatorial optimization problem. It is observed that semi-definite programming (SDP), as a convex relaxation of the combinatorial optimization, leads to an efficient solver. Therefore, it is of interest to quantify the performance limits of SDP and understand when the convex relaxation is tight or not.

- From the perspective of machine learning and social network analysis, the study of the fundamental limits provides a principled approach to develop efficient algorithms with strong theoretical performance guarantees.

1.4 Notation

Throughout this thesis, we use the following notation. All logarithms are natural unless the base is explicitly specified. We use the convention $0 \log 0 = 0$. For any positive integer N, let $[N] = \{1, \ldots, N\}$. For $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For any set S, let $|S|$ denote its cardinality. Let $s_1^n = \{s_1, \ldots, s_n\}$. We use standard big O notations, e.g., for any sequences $\{a_n\}$ and

$\{b_n\}$, $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if there is an absolute constant $c > 0$ such that $1/c \le a_n/b_n \le c$.

Let I denote the identity matrix, and J denote the all-one matrix. We write $X \succeq 0$ if X is positive semidefinite and $X \ge 0$ if all the entries of X are nonnegative. Let $\mathcal{S}^n$ denote the set of all $n \times n$ symmetric matrices. For $X \in \mathcal{S}^n$, let $\lambda_2(X)$ denote its second smallest eigenvalue. For any matrix Y, let $\|Y\|$ denote its spectral norm.

Let $\mathrm{Bern}(p)$ denote the Bernoulli distribution with mean p and $\mathrm{Binom}(N, p)$ denote the binomial distribution with N trials and success probability p. For random variables X, Y, we write $X \perp\!\!\!\perp Y$ if X is independent with Y. For probability measures $\mathbb{P}$ and $\mathbb{Q}$, let $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}|$ denote the total variation distance and $\chi^2(\mathbb{P}\|\mathbb{Q}) = \int \frac{(\mathrm{d}\mathbb{P}-\mathrm{d}\mathbb{Q})^2}{\mathrm{d}\mathbb{Q}}$ the χ^2-divergence. The Kullback-Leibler (KL) divergence between two Bernoulli distributions with means $u \in [0,1]$ and $v \in [0,1]$ is denoted by $D(u\|v) \triangleq u \log \frac{u}{v} + (1-u)\log \frac{1-u}{1-v}$. The distribution of a random variable X is denoted by P_X. We write $X \sim \mathbb{P}$ if $P_X = \mathbb{P}$. Let c_1, c_2 etc. denote universal constants whose values can be made explicit and are independent of the model parameters. We say a sequence of events $\{A_n\}$ holds with high probability if $\mathbb{P}[A_n] \ge 1 - c_1 n^{-c_2}$ for two universal positive constants c_1, c_2.

CHAPTER 2

FINDING COMMUNITIES WITHIN A NETWORK

In this chapter we formally introduce the problem of finding tight-knit communities within a network, summarize the previous work and overview main results given in Chapters 3, 4, 5, 6.

2.1 Generative Network Model

Consider the following model for generating a network with some underlying community structure with five parameters $n, r, K \in \mathbb{N}$ with $n \geq rK$, and $p, q \in [0, 1]$.

Definition 1 (Planted Cluster Model)**.** *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 (called outlier nodes); A random graph G is generated based on the cluster structure: Each pair of nodes is connected independently of all others by an edge with probability p (called in-cluster edge density) if they are in the same cluster, and otherwise with probability q (called out-cluster edge density).*

The goal is to recover the underlying clusters (up to a permutation of cluster indices) from the observation of the graph. Let $\{C_i^*\}_{i=1}^r$ denote the r true clusters. We say an estimator $\mathcal{A} : G \to \{\widehat{C}_i\}_{i=1}^r$ achieves $(1-\epsilon)$-approximation of the true clusters if there exists a permutation $\pi : [r] \to [r]$ such that $\sum_{i=1}^r |\widehat{C}_{\pi(i)} \cap C_i^*| \geq (1-\epsilon)rK$. In this thesis, we will primarily focus on exact recovery where $\epsilon = 0$, but our computational lower bounds hold for any $\epsilon < 1$. Correlated recovery defined in [15, 31] refers to $\epsilon < 1 - \frac{1}{r}$.

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 [32, 9]. It is also possible to allow for non-uniform cluster sizes [33], and heterogeneous edge probabil-

ities [34] and node degrees [35, 32]. These extensions are certainly important in practical applications; we will discuss them in Chapter 6.

By varying the values of the model parameters, the planted cluster model covers several classical models including planted clique, planted coloring, planted densest subgraph, planted partition, and the stochastic block models.

- **Planted r-Disjoint-Clique** [36]. Here $p = 1$ and $0 < q < 1$, so r cliques of size K are planted into an Erdős-Rényi random graph $\mathcal{G}(n, q)$. The special case with $r = 1$ is known as the planted clique (PC) problem [7].

- **Planted Densest Subgraph** [9]. Here $0 < q < p < 1$ and $r = 1$, so there is a subgraph of size K and density p planted into a $\mathcal{G}(n, q)$ graph.

- **Planted Partition** [10]. Also known as the *stochastic block model* [11]. Here $n = rK$ and $p, q \in (0, 1)$. The special case with $r = 2$ can be called *planted bisection* [10]. The case with $p < q$ is sometimes called *planted noisy coloring* or *planted r-cut* [15, 22].

- **Planted r-Coloring** [8]. 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.

Reduction to the $p > q$ case. 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) can be considered as generated with in/out-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.

2.2 Previous Work

There is a vast literature on community detection (see, e.g., the exposition [4] for a comprehensive survey), here we cover the fraction we see as most relevant. Detailed comparisons of existing results are given after the main theorems. Our

investigation of the information and computational limits of cluster recovery is inspired by the following four lines of research at the intersection of theoretical computer science, statistics, physics, information theory, machine learning and social network analysis.

Fundamental limits in planted clique model In the planted clique model, a clique of size K is planted in an Erdős-Rényi random graph $\mathcal{G}(n, 1/2)$. If the planted clique has size $K = o(\log n)$, recovery is impossible because $\mathcal{G}(n, \frac{1}{2})$ will have a clique with size at least K; if $K \geq 2(1+\epsilon)\log_2(n)$ for any $\epsilon > 0$, an exhaustive search succeeds [7]; if $K = \Omega(\sqrt{n})$, the state-of-the-art polynomial-time algorithms work [7, 37, 36, 38, 39, 40, 41]; if $K = \Omega(\sqrt{n \log n})$, the nodes in the clique can be easily identified by counting degrees [42]. It is an open problem to find polynomial-time algorithms for the $K = o(\sqrt{n})$ regime, and it is believed that this cannot be done [43, 44, 45, 46]. The same results also hold for detecting the planted clique in Erdős-Rényi $\mathcal{G}(n, \gamma)$ for any fixed $\gamma \in (0, 1)$, i.e., distinguishing $G(n, \gamma)$ from a model with a cliuqe of size K planted in $\mathcal{G}(n, \gamma)$. In particular, the hardness assumption of detecting the planted clique of size $o(\sqrt{n})$ in $\mathcal{G}(n, \gamma)$ with a positive constant γ is known as Planted Clique Hypothesis [47, 48]. Various hardness results in the theoretical computer science literature have been established based on the PC Hypothesis with $\gamma = \frac{1}{2}$, e.g. cryptographic applications [44], approximating Nash equilibrium [43], testing k-wise independence [45], etc. In this thesis, we will show the hardness of recovering a single cluster below a certain limit based on the Planted Clique Hypothesis. In contrast to most previous works, our computational lower bounds rely on the stronger assumption that the PC Hypothesis holds for any positive constant γ. An even stronger assumption that the PC Hypothesis holds for $\gamma = 2^{-(\log n)^{0.99}}$ has been used in [49] for public-key cryptography. It is an interesting open problem to prove that the PC Hypothesis for a fixed $\gamma \in (0, \frac{1}{2})$ follows from that for $\gamma = \frac{1}{2}$.

Fundamental limits in planted bisection model A recent line of research identifies the sharp cluster recovery threshold under the planted bisection model with two approximately equal-sized clusters. In the very sparse regime where $p = a/n$ and $q = b/n$ for two constants a and b, it was conjectured in [15] that correlated recovery (i.e., $(1-\epsilon)$-approximation with $\epsilon < \frac{1}{2}$) is possible if and only if $(a-b)^2 > 2(a+b)$; the converse part is proved in [31] and the achievability part is proved independently in [16, 50]. In the relatively sparse regime where

$p = \frac{a \log n}{n}$ and $q = \frac{b \log n}{\log n}$ for two constants a and b, it is shown in [51, 52] that exact cluster recovery is possible if and only if $\frac{a+b}{2} - \sqrt{ab} > 1$. Remarkably, the recovery in both cases can be achieved by a polynomial-time algorithm whenever it is possible, showing that there is no gap between the information and computational limits. The polynomial time algorithms in [51, 52] is obtained using an approximate cluster recovery algorithm followed by a local, greedy improvement procedure. It is open to find a one-step polynomial-time procedure to achieve the recovery threshold $\frac{a+b}{2} - \sqrt{ab} = 1$. It is conjectured in [52] that the semidefinite programming succeeds if $\frac{a+b}{2} - \sqrt{ab} > 1$, backed by compelling simulation results. We prove the conjecture is correct in Chapter 5. However, if there are more than two clusters, [15] and [31] conjecture a hard regime exists when constant factors are concerned. In Chapter 3, we conjecture a hard regime exists even when only polynomial factors are concerned, if the cluster size K is sub-linear in n.

Fundamental limits in minimax inference problems There is an emerging line of research (see, e.g.,[53, 54, 47, 55, 9, 56, 48]) which examines high dimensional inference problems from both the statistical and computational perspectives. Gaps between the information and computational limits are observed in detecting or recovering a single sparse principal component [57, 58], detecting or localizing a single sparse submatrix [54, 59, 60, 61], and detecting a single cluster [9, 56]. Computational lower bounds for detecting sparse principal components [47] and noisy biclustering (submatrix detection) [48] have been recently proved based on the PC Hypothesis with $\gamma = \frac{1}{2}$.

Polynomial-time cluster recovery algorithms Most efficient cluster recovery algorithms fall into one of the following three categories: spectral method, convex method, or approximate Bayesian inference. The spectral method extracts the eigenvectors corresponding to the top eigenvalues and applies K-means or some thresholding algorithm on the eigenvectors to recover the clusters (see, e.g., [62, 63] for a comprehensive survey). A rigorous analysis with performance guarantees can be found in [36, 13, 64]. The tensor method, as an interesting variant of the spectral method, is studied in [19] and shown to be even able to find overlapping clusters. The convex method is based on taking a semidefinite relaxation of the ML estimation (see, e.g., [32, 65, 34]). Exact MAP inference under the planted cluster model is computationally challenging; an efficient message-passing algorithm is proposed in [15] to approximate the MAP estimator. Our results on the

convex method improve upon the previously known statistical performance of polynomial-time algorithms.

2.3 Overview of Main Results

In Chapter 3, we present our general results allowing for any values of p, q, K and r, which are accurate up to constant factors. Here for ease of presentation, we specialize our general results to the following asymptotic regime:

$$p = cq = \Theta(n^{-\alpha}), \quad K = \Theta(n^{\beta}), \quad n \to \infty, \tag{2.1}$$

where $c > 1$ is a fixed constant, $\alpha \in [0, 1]$ governs the sparsity of the graph and $\beta \in [0, 1]$ captures the size of communities. Clearly the cluster recovery problem becomes more difficult if either α increases or β decreases. We show that the parameter space of (α, β) is partitioned into three regimes as depicted in Fig. 2.1:

- **The Easy Regime:** $\beta > \frac{1}{2} + \frac{\alpha}{2}$. The planted cluster can be perfectly recovered in polynomial-time with high probability via the semidefinite programming relaxation of the ML estimation.
- **The Hard Regime:** $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{2}$. The planted cluster can be exactly recovered with high probability by ML estimation using an exhaustive search.
- **The Impossible Regime:** $\beta < \alpha$. Regardless of the computational costs, no algorithm can exactly recover the planted cluster with vanishing probability of error.

The impossible and hard regimes are separated by the *information limit* below which the graph does not carry enough information about the underlying clusters, so recovery is impossible. Our semidefinite programming improves the best known performance guarantees of polynomial-time algorithms. We conjecture no polynomial-time algorithm can succeed in the hard regime. If the conjecture is true, then the easy and hard regimes are separated by the *computational limit* below which finding the clusters is computationally intractable.

In Chapter 4, we bring evidence to the conjecture by showing that achieving $(1 - \epsilon)$-approximation of a single planted cluster for any $\epsilon < 1$ is at least as hard as detecting a clique of size $o(\sqrt{n})$ planted in $\mathcal{G}(n, \gamma)$ with a constant edge

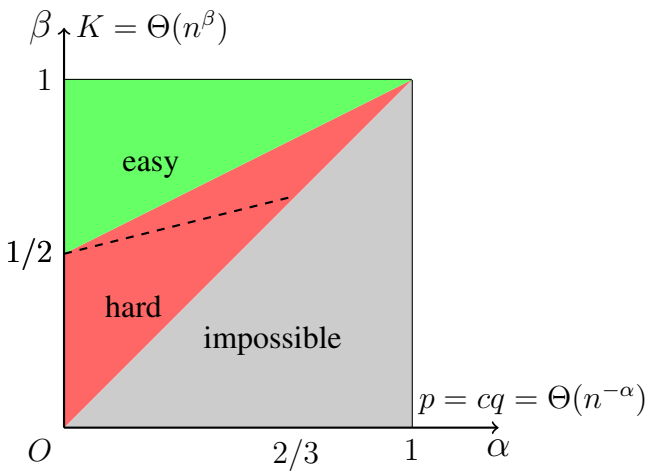

Figure 2.1: The simple (green), hard (red), impossible (gray) regimes for cluster recovery; $c > 1$ is a fixed constant.

probability $\gamma > 0$ when $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$ (red regime below the dashed line in Fig. 2.1). Since it is believed to be computationally intractable to detect a clique of size $o(\sqrt{n})$ planted in $\mathcal{G}(n, \gamma)$, our results imply that it is also computationally intractable to achieve $(1 - \epsilon)$-approximation of a single planted cluster for any $\epsilon < 1$ when $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$. Interestingly, if $\frac{1}{2} + \frac{\alpha}{4} < \beta < \frac{1}{2} + \frac{\alpha}{2}$ (red regime above the dashed line in Fig. 2.1), *detecting* the single planted cluster is computationally easy by thresholding the total number of edges, but we conjecture that *finding* the single planted cluster is computationally intractable. It is an interesting future work to prove our conjecture in this regime.

From Fig. 2.1, we also see that if $K = \Theta(n)$, the hard regime disappears. In fact, our general results in Chapter 3 show our semidefinite programming (SDP) achieves the information limit up to constant factors. In Chapter 5, we show the SDP even succeeds all the way down to the information limit without constant gaps in the following two special scaling regimes:

1. Binary symmetric stochastic block model (assuming n is even): $K = \frac{n}{2}$, $r = 2$, $p = \frac{a \log n}{n}$, and $q = \frac{b \log n}{n}$;
2. Planted dense subgraph model: $K = \lfloor \rho n \rfloor$, $r = 1$, $p = \frac{a \log n}{n}$, and $q = \frac{b \log n}{n}$,

where $a > b > 0$ and $0 < \rho < 1$ are fixed constants More specifically, under the binary symmetric stochastic block model, if $\sqrt{a} - \sqrt{b} > \sqrt{2}$, the SDP exactly

recovers the clusters with probability converging to one; if $\sqrt{a}-\sqrt{b} < \sqrt{2}$, any algorithms fails with probability converging to one regardless of the computational complexity. Under the planted dense subgraph model, if $\rho\left(a - \tau^* \log \frac{ea}{\tau^*}\right) > 1$ with $\tau^* = (a-b)/\log(a/b)$, the SDP exactly recovers the planted cluster with probability converging to one; if $\rho\left(a - \tau^* \log \frac{ea}{\tau^*}\right) < 1$, any algorithm fails with probability converging to one regardless of the computational complexity. It is an interesting open problem to understand if SDP achieves the sharp recovery threshold when there are a constant number of, but more than two, clusters of possibly unequal sizes.

In Chapter 6, we extend the analysis of the semidefinite program to the degree-corrected SBM (DCSBM), which is an extension of the SBM with heterogeneous cluster sizes and node degrees, and obtain a stronger performance guarantee than previously known. We also test the semidefinite program on the political blog datasets [66]: The empirical performance is comparable to the best known results in the literature.

CHAPTER 3

INFORMATION LIMITS AND EFFICIENT ALGORITHMS FOR FINDING COMMUNITIES

In this chapter, we first derive the information limit above which finding communities is possible and below which it is impossible for any algorithm to reliably find communities. Then, based on taking a semidefinite relaxation of the ML estimator, we derive an efficient convex program for finding communities and characterize its performance limit. We find that the performance limit depends on the spectral gap of the graph: Our convex method succeeds if and only if the spectral gap is large.

3.1 The Impossible Regime

In this section, we characterize the necessary conditions for cluster recovery. To facilitate subsequent discussion, we introduce a matrix representation of the cluster structure. For $k \in [r]$, let $\sigma_k^* \in \{0,1\}^n$ denote the indicator function of the true cluster k such that $\sigma_{ki}^* = 1$ if vertex i is in the true cluster k and $\sigma_{ki}^* = 0$ otherwise. We represent the true cluster structure by a *cluster matrix* $Y^* \in \{0,1\}^{n\times n}$ such that $Y^* = \sum_{k=1}^r \sigma_k^* \left(\sigma_k^*\right)^\top$. Notice that $Y_{ii}^* = 1$ if i is not an outlier node; otherwise $Y_{ii}^* = 0$, and $Y_{ij}^* = 1$ if and only if nodes i and j are in the same true cluster. The rank of Y^* equals the number of clusters r. The adjacency matrix of the graph is denoted as A, with the convention $A_{ii} = 0, \forall i \in [n]$. Under the planted cluster 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.

Let $\mathcal{Y}$ be the set of cluster matrices corresponding to r clusters of size K; i.e.,

$$\mathcal{Y} = \left\{ Y \mid Y = \sum_{k=1}^{r} \sigma_k \left(\sigma_k\right)^\top, \sigma_k \in \{0,1\}^n, \sigma_k \perp\!\!\!\perp \sigma_{k'}, \forall k \neq k', \right\}.$$

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 $D(u\|v) = u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v}$.

Theorem 3.1.1. *Suppose $128 \le K \le n/2$. Under the planted cluster model with $p > q$, if one of the following two conditions holds:*

$$K \cdot D(q\|p) \le \frac{1}{192} \left[\log(rK) \wedge K\right], \tag{3.1}$$

$$K \cdot D(p\|q) \le \frac{1}{192} \log n, \tag{3.2}$$

then

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

The theorem shows reliable cluster recovery is fundamentally impossible in the regime where (3.1) or (3.2) holds, which is thus called the *impossible regime*. This regime arises from an *information barrier*: The left hand sides of (3.1) and (3.2) measure how much information of Y^* is contained in the data A via KL divergence; while the right hand sides measure the amount of ambiguity in Y^* via the entropy. If the in-cluster and out-cluster edge distributions are close 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 (3.1) and (3.2) using its Taylor expansion. Applying the corollary to the asymptotic regime (2.1) implies that exact recovery is impossible if $\beta < \alpha$, as shown in Fig. 2.1.

Corollary 3.1.2. *Suppose $128 \le K \le n/2$. Under the planted clustering model with $p > q$, if any one of the following three conditions holds:*

$$K(p-q)^2 \le \frac{1}{192} q(1-q) \log n, \tag{3.3}$$

$$Kp \le \frac{1}{193} \left[\log(rK) \wedge K\right], \tag{3.4}$$

$$Kp \log \frac{p}{q} \le \frac{1}{192} \log n, \tag{3.5}$$

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 (3.1) and (3.2); this is made apparent in Corollary 3.1.2. To see why the asymmetry is natural, recall that by a classical result of [67], 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 (3.5). When $q = 0$, cluster recovery requires $p \gtrsim \frac{\log(rK)}{K}$ to ensure all true clusters are connected within themselves, matching the condition (3.4). The term K on the RHS of (3.1) and (3.4) is relevant only when $K \leq \log(rK)$.

Comparison to 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 [35, 68] establishes the necessary condition $p-q \lesssim \sqrt{p/n}$; our result is stronger by a logarithmic factor.

3.2 The Hard Regime

In this subsection, we characterize the sufficient conditions for cluster recovery which match the necessary conditions given in Theorem 3.1.1 up to constant factors. We consider the ML estimator of Y^* under the planted cluster model. The log-likelihood of observing the graph A given a cluster matrix $Y \in \mathcal{Y}$ is

$$\begin{aligned}\log \mathbb{P}_Y(A) &= \log \prod_{i<j} p^{A_{ij}Y_{ij}} q^{A_{ij}(1-Y_{ij})} (1-p)^{(1-A_{ij})Y_{ij}} (1-q)^{(1-A_{ij})(1-Y_{ij})} \\ &= \log \frac{p(1-q)}{q(1-p)} \sum_{i<j} A_{ij}Y_{ij} + \log \frac{1-p}{1-q} \sum_{i<j} Y_{ij} + \log \frac{q}{1-q} \sum_{i<j} A_{ij} \\ &\quad + \sum_{i<j} \log(1-q). \end{aligned} \tag{3.6}$$

Given A, the ML estimation maximizes the log-likelihood over the set $\mathcal{Y}$ of 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 (3.6) are independent of Y. Therefore, the ML estimation for the $p > q$ case is given as in Algorithm 1.

Algorithm 1 ML Estimator ($p > q$)

$$\widehat{Y} = \arg\max_{Y} \sum_{i,j} A_{ij} Y_{ij} \tag{3.7}$$

$$\text{s.t. } Y \in \mathcal{Y}. \tag{3.8}$$

Algorithm 1 is equivalent to finding r disjointclusters of size K that maximize the number of edges inside the clusters (similar to Densest K-Subgraph), or minimize number of edges outside the clusters (similar to Balanced Cut) or the disagreements between A and Y (similar to Correlation Clustering in [69]). Therefore, while Algorithm 1 is derived from the planted cluster 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 $|\mathcal{Y}| = \Omega(e^n)$.

The following theorem provides a success condition for the ML estimator.

Theorem 3.2.1. *Under the planted cluster model with $p > q$, there exists a universal constant c_1 such that the optimal solution $\widehat{Y}$ to the problem* (3.7)–(3.8) *is unique and equal to Y^* with probability at least $1 - 16(rK)^{-1} - 256n^{-1}$ if both of the following hold:*

$$\begin{aligned} K \cdot D(q\|p) &\geq c_1 \log(rK), \\ K \cdot D(p\|q) &\geq c_1 \log n. \end{aligned} \tag{3.9}$$

We refer to the regime in which the condition (3.9) holds but (3.14) below fails as the *hard regime*, because cluster recovery is statistically possible but conjectured to be computationally hard (cf. Section 3.3 and Conjecture 4.1.1). The conditions (3.9) and (3.1)–(3.2) in Theorem 3.1.1 match up to a constant factor under the mild assumption $K \geq \log(rK)$, establishing the information limit up to constant factors.

By lower bounding the KL divergence, we obtain the following corollary.

Corollary 3.2.2. *For the planted cluster model with $p > q$, there exists a universal constant c_2 such that the optimal solution $\widehat{Y}$ to the problem* (3.7)–(3.8) *is unique and equal to Y^* with probability at least $1 - 4(rK)^{-1} - 16n^{-1}$ provided*

$$K(p-q)^2 \geq c_2 q(1-q)\log n, \; Kp \geq c_2 \log(\gamma r K) \text{ and } Kp \log\frac{p}{q} \geq c_2 \log n. \tag{3.10}$$

The condition (3.10) 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 3.1.2 up to constant factors. Applying Corollary 3.2.2 to the asymptotic regime (2.1) implies that the ML estimator exactly recovers the clusters if $\beta > \alpha$, and thus the information limit is given by $\beta = \alpha$, as shown in Fig. 2.1.

Comparison to previous work Theorem 3.2.1 establishes the information limit tight up to constant factors. Interestingly, for a fixed cluster size, the recovery limit (3.9) depends only weakly on the number of clusters r through the logarithmic term. Our results recover the information limit $K \asymp \log n$ for the planted clique problem. For the planted densest subgraph model where $p/q = \Theta(1)$, p bounded away from 1 and $Kq \gg 1$, the detection limit is shown in [9] to be $\frac{(p-q)^2}{q} \asymp \min\{\frac{1}{K}\log\frac{n}{K}, \frac{n^2}{K^4}\}$; while our results show that the recovery limit is $\frac{(p-q)^2}{q} \asymp \frac{\log n}{K}$, which is strictly above the detection limit because $\frac{n^2}{K^4}$ can be much smaller than $\frac{\log n}{K}$. For the planted bisection model with two approximately equal-size clusters: if $p, q = \Theta(\log(n)/n)$, the recovery limit is found in [52] and [51] 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 [31, 50, 16] to be $K(p-q)^2 > p+q$, which is consistent with our results up to a logarithmic factor.

3.3 The Easy Regime

In this subsection, we present a polynomial-time semidefinite program based on taking a convex relaxation of the ML estimation in Algorithm 1. Note that the objective function (3.7) in the ML estimation is linear, but the constraint $Y \in \mathcal{Y}$ is non-convex. We replace this non-convex constraint with a trace norm (also known as nuclear norm) constraint and a set of linear constraints. This leads to a convex relaxation of ML estimation 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 (3.11)–(3.13) because $\|Y^*\|_* = \text{trace}(Y^*) = rK$.

Algorithm 2 Convex Relaxation of ML Estimator ($p > q$)

$$\widehat{Y} = \arg\max_{Y} \sum_{i,j} A_{ij} Y_{ij} \tag{3.11}$$

$$\text{s.t. } \|Y\|_* \leq rK, \tag{3.12}$$

$$\sum_{i,j} Y_{ij} = rK^2, \quad 0 \leq Y_{ij} \leq 1, \forall i, j \tag{3.13}$$

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 [70, 71]. Similarly to Algorithm 1, this algorithm is not strictly tied to the planted cluster model because 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 (3.12) and (3.13) with an appropriately weighted objective function [32]; we will discuss them in Chapter 6.

The following theorem provides a sufficient condition for the success of the convex relaxation of the ML estimator.

Theorem 3.3.1. *Under the planted cluster 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* (3.11)–(3.13) *is unique and equal to Y^* provided*

$$K^2(p-q)^2 \geq c_1 \left[p(1-q)K \log n + q(1-q)n\right]. \tag{3.14}$$

Since the convex relaxation of the ML estimator can be solved in polynomial-time, we refer to the regime where the condition (3.14) holds as the *easy regime*. If $p/q = \Theta(1)$, it is easy to see that the smallest possible cluster size allowed by (3.14) 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. In contrast, if $q = o(p)$ (we call it the high SNR setting), the condition (3.14) becomes $Kp \gtrsim \max\{\log n, \sqrt{qn}\}$. 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 $\sqrt{n}$.

Theorem 3.3.1 immediately implies guarantees for other tighter convex relax-

ations. Define the sets $\mathcal{B} := \{Y | (3.13) \text{ holds}\}$ and

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

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 the 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 (3.12) with $Y \in \mathcal{S}_2$, we obtain a *tighter* relaxation of the ML estimator, and Theorem 3.3.1 guarantees that it also recovers Y^* under the condition (3.14) with high probability.

We have a partial converse to the achievability result in Theorem 3.3.1. The following theorem characterizes the conditions under which the trace norm relaxation (3.11)–(3.13) 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 3.3.2 (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 (3.11)–(3.13)*.*

Theorem 3.3.2 proves the failure of our trace norm relaxation that has access to the *exact* number and sizes of the clusters. Consequently, replacing the constraints (3.12) and (3.13) with a Lagrangian penalty term in the objective would not help for *any* value of the Lagrangian multipliers. Theorems 3.3.1 and 3.3.2 together establish that under the assumptions of both theorems, the *sufficient and necessary* condition for the success of our trace norm relaxation is

$$\frac{K^2(p-q)^2}{pK + qn} \stackrel{\cdot}{\gtrsim} 1, \tag{3.15}$$

where $\stackrel{\cdot}{\gtrsim}$ ignores the logarithmic factors. Specializing (3.15) to the asymptotic regime (2.1) implies that the convexified ML estimator succeeds if $\beta > \frac{1}{2} + \frac{\alpha}{2}$ and fails if $\beta < \frac{1}{2} + \frac{\alpha}{2}$, as shown in Fig. 2.1.

Condition (3.15) is known as "spectral barrier" [14] in the literature. Let J denote the all-one matrix and I denote the identity matrix. Let $\widetilde{A} \triangleq A - qJ + qI$ be the centered adjacency matrix, which can be decomposed into a mean term plus a noise term:

$$\widetilde{A} = \mathbb{E}\widetilde{A} + \widetilde{A} - \mathbb{E}\widetilde{A} = (p-q)Y^* + A - \mathbb{E}A,$$

where here we adopt a different convention that $Y^*_{ii} = 0, \forall i$.

Observe that $\mathbb{E}\widetilde{A}$ contains the information about the true cluster matrix Y^* and the r largest eigenvalues of $\mathbb{E}\widetilde{A}$ are around $K(p-q)$, whereas $A - \mathbb{E}A$ is induced by the random noise and its largest eigenvalue is around $\Theta(\sqrt{Kp+qn})$ (see e.g., [72]). Therefore, the left hand side of (3.15) can be interpreted as the "spectral signal-to-noise ratio". If the ratio is much smaller than 1, then there is no spectral gap between the top r eigenvalues and the rest of the eigenvalues of A; therefore the spectrum of A cannot reveal useful information about the hidden cluster structure. In contrast, if the ratio is much larger than 1, then there is a large spectral gap and the top r eigenvectors of A can be exploited to estimate the hidden cluster structure. In fact, our proofs for Theorems 3.3.1 and 3.3.2 are built on this intuition.

Comparison to previous work We refer to [32] for a survey of the performance of state-of-the-art polynomial-time algorithms under various planted models. Theorem 3.3.1 matches and in many cases improves upon existing results in terms of the scaling. For example, for planted partition, the previous best results are $(p-q)^2 \gtrsim p(K\log^4 n + n)/K^2$ in [32] and $(p-q)^2 \gtrsim pn\,\mathrm{polylog}\, n/K^2$ in [19]. Theorem 3.3.1 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 $\Omega((rn + rK\log n)/K^2)$ [36], $\Omega(\sqrt{n}/K)$ [65] or $\Omega((n + K\log^4 n)/K^2)$ [32]. We improve them to $\Omega((n + K\log n)/K^2)$.

Our converse result in Theorem 3.3.2 is inspired by, and improves upon, the recent work in [73], which focuses on the special case $p > 1/2 > q$, and considers a convex relaxation approach that is equivalent to our relaxation (3.11)–(3.13) but without the additional equality constraint in (3.13). 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.

CHAPTER 4

COMPUTATIONAL LOWER BOUNDS FOR FINDING COMMUNITIES

By comparing the information limit established in Theorems 3.1.1 and 3.2.1 with the performance limit of our convex method established in Theorem 3.3.1, we get two strikingly different observations. On one hand, if $K = \Theta(n)$, the convex relaxation is tight and the hard regime disappears up to constants, even though the hard regime may still exist [31, 15] when constant factors are concerned. In this case, we get a computationally efficient and statistically order-optimal estimator. On the other hand, if $K = o(n)$, there exists a substantial gap between the information limit and performance limit of our convex method.

4.1 Conjecture on Computational Limits

We conjecture that no polynomial-time algorithm has order-wise better statistical performance than our convex method and succeeds beyond the performance limit of our convex method given in (3.14).

Conjecture 4.1.1. *For any given constant $\delta > 0$, 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*

$$(p-q)^2K^2 \leq n^{-\delta}\left(Kp(1-p) + q(1-q)n\right). \tag{4.1}$$

If the conjecture is true, then in the asymptotic regime (2.1), the *computational limit* for the cluster recovery is given by $\beta = \frac{\alpha}{2} + \frac{1}{2}$, i.e., the boundary between the green regime and red regime in Fig. 2.1.

A direct proof of Conjecture 4.1.1 seems difficult with current techniques. There are many possible convex formulations for cluster recovery. 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 3.3.2. Recall

that the planted dense subgraph (PDS) model is a special case of the planted cluster model where $0 < q < p < 1$ and $r = 1$. Let PDS-R (n, K, p, q, ϵ) denote the problem of *recovering* a $(1 - \epsilon)$-approximation of the planted cluster under the PDS model. We prove the computational lower bounds for PDS-R (n, K, p, q, ϵ) problem for any $\epsilon < 1$ conditional on the PC Hypothesis.

The proof is divided into two steps. First, we show that PDS-R (n, K, p, q, ϵ) problem for any $\epsilon < 1$ is at least as hard as the planted *binomial-sized* dense subgraph *detection* (PBDS-D) problem, i.e., distinguishing between Erdős-Rényi random graph model and the planted binomial-sized dense subgraph (PBDS) model. Second, adopting the standard reduction approach in complexity theory, we show that the PBDS-D problem in some regime below the (conjectured) computational limit is computationally intractable conditional on the PC Hypothesis.

4.2 Planted Binomial-sized Dense Subgraph Detection

In this section, we formally introduce the planted binomial-sized dense subgraph detection (PBDS-D) problem and connect it to PDS-R (n, K, p, q, ϵ).

Let $\mathcal{G}(N, q)$ denote the Erdős-Rényi random graph with N vertices, where each pair of vertices is connected independently with probability q. Let $\mathcal{G}(N, K, p, q)$ denote the PBDS model with N vertices where: (1) each vertex is included in the random set S independently with probability $\frac{K}{N}$; (2) for any two vertices, they are connected independently with probability p if both of them are in S and with probability q otherwise, where $p > q$. In this case, the vertices in S form a community with higher connectivity than elsewhere and $|S| \sim \text{Binom}(N, K/N)$. The planted dense subgraph here has a random size with mean K, which is similar to the models adopted in [15, 31, 16, 74, 50, 75], instead of a deterministic size K as assumed in [9, 56, 76].

Definition 2. *The planted binomial-sized dense subgraph detection problem with parameters* (N, K, p, q)*, henceforth denoted by* PBDS-D (N, K, p, q), refers to the problem of distinguishing hypotheses:

$$\begin{aligned} H_0 : &\quad G \sim \mathcal{G}(N, q) \triangleq \mathbb{P}_0, \\ H_1 : &\quad G \sim \mathcal{G}(N, K, p, q) \triangleq \mathbb{P}_1. \end{aligned}$$

The following theorem shows that PDS-R (N, K, p, q, ϵ) is at least as hard as

PBDS-D (N, K, p, q). Notice that in PDS-R (N, K, p, q, ϵ), the planted cluster has a deterministic size K, while in PBDS-D (N, K, p, q), the size of the planted cluster is binomially distributed with mean K.

Theorem 4.2.1. *For any given constant $\epsilon < 1$ and $c > 0$, suppose there is an algorithm $\mathcal{A}_N$ with running time T_N that with probability $1 - \eta_N$ solves the* PDS-R (N, K, cq, q, ϵ) problem. Then there exists a test ϕ_N with running time at most $N^2 + NT_N + NK^2$ that solves PBDS-D $(N, 2K, cq, q)$ with Type-I+II error probabilities upper bounded by $\eta_N + e^{-CK} + Ne^{-CK^2q}$ for a constant $C > 0$ only depending on ϵ and c.

4.3 Computational Lower Bounds for PBDS-D

In this section, we prove the computational lower bounds for PBDS-D (N, K, cq, q) assuming the intractability of the planted clique detection (PC-D) problem, which further implies the hardness of PDS-R (N, K, cq, q, ϵ) for any $\epsilon < 1$ in view of Theorem 4.2.1.

We first formally introduce the PC Hypothesis as our hardness assumption. Let $\mathcal{G}(n, k, \gamma)$ denote the planted clique model in which we add edges to k vertices uniformly chosen from $\mathcal{G}(n, \gamma)$ to form a clique.

Definition 3. *The planted clique detection problem with parameters (n, k, γ), denoted by* PC-D (n, k, γ) henceforth, refers to the problem of distinguishing hypotheses:

$$\begin{aligned} H_0^{\mathrm{C}} : &\quad G \sim \mathcal{G}(n, \gamma), \\ H_1^{\mathrm{C}} : &\quad G \sim \mathcal{G}(n, k, \gamma). \end{aligned}$$

The problem of finding or detecting the planted clique has been extensively studied for $\gamma = \frac{1}{2}$ and there is no known polynomial-time solver for either the PC recovery or detection problems for $k = o(\sqrt{n})$ and any constant $\gamma > 0$. It is conjectured [77, 43, 44, 45, 74] that the PC detection problem cannot be solved in polynomial time for $k = o(\sqrt{n})$ with $\gamma = \frac{1}{2}$, which we refer to as the PC Hypothesis.

Hypothesis 1 (PC Hypothesis)**.** *Fix some constant* $0 < \gamma \leq \frac{1}{2}$*. For any sequence of randomized polynomial-time tests* $\{\psi_{n,k_n}\}$ *such that* $\limsup_{n\to\infty} \frac{\log k_n}{\log n} < 1/2$*,*

$$\liminf_{n\to\infty} \mathbb{P}_{H_0^{\mathrm{C}}}\{\psi_{n,k}(G) = 1\} + \mathbb{P}_{H_1^{\mathrm{C}}}\{\psi_{n,k}(G) = 0\} \geq \frac{1}{2}.$$

The PC Hypothesis with $\gamma = \frac{1}{2}$ is similar to [48, Hypothesis 1] and [47, Hypothesis $\mathbf{B_{PC}}$]. Our computational lower bounds require that the PC Hypothesis holds for any positive constant γ. An even stronger assumption that PC Hypothesis holds for $\gamma = 2^{-(\log n)^{0.99}}$ has been used in [78, Theorem 10.3] for public-key cryptography. Furthermore, [74, Corollary 5.8] shows that under a statistical query model, any statistical algorithm requires at least $n^{\Omega(\frac{\log n}{\log(1/\gamma)})}$ queries for detecting the planted bi-clique in an Erdős-Rényi random bipartite graph with edge probability γ.

4.3.1 Polynomial-time Reduction from PC-D to PBDS-D

In this subsection, we show the PBDS-D problem can be approximately reduced from the PC-D problem of appropriately chosen parameters in randomized polynomial time. Based on this reduction scheme, we establish a formal connection between the PC-D problem and the PBDS-D problem in Proposition 4.3.1, and the desired computational lower bounds follow as Theorem 4.3.3.

We aim to reduce the PC-D (n, k, γ) problem to the PBDS-D (N, K, cq, q) problem. For simplicity, we focus on the case of $c = 2$; the general case follows similarly with a change in some numerical constants that come up in the proof. We are given an adjacency matrix $A \in \{0,1\}^{n\times n}$, or equivalently, a graph G, and with the help of additional randomness, will map it to an adjacency matrix $\widetilde{A} \in \{0,1\}^{N\times N}$, or equivalently, a graph $\widetilde{G}$ such that the hypothesis H_0^{C} (resp. H_1^{C}) in Definition 3 is mapped to H_0 exactly (resp. H_1 approximately) in Definition 2. In other words, if A is drawn from $\mathcal{G}(n,\gamma)$, then $\widetilde{A}$ is distributed according to $\mathbb{P}_0$; if A is drawn from $\mathcal{G}(n,k,1,\gamma)$, then the distribution of $\widetilde{A}$ is close in total variation to $\mathbb{P}_1$.

Our reduction scheme works as follows. Each vertex in $\widetilde{G}$ is randomly assigned a parent vertex in G, with the choice of parent being made independently for different vertices in $\widetilde{G}$, and uniformly over the set $[n]$ of vertices in G. Let V_s denote the set of vertices in $\widetilde{G}$ with parent $s \in [n]$ and let $\ell_s = |V_s|$. Then the sets

of children nodes $\{V_s : s \in [n]\}$ form a random partition of $[N]$. For any $1 \leq s \leq t \leq n$, the number of edges, $E(V_s, V_t)$, from vertices in V_s to vertices in V_t in $\widetilde{G}$ will be selected randomly with a conditional probability distribution specified below. Given $E(V_s, V_t)$, the particular set of edges with cardinality $E(V_s, V_t)$ is chosen uniformly at random.

It remains to specify, for $1 \leq s \leq t \leq n$, the conditional distribution of $E(s,t)$ given l_s, l_t, and $A_{s,t}$. Ideally, conditioned on ℓ_s and ℓ_t, we want to construct a Markov kernel from $A_{s,t}$ to $E(s,t)$ which maps $\mathrm{Bern}(1)$ to the desired edge distribution $\mathrm{Binom}(\ell_s\ell_t, p)$, and $\mathrm{Bern}(1/2)$ to $\mathrm{Binom}(\ell_s\ell_t, q)$, depending on whether both s and t are in the clique or not, respectively. Such a kernel, unfortunately, provably does not exist. Nonetheless, this objective can be accomplished approximately in terms of the total variation. For $s = t \in [n]$, let $E(V_s, V_t) \sim \mathrm{Binom}(\binom{\ell_t}{2}, q)$. For $1 \leq s < t \leq n$, denote $P_{\ell_s\ell_t} \triangleq \mathrm{Binom}(\ell_s\ell_t, p)$ and $Q_{\ell_s\ell_t} \triangleq \mathrm{Binom}(\ell_s\ell_t, q)$. Fix $0 < \gamma \leq \frac{1}{2}$ and put $m_0 \triangleq \lfloor \log_2(1/\gamma) \rfloor$. Define

$$P'_{\ell_s\ell_t}(m) = \begin{cases} P_{\ell_s\ell_t}(m) + a_{\ell_s\ell_t} & \text{for } m = 0 \\ P_{\ell_s\ell_t}(m) & \text{for } 1 \leq m \leq m_0 \\ \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) & \text{for } m_0 < m \leq \ell_s\ell_t \end{cases}$$

where $a_{\ell_s\ell_t} = \sum_{m_0 < m \leq \ell_s\ell_t} [P_{\ell_s\ell_t}(m) - \frac{1}{\gamma} Q_{\ell_s\ell_t}(m)]$. Let $Q'_{\ell_s\ell_t} = \frac{1}{1-\gamma}(Q_{\ell_s\ell_t} - \gamma P'_{\ell_s\ell_t})$. As we show later, $Q'_{\ell_s\ell_t}$ and $P'_{\ell_s\ell_t}$ are well-defined probability distributions as long as $\ell_s, \ell_t \leq 2\ell$ and $16q\ell^2 \leq 1$, where $\ell = N/n$. Then, for $1 \leq s < t \leq n$, let the conditional distribution of $E(V_s, V_t)$ given ℓ_s, ℓ_t, and $A_{s,t}$ be given by

$$E(V_s, V_t) \sim \begin{cases} P'_{\ell_s\ell_t} & \text{if } A_{st} = 1, \ell_s, \ell_t \leq 2\ell \\ Q'_{\ell_s\ell_t} & \text{if } A_{st} = 0, \ell_s, \ell_t \leq 2\ell \\ Q_{\ell_s\ell_t} & \text{if } \max\{\ell_s, \ell_t\} > 2\ell. \end{cases} \tag{4.2}$$

The next proposition (proved in Section A.6) shows that the randomized reduction defined above maps $\mathcal{G}(n, \gamma)$ into $\mathcal{G}(N, q)$ under the null hypothesis and $\mathcal{G}(n, k, \gamma)$ approximately into $\mathcal{G}(N, K, p, q)$ under the alternative hypothesis, respectively. The intuition behind the reduction scheme is as follows: By construction, $(1-\gamma)Q'_{\ell_s\ell_t} + \gamma P'_{\ell_s\ell_t} = Q_{\ell_s\ell_t} = \mathrm{Binom}(\ell_s\ell_t, q)$ and therefore the null distribution of the PC problem is exactly matched to that of the PDS problem, i.e., $P_{\widetilde{G}|H_0^C} = \mathbb{P}_0$. The core of the proof lies in establishing that the alternative distributions are approximately matched. The key observation is that $P'_{\ell_s\ell_t}$ is close

to $P_{\ell_s\ell_t} = \mathrm{Binom}(\ell_s\ell_t, p)$ and thus for nodes with distinct parents $s \neq t$ in the planted clique, the number of edges $E(V_s, V_t)$ is approximately distributed as the desired $\mathrm{Binom}(\ell_s\ell_t, p)$; for nodes with the same parent s in the planted clique, even though $E(V_s, V_s)$ is distributed as $\mathrm{Binom}(\binom{\ell_s}{2}, q)$ which is not sufficiently close to the desired $\mathrm{Binom}(\binom{\ell_s}{2}, p)$, after averaging over the random partition $\{V_s\}$, the total variation distance becomes negligible.

Proposition 4.3.1. *Let $\ell, n \in \mathbb{N}$, $k \in [n]$ and $\gamma \in (0, \frac{1}{2}]$. Let $N = \ell n$, $K = k\ell$, $p = 2q$ and $m_0 = \lfloor \log_2(1/\gamma) \rfloor$. Assume that $16q\ell^2 \leq 1$ and $k \geq 6e\ell$. If $G \sim \mathcal{G}(n, \gamma)$, then $\widetilde{G} \sim \mathcal{G}(N, q)$, i.e., $P_{\widetilde{G}|H_0^C} = \mathbb{P}_0$. If $G \sim \mathcal{G}(n, k, 1, \gamma)$, then*

$$\begin{aligned} &d_{\mathrm{TV}}\left(P_{\widetilde{G}|H_1^C}, \mathbb{P}_1\right) \\ &\leq e^{-\frac{K}{12}} + 1.5ke^{-\frac{\ell}{18}} + 2k^2(8q\ell^2)^{m_0+1} + 0.5\sqrt{e^{72e^2q\ell^2} - 1} + \sqrt{0.5k}e^{-\frac{\ell}{36}}. \quad (4.3) \end{aligned}$$

An immediate consequence of Proposition 4.3.1 is the following result (proved in Section A.7) showing that any PBDS-D solver induces a solver for a corresponding instance of the PC-D problem.

Proposition 4.3.2. *Let the assumption of Proposition 4.3.1 hold. Suppose $\phi : \{0,1\}^{\binom{N}{2}} \to \{0,1\}$ is a test for* PBDS-D $(N, K, 2q, q)$ with Type-I+II error probability η. Then $G \mapsto \phi(\widetilde{G})$ is a test for the PC-D (n, k, γ) whose Type-I+II error probability is upper bounded by $\eta + \xi$ with ξ given by the right-hand side of (4.3).

The following theorem establishes the computational limit of the PBDS-D problem.

Theorem 4.3.3. *Assume Hypothesis 1 holds for a fixed $0 < \gamma \leq 1/2$. Let $m_0 = \lfloor \log_2(1/\gamma) \rfloor$. Let $\alpha > 0$ and $0 < \beta < 1$ be such that*

$$\alpha < \beta < \frac{1}{2} + \frac{m_0\alpha + 4}{4m_0\alpha + 4}\alpha - \frac{2}{m_0\alpha}. \quad (4.4)$$

Then there exists a sequence $\{(N_\ell, K_\ell, q_\ell)\}_{\ell \in \mathbb{N}}$ satisfying

$$\lim_{\ell\to\infty} \frac{\log \frac{1}{q_\ell}}{\log N_\ell} = \alpha, \quad \lim_{\ell\to\infty} \frac{\log K_\ell}{\log N_\ell} = \beta$$

such that for any sequence of randomized polynomial-time tests $\phi_\ell : \{0,1\}^{\binom{N_\ell}{2}} \to \{0,1\}$ for the PBDS-D $(N_\ell, K_\ell, 2q_\ell, q_\ell)$ problem, the Type-I+II error probability

is lower bounded by

$$\liminf_{\ell\to\infty} \mathbb{P}_0\{\phi_\ell(G') = 1\} + \mathbb{P}_1\{\phi_\ell(G') = 0\} \geq \frac{1}{2},$$

where $G' \sim \mathcal{G}(N, q)$ under H_0 and $G' \sim \mathcal{G}(N, K, p, q)$ under H_1. Consequently, if Hypothesis 1 holds for all $0 < \gamma \leq 1/2$, then the above holds for all $\alpha > 0$ and $0 < \beta < 1$ such that

$$\alpha < \beta < \beta^\sharp(\alpha) \triangleq \frac{1}{2} + \frac{\alpha}{4}. \tag{4.5}$$

Consider the asymptotic regime (2.1). The function $\beta^\sharp$ in (4.5) gives the computational barrier for the PBDS-D (N, K, cq, q) problem. Notice that PBDS-D (N, K, cq, q) problem can be solved in linear time by thresholding based on the total number of edges if $\beta > \beta^\sharp$ [79].

In view of Theorem 4.2.1, Theorem 4.3.3 further implies the computational lower bounds for PDS-R $(N, K, 2q, q)$.

Corollary 4.3.4. *Assume Hypothesis 1 holds for a fixed $0 < \gamma \leq 1/2$. Given any constant $\epsilon < 1$ and $c > 0$. Let $m_0 = \lfloor \log_c(1/\gamma) \rfloor$. Let $\alpha > 0$ and $0 < \beta < 1$ be such that*

$$\alpha < \beta < \frac{1}{2} + \frac{m_0\alpha + 4}{4m_0\alpha + 4}\alpha - \frac{2}{m_0\alpha}. \tag{4.6}$$

Then there exists a sequence $\{(N_\ell, K_\ell, q_\ell)\}_{\ell\in\mathbb{N}}$ satisfying

$$\lim_{\ell\to\infty} \frac{\log \frac{1}{q_\ell}}{\log N_\ell} = \alpha, \quad \lim_{\ell\to\infty} \frac{\log K_\ell}{\log N_\ell} = \beta$$

such that any sequence of polynomial-time algorithms $\mathcal{A}_\ell : \{0,1\}^{\binom{N_\ell}{2}} \to S \subset [N_\ell]$ with $|S| = K_\ell$ fails to solve PDS-R *$(N_\ell, K_\ell, 2q_\ell, q_\ell)$* problem with probability at least $1/2$. Consequently, if Hypothesis 1 holds for all $0 < \gamma \leq 1/2$, then the above holds for all $\alpha > 0$ and $0 < \beta < 1$ such that

$$\alpha < \beta < \beta^\sharp(\alpha) \triangleq \frac{1}{2} + \frac{\alpha}{4}. \tag{4.7}$$

Corollary 4.3.4 implies that in the asymptotic regime (2.1), PDS-R (n, K, p, q) is computational intractable if $\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}$ conditional on the PC Hypothesis

(see Fig. 2.1).

Comparison to previous work on reduction from the PC **Problem** Most previous work [43, 45, 80, 78] in the theoretical computer science literature uses the reduction from the PC-D problem to generate computationally hard instances of problems and establish *worst-case* hardness results; the underlying distributions of the instances could be arbitrary. Similarly, in the recent works [47, 48] on the computational limits of certain *minimax* inference problems, the reduction from the PC-D problem is used to generate computationally hard but statistically feasible instances of their problems; the underlying distributions of the instances can also be arbitrary as long as they are valid priors on the parameter spaces. In contrast, here our goal is to establish the average-case hardness of the PBDS-D problem based on that of the PC-D problem. Thus the underlying distributions of the problem instances generated from the reduction must be close to the desired distributions under both the null and alternative hypotheses. To this end, we start with a small dense graph generated from $\mathcal{G}(n, \gamma)$ under H_0 and $\mathcal{G}(n, k, \gamma)$ under H_1, and arrive at a large sparse graph whose distribution is exactly $\mathcal{G}(N, q)$ under H_0 and approximately equal to $\mathcal{G}(N, K, p, q)$ under H_1. Notice that simply sparsifying the PC-D problem does not capture the desired tradeoff between the graph sparsity and the cluster size. Our reduction scheme differs from those used in [47, 48] which start with a large dense graph. Similar to ours, the reduction scheme in [80] also enlarges and sparsifies the graph by taking its subset power; but the distributions of the resulting random graphs are rather complicated and not close to Erdős-Rényi type graphs.

CHAPTER 5

ACHIEVING SHARP RECOVERY THRESHOLD VIA SDP

By comparing the information limit established in Theorems 3.1.1 and 3.2.1 and the performance limit of our convex method established in Theorem 3.3.1, we see the SDP achieves the information limit up to constants if $K = \Theta(n)$. It is tempting to ask whether SDP achieves the sharp information limit without constant gaps if $K = \Theta(n)$. In this chapter, we focus on the following particular cases:

- Binary symmetric stochastic block model (assuming n is even):

$$r = 2,\ K = \frac{n}{2},\ p = \frac{a \log n}{n},\ q = \frac{b \log n}{n},\ n \to \infty \tag{5.1}$$

- Planted dense subgraph model:

$$r = 1,\ K = \lfloor \rho n \rfloor,\ p = \frac{a \log n}{n},\ q = \frac{b \log n}{n},\ n \to \infty, \tag{5.2}$$

where $a \neq b$ and $0 < \rho < 1$ are fixed constants, and show the SDP achieves the sharp recovery thresholds.

5.1 Binary Symmetric Stochastic Block Model

Exact cluster recovery under the binary symmetric stochastic block model is studied in [52, 51] and a sharp recovery threshold is found.

Theorem 5.1.1 ([52, 51])**.** *Under the binary symmetric stochastic block model (5.1), if* $(\sqrt{a} - \sqrt{b})^2 > 2$*, clusters can be exactly recovered up to a permutation of cluster indices with probability converging to* 1*; if* $(\sqrt{a} - \sqrt{b})^2 < 2$*, any algorithm fails to exactly recover clusters with probability converging to* 1*.*

The optimal reconstruction threshold in Theorem 5.1.1 is achieved by the maximum likelihood (ML) estimator, which entails finding the minimum bisection

of the graph, a problem known to be NP-hard in the worst case [2, Theorem 1.3]. Nevertheless, it has been shown that the optimal recovery threshold can be attained in polynomial time using a two-step procedure [52, 51]: First, apply the partial recovery algorithms in [16, 50] to correctly cluster all but $o(n)$ vertices; Second, flip the cluster memberships of those vertices who do not agree with the majority of their neighbors. This two-step procedure has two limitations: a) the partial recovery algorithms used in the first step are sophisticated; b) the original graph needs to be split to implement the two steps to ensure their independence. It was left open to find a simple direct approach to achieve the exact recovery threshold in polynomial time. It was proved in [52] that a semidefinite programming (SDP) relaxation of the ML estimator succeeds if $(a-b)^2 > 8(a+b) + 8/3(a-b)$. Backed by compelling simulation results, it was further conjectured in [52] that the SDP relaxation can achieve the optimal recovery threshold. In this paper, we resolve this conjecture in the positive.

The cluster structure under the binary symmetric stochastic block model can be represented by a vector $\sigma \in \{\pm 1\}^n$ such that $\sigma_i = 1$ if vertex i is in the first cluster and $\sigma_i = -1$ otherwise. Let σ^* correspond to the true clusters. Then the ML estimator of σ^* for the case $a > b$ can be simply stated as

$$\begin{aligned} \max_{\sigma} \ & \sum_{i,j} A_{ij}\sigma_i\sigma_j \\ \text{s.t.} \ & \sigma_i \in \{\pm 1\}, \quad i \in [n] \\ & \sigma^\top \mathbf{1} = 0, \end{aligned} \tag{5.3}$$

which maximizes the number of in-cluster edges minus the number of out-cluster edges. This is equivalent to solving the NP-hard minimum graph bisection problem. Instead, let us consider its convex relaxation similar to the SDP relaxation studied in [81, 52]. Let $Y = \sigma\sigma^\top$. Then $Y_{ii} = 1$ is equivalent to $\sigma_i = \pm 1$ and $\sigma^\top \mathbf{1} = 0$ if and only if $\langle Y, J\rangle = 0$. Therefore, (5.3) can be recast as

$$\begin{aligned} \max_{Y,\sigma} \ & \langle A, Y\rangle \\ \text{s.t.} \ & Y = \sigma\sigma^\top \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle J, Y\rangle = 0. \end{aligned} \tag{5.4}$$

Notice that the matrix $Y = \sigma\sigma^\top$ is a rank-one positive semidefinite matrix. If we relax this condition by dropping the rank-one restriction, we obtain the following convex relaxation of (5.4), which is a semidefinite program:

$$\begin{aligned} \widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle J, Y \rangle = 0. \end{aligned} \tag{5.5}$$

We remark that (5.5) does not rely on any knowledge of the model parameters except that $a > b$; for the case $a < b$, we replace $\arg\max$ in (5.5) by $\arg\min$.

Let $Y^* = \sigma^*(\sigma^*)^\top$ and $\mathcal{Y}_n \triangleq \{\sigma\sigma^\top : \sigma \in \{-1, 1\}^n, \sigma^\top \mathbf{1} = 0\}$. The following result establishes the optimality of the SDP procedure:

Theorem 5.1.2. *If* $(\sqrt{a} - \sqrt{b})^2 > 2$*, then* $\min_{Y^* \in \mathcal{Y}_n} \mathbb{P}\{\widehat{Y}_{\mathrm{SDP}} = Y^*\} \to 1$ *as* $n \to \infty$.

5.2 Planted Dense Subgraph Model

In this section we turn to the planted dense subgraph model in the asymptotic regime (5.2), where there exists a single cluster of size ρN. To specify the optimal reconstruction threshold, define the following function: For $a, b \geq 0$, let

$$f(a,b) = \begin{cases} a - \tau^* \log \frac{ea}{\tau^*} & \text{if } a, b > 0, a \neq b \\ a & \text{if } b = 0 \\ b & \text{if } a = 0 \\ 0 & \text{if } a = b \end{cases}, \tag{5.6}$$

where $\tau^* \triangleq \frac{a-b}{\log a - \log b}$ if $a, b > 0$ and $a \neq b$. We show that if $\rho f(a,b) > 1$, exact recovery is achievable in polynomial-time via SDP with probability tending to one; if $\rho f(a,b) < 1$, any estimator fails to recover the cluster with probability tending to one regardless of the computational costs. The sharp threshold $\rho f(a,b) = 1$ is plotted in Fig. 5.1 for various values of ρ.

We first introduce the maximum likelihood estimator and its convex relaxation. For ease of notation, in this section we use a vector $\xi \in \{0,1\}^n$, as opposed to $\sigma \in \{\pm 1\}^n$ used in Section 5.1 for the SBM, as the indicator function of the

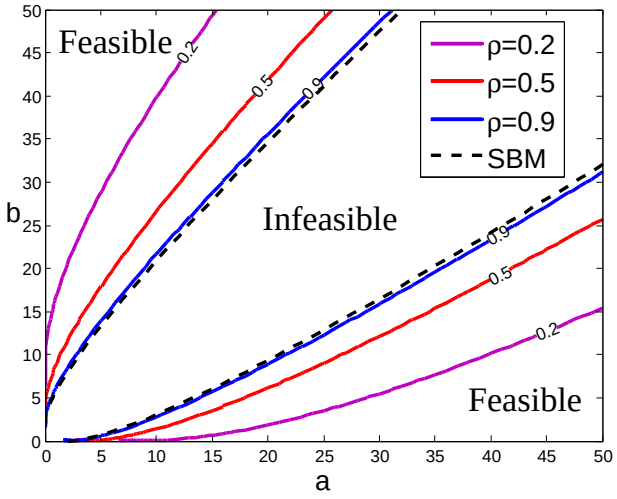

Figure 5.1: The recovery threshold: $\rho f(a,b) = 1$ (solid curves) for the planted dense subgraph model (5.2); $(\sqrt{a} - \sqrt{b})^2 = 2$ (dashed curve) for the stochastic block model (5.1).

cluster, such that $\xi_i = 1$ if vertex i is in the cluster and $\xi_i = 0$ otherwise. Let ξ^* be the indicator of the true cluster. Assuming $a > b$, i.e., the nodes in the cluster are more densely connected, the ML estimation of ξ^* is simply

$$\begin{aligned} \max_{\xi} \ & \sum_{i,j} A_{ij}\xi_i\xi_j \\ \text{s.t.} \ & \xi \in \{0,1\}^n \\ & \xi^\top \mathbf{1} = K, \end{aligned} \tag{5.7}$$

which maximizes the number of in-cluster edges. Due to the integrality constraints, it is computationally difficult to solve (5.7), which prompts us to consider

its convex relaxation. Note that (5.7) can be equivalently[1] formulated as

$$\begin{aligned}
\max_{Z,\xi}\ & \langle A, Z\rangle \\
\text{s.t.}\ & Z = \xi\xi^\top \\
& Z_{ii} \le 1, \quad \forall i \in [n] \\
& Z_{ij} \ge 0, \quad \forall i,j \in [n] \\
& \langle I, Z\rangle = K \\
& \langle J, Z\rangle = K^2,
\end{aligned} \tag{5.8}$$

where the matrix $Z = \xi\xi^\top$ is positive semidefinite and rank-one. Removing the rank-one restriction leads to the following convex relaxation of (5.8), which is a semidefinite program.

$$\begin{aligned}
\widehat{Z}_{\mathrm{SDP}} = \arg\max_{Z}\ & \langle A, Z\rangle \\
\text{s.t.}\ & Z \succeq 0 \\
& Z_{ii} \le 1, \quad \forall i \in [n] \\
& Z_{ij} \ge 0, \quad \forall i,j \in [n] \\
& \langle I, Z\rangle = K \\
& \langle J, Z\rangle = K^2.
\end{aligned} \tag{5.9}$$

We note that, apart from the assumption that $a > b$, the only model parameter needed by the estimator (5.9) is the cluster size K; for the case $a < b$, we replace $\arg\max$ in (5.9) by $\arg\min$.

Let $Z^* = \xi^*(\xi^*)^\top$ correspond to the true cluster and define $\mathcal{Z}_n = \big\{\xi\xi^\top : \xi \in \{0,1\}^n, \xi^\top \mathbf{1} = K\big\}$. The recovery threshold for the SDP (5.9) is given as follows.

Theorem 5.2.1. *Under the planted dense subgraph model (5.2), if*

$$\rho f(a,b) > 1, \tag{5.10}$$

then $\min_{Z^* \in \mathcal{Z}_n} \mathbb{P}\{\widehat{Z}_{\mathrm{SDP}} = Z^*\} \to 1$ *as* $n \to \infty$.

Next we prove a converse for Theorem 5.2.1 which shows that the recovery threshold achieved by the SDP relaxation is in fact optimal.

[1]Here (5.7) and (5.8) are equivalent in the following sense: for any feasible ξ for (5.7), $Z = \xi\xi^\top$ is feasible for (5.8); for any feasible Z, ξ for (5.8), either ξ or $-\xi$ is feasible for (5.7).

Theorem 5.2.2. *If*

$$\rho f(a,b) < 1, \tag{5.11}$$

then for any sequence of estimators $\widehat{Z}_n$, $\max_{Z^* \in \mathcal{Z}_n} \mathbb{P}\{\widehat{Z}_n = Z^*\} \to 0$.

Under the planted dense subgraph model, our investigation of the exact cluster recovery problem in this section has been focused on the regime where the cluster size K grows **linearly** with n, and $p, q = \Theta(\frac{\log n}{n})$, where the statistically optimal threshold can be attained by SDP in polynomial time. However, this need *not* be the case if K grows **sublinearly** in n. Recall our main results of Chapter 3 and 4 in the asymptotic regime (2.1). Assuming the PC Hypothesis, when $\alpha \in (0, \frac{2}{3})$ (and, quite possibly, the entire range $(0, 1)$), there exists a significant gap between the information limit (recovery threshold of the optimal procedure) and the computational limit (recovery threshold for polynomial-time algorithms). In contrast, in the asymptotic regime of (5.2), the computational constraint imposes no penalty on the statistical performance, in that the optimal threshold can be attained by SDP relaxation in view of Theorem 5.2.1.

CHAPTER 6

DEGREE-CORRECTED SBM AND EMPIRICAL STUDY ON REAL DATA

In this chapter, we analyze the convex method under the degree-corrected SBM (DCSBM), which includes heterogeneous cluster sizes and node degrees, and test the convex method on real data.

6.1 Degree-corrected SBM and Convex Method

The DCSBM is defined by five key parameters $n, r \in \mathbb{N}$, $p \geq q \in [0,1]$ and $\theta \in \mathbb{R}_+^n$, where θ_i controls the expected degree of node $i \in [n]$. Assume $p\theta_i\theta_j \geq 1$ for all $i \neq j$.

Definition 4 (DCSBM)**.** *Suppose there are n nodes indexed by $[n]$ and each node i is associated with a parameter θ_i. Assume the nodes are partitioned into r disjoint clusters $C_1^*, \ldots, C_r^*$ (called* true clusters*). A random graph is generated based on the cluster structure: Each pair of nodes i and j are connected independently of all others by an edge with probability $\theta_i\theta_j p$ (p is called* in-cluster edge density*) if they are in the same cluster, and otherwise with probability $\theta_i\theta_j q$ (q is called* out-cluster edge density*).*

Note that p, q, r and θ are allowed to be functions of n. The goal is to exactly recover the true clusters $\{C_m^*\}_{m=1}^r$ given the random graph. Recall that Y^* is the cluster matrix. Under DCSBM, we have, for all $i \neq j$, $\mathbb{P}(A_{ij} = 1) = \theta_i\theta_j p$ if $Y_{ij}^* = 1$ and $\mathbb{P}(A_{ij} = 1) = \theta_i\theta_j q$ if $Y_{ij}^* = 0$. The cluster recovery problem reduces to recovering Y^* given A. We can derive an efficient SDP by taking the convex relaxation of the ML estimation.

Algorithm 3 Convex Relaxation of ML Estimator under DCSBM with $p > q$

$$\widehat{Y} = \arg\max_{Y} \sum_{i<j} (A_{ij} - \theta_i \theta_j \lambda) Y_{ij} \tag{6.1}$$

$$\text{s.t. } Y \succeq 0, \tag{6.2}$$

$$Y_{ii} = 1, \forall i,\ 0 \le Y_{ij} \le 1, \forall i \ne j. \tag{6.3}$$

The following theorem provides a sufficient condition for the success of the convex method. For each $k \in [r]$, define $\theta^{(k)} \in \mathbb{R}^n_+$ such that $\theta^{(k)}_i = \theta_i$ if $i \in C^*_k$ and $\theta^{(k)}_i = 0$ otherwise. Let $\theta_{\min} = \min_i \theta_i$ and $\theta_{\max} = \max_i \theta_i$.

Theorem 6.1.1. *Consider the* DCSBM with $p > q$. Assume the tuning parameter λ in the problem (6.1)–(6.3) satisfies

$$\frac{1}{4}p + \frac{3}{4}q \le \lambda \le \frac{3}{4}p + \frac{1}{4}q.$$

Then, there exists a universal constant c_1 such that with high probability, the optimal solution to Algorithm 3 is unique and equal to Y^* provided

$$(p-q)^2 \min_k \|\theta^{(k)}\|_1 \theta_{\min} \ge c_1 p(1-q)\log n, \tag{6.4}$$

$$(p-q)^2 \left(\min_k \|\theta^{(k)}\|_1\right)^2 \ge c_1 q(1-q) n \log n. \tag{6.5}$$

In the special case with $\theta_i = 1$ for all $i \in [n]$ and clusters of equal size K, the DCSBM reduces to the classical SBM and conditions (6.4) and (6.5) reduce to condition (3.14):

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

6.2 Empirical Study on Political Blog Network

Note that Algorithm 3 involves the model parameter θ, which is unobservable in practice. Nevertheless we can get an estimator of $\theta_i\theta_j$ based on the degree sequence.

Lemma 1. *Consider the* DCSBM with $p > q$. Assume

$$\liminf_{n\to\infty} \min_k \|\theta^{(k)}\|_1/\|\theta\|_1 = \alpha, \quad \limsup_{n\to\infty} \max_k \|\theta^{(k)}\|_1/\|\theta\|_1 = \beta,$$

and $\limsup_{n\to\infty} \frac{rp\theta_{\max}}{(p+(r-1)q)\|\theta\|_1} = 0$. Then for all $i \neq j$,

$$\frac{((p-q)\alpha+q)^2}{(p-q)\beta+q} \leq \liminf_{n\to\infty} \frac{d_i d_j}{\theta_i\theta_j \sum_{i'} d_{i'}} \leq \limsup_{n\to\infty} \frac{d_i d_j}{\theta_i\theta_j \sum_{i'} d_{i'}} \leq \frac{((p-q)\beta+q)^2}{(p-q)\alpha+q}. \tag{6.6}$$

To interpret Lemma 1, consider the simple case with two symmetric communities where $r = 2$, $\alpha = \beta = \frac{1}{2}$ and $\theta_{\max} = o(\|\theta\|_1)$, then Lemma 1 implies that $\lim_{n\to\infty} \frac{d_i d_j}{\sum_{i'} d_{i'}} = \frac{p+q}{2}\theta_i\theta_j, \forall i \neq j$. Therefore, we can get a completely data-driven algorithm by replacing the term $\lambda\theta_i\theta_j$ in Algorithm 3 with $\frac{d_i d_j}{\sum_{i'} d_{i'}}$. More generally, we have the following data-driven SDP.

Algorithm 4 Data-driven SDP under DCSBM with $p > q$

$$\widehat{Y} = \arg\max_Y \sum_{i<j} \left(A_{ij} - \tau \frac{d_i d_j}{\sum_{i'} d_{i'}}\right) Y_{ij} \tag{6.7}$$

$$\text{s.t. } Y \succeq 0, \tag{6.8}$$

$$Y_{ii} = 1, \forall i,\ 0 \leq Y_{ij} \leq 1, \forall i \neq j, \tag{6.9}$$

where τ is called resolution tuning parameter.

Next, we connect theory with practice and test the empirical performance of Algorithm 4 on the US political blog network dataset [66]. This network dataset collected in 2005 consists of 19090 hyperlinks (directed edges) between 1490 political blogs. Also, the political leaning of all blogs (either liberal or conservative) is labeled manually based on blog directories, incoming and outgoing links and posts around the time of the 2004 presidential election. We treat these labels as the true community memberships. We pre-process the data by ignoring the edge directions and focus on the largest connected component with 1222 nodes, $16,714$ edges. Note that the graph has high degree variation: The max degree is 351 and the mean degree is around 27.

We solve for $\widehat{Y}$ in Algorithm 4 using the alternating direction method of multipliers as suggested in [34] and output clusters using k-means with $k = 2$ on $\widehat{Y}$.

Interestingly, letting $B_{ij} = A_{ij} - \frac{d_i d_j}{\sum_k d_k}, \forall i \neq j$ and $B_{ii} = 0, \forall i$, B is known as the modularity matrix [3]. The modularity maximization in [3] maximizes the objective function $\sum_{i<j} B_{ij} Y_{ij}$ over all possible cluster matrices $Y \in \mathcal{Y}$. Thus, Algorithm 4 with $\tau = 1$ can be interpreted as a convex relaxation of the modularity maximization. We find Algorithm 4 with $\tau = 1$ only misclassifies 62 nodes among 1222 nodes in total, which is comparable to the best known results in the literature: The SCORE method proposed in [64] misclassifies 58 nodes.

CHAPTER 7

INFERRING PREFERENCE FROM PARTIAL RANKINGS

Given a set of partial rankings from multiple decision makers or judges, in this chapter we study the problem of inferring the inherent preference of the whole population.

7.1 Problem Setup

We describe our model in the context of recommender systems, but it is applicable to other systems with partial rankings. Consider a recommender system with m users indexed by $[m]$ and n items indexed by $[n]$. For each item $i \in [n]$, there is a hidden parameter θ_i^* measuring the underlying preference. Each user j, independent of everyone else, randomly generates a partial ranking σ_j over a subset of items $S_j \subseteq [n]$ according to the PL model with the underlying preference vector $\theta^* = (\theta_1^*, \ldots, \theta_n^*)$.

Definition 5 (PL model)**.** *A partial ranking $\sigma : [|S|] \to S$ is generated from $\{\theta_i^*, i \in S\}$ under the* PL *model in two steps: (1) independently assign each item $i \in S$ an unobserved value X_i, exponentially distributed with mean $e^{-\theta_i^*}$; (2) select σ so that $X_{\sigma(1)} \leq X_{\sigma(2)} \leq \cdots \leq X_{\sigma(|S|)}$.*

The PL model can be equivalently described in the following sequential manner. To generate a partial ranking σ, first select $\sigma(1)$ in S randomly from the distribution $e^{\theta_i^*} / \big(\sum_{i' \in S} e^{\theta_{i'}^*} \big)$; secondly, select $\sigma(2)$ in $S \setminus \{\sigma(1)\}$ with the probability distribution $e^{\theta_i^*} / \big(\sum_{i' \in S \setminus \{\sigma(1)\}} e^{\theta_{i'}^*} \big)$; continue the process in the same fashion until all the items in S are assigned. The PL model is a special case of the following class of models.

Definition 6 (Thurstone model, or random utility model (RUM))**.** *A partial ranking $\sigma : [|S|] \to S$ is generated from $\{\theta_i^*, i \in S\}$ under the Thurstone model for*

a given CDF F in two steps: (1) independently assign each item $i \in S$ an unobserved utility U_i, with CDF $F(c - \theta_i^*)$; (2) select σ so that $U_{\sigma(1)} \geq U_{\sigma(2)} \geq \cdots \geq U_{\sigma(|S|)}$.

To recover the PL model from the Thurstone model, take F to be the CDF for the standard Gumbel distribution: $F(c) = e^{-(e^{-c})}$. Equivalently, take F to be the CDF of $-\log(X)$ such that X has the exponential distribution with mean one. For this choice of F, the utility U_i having CDF $F(c - \theta_i^*)$ is equivalent to $U_i = -\log(X_i)$ such that X_i is exponentially distributed with mean $e^{-\theta_i^*}$. The corresponding partial permutation σ is such that $X_{\sigma(1)} \leq X_{\sigma(2)} \leq \cdots \leq X_{\sigma(|S|)}$, or equivalently, $U_{\sigma(1)} \geq U_{\sigma(2)} \geq \cdots \geq U_{\sigma(|S|)}$. (Note the opposite ordering of X's and U's.)

Given the observation of all partial rankings $\{\sigma_j\}_{j\in[m]}$ over the subsets $\{S_j\}_{j\in[m]}$ of items, the task is to infer the underlying preference vector θ^*. For the PL model, and more generally for the Thurstone model, we see that θ^* and $\theta^* + a\mathbf{1}$ for any $a \in \mathbb{R}$ are statistically indistinguishable, where $\mathbf{1}$ is an all-ones vector. Indeed, under our model, the preference vector θ^* is the equivalence class $[\theta^*] = \{\theta : \exists a \in \mathbb{R}, \theta = \theta^* + a\mathbf{1}\}$. To get a unique representation of the equivalence class, we assume $\sum_{i=1}^n \theta_i^* = 0$. Then the space of all possible preference vectors is given by $\Theta = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \theta_i = 0\}$. Moreover, if $\theta_i^* - \theta_{i'}^*$ becomes arbitrarily large for all $i' \neq i$, then with high probability item i is ranked higher than any other item i' and there is no way to estimate θ_i to any accuracy. Therefore, we further put the constraint that $\theta^* \in [-b, b]^n$ for some $b \in \mathbb{R}$ and define $\Theta_b = \Theta \cap [-b, b]^n$. The parameter b characterizes the dynamic range of the underlying preference. In this chapter, we assume b is a fixed constant. As observed in [27], if b were scaled with n, then it would be easy to rank items with high preference versus items with low preference and one can focus on ranking items with close preference.

We denote the number of items assigned to user j by $k_j := |S_j|$ and the average number of assigned items per use by $k = \frac{1}{m}\sum_{j=1}^m k_j$; parameter k may scale with n in this chapter. We consider two scenarios for generating the subsets $\{S_j\}_{j=1}^m$: the random item assignment case where the S_j's are chosen independently and uniformly at random from all possible subsets of $[n]$ with sizes given by the k_j's, and the deterministic item assignment case where the S_j's are chosen deterministically.

Our main results depend on the structure of a weighted undirected graph G

defined as follows.

Definition 7 (Comparison graph G)**.** *Each item $i \in [n]$ corresponds to a vertex $i \in [n]$. For any pair of vertices i, i', there is a weighted edge between them if there exists a user who ranks both items i and i'; the weight equals $\sum_{j:i,i' \in S_j} \frac{1}{k_j - 1}$.*

Let A denote the weighted adjacency matrix of G. Let $d_i = \sum_j A_{ij}$, so d_i is the number of users who rank item i, and without loss of generality assume $d_1 \leq d_2 \leq \cdots \leq d_n$. Let D denote the $n \times n$ diagonal matrix formed by $\{d_i, i \in [n]\}$ and define the graph Laplacian L as $L = D - A$. Observe that L is positive semi-definite and the smallest eigenvalue of L is zero with the corresponding eigenvector given by the normalized all-one vector. Let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ denote the eigenvalues of L in ascending order.

Summary of main results. Theorem 7.3.1 gives a lower bound for the estimation error that scales as $\sum_{i=2}^{n} \frac{1}{d_i}$. The lower bound is derived based on a genie-argument and holds for both the PL model and the more general Thurstone model. Theorem 7.4.1 shows that the Cramér-Rao lower bound scales as $\sum_{i=2}^{n} \frac{1}{\lambda_i}$. Theorem 7.5.1 gives an upper bound for the squared error of the maximum likelihood (ML) estimator that scales as $\frac{mk \log n}{(\lambda_2 - \sqrt{\lambda_n})^2}$. Under the full rank breaking scheme that decomposes a k-way comparison into $\binom{k}{2}$ pairwise comparisons, Theorem 7.6.2 gives an upper bound that scales as $\frac{mk \log n}{\lambda_2^2}$. If the comparison graph is an expander graph, i.e., $\lambda_2 \sim \lambda_n$ and $mk = \Omega(n \log n)$, our lower and upper bounds match up to a $\log n$ factor. This follows from the fact that $\sum_i \lambda_i = \sum_i d_i = mk$, and for expanders $mk = \Theta(n\lambda_2)$. Since the Erdős-Rényi random graph is an expander graph with high probability for average degree larger than $\log n$, when the system is allowed to choose the item assignment, we propose a random assignment scheme under which the items for each user are chosen *independently and uniformly at random*. It follows from Theorem 7.3.1 that $mk = \Omega(n)$ is *necessary* for any item assignment scheme to reliably infer the underlying preference vector, while our upper bounds imply that $mk = \Omega(n \log n)$ is *sufficient* with the random assignment scheme and can be achieved by either the ML estimator or the full rank breaking or the independence-preserving breaking that decompose a k-way comparison into $\lfloor k/2 \rfloor$ non-intersecting pairwise comparisons, proving that rank breaking schemes are also nearly optimal.

7.2 Related Work

In this chapter, we study a statistical learning approach, assuming the observed ranking data is generated from a probabilistic model. Various probabilistic models on permutations have been studied in the ranking literature (see, e.g., [82, 26]). A nonparametric approach to modeling distributions over rankings using sparse representations has been studied in [83]. Most of the parametric models fall into one of the following three categories: noisy comparison model, distance based model, and random utility model. The noisy comparison model assumes that there is an underlying true ranking over n items, and each user independently gives a pairwise comparison which agrees with the true ranking with probability $p > 1/2$. It is shown in [84] that $O(n \log n)$ pairwise comparisons, when chosen adaptively, are sufficient for accurately estimating the true ranking.

The Mallows model is a distance-based model, which randomly generates a full ranking σ over n items from some underlying true ranking σ^* with probability proportional to $e^{-\beta d(\sigma,\sigma^*)}$, where β is a fixed spread parameter and $d(\cdot,\cdot)$ can be any permutation distance such as the Kemeny distance. It is shown in [84] that the true ranking σ^* can be estimated accurately given $O(\log n)$ independent full rankings generated under the Mallows model with the Kemeny distance.

In this chapter, we study a special case of random utility models (RUMs) known as the Plackett-Luce (PL) model. It is shown in [24] that the likelihood function under the PL model is concave and the ML estimator can be efficiently found using a minorization-maximization (MM) algorithm which is a variation of the general EM algorithm. We give an upper bound on the error achieved by such an ML estimator, and prove that this is matched by a lower bound. The lower bound is derived by comparing to an oracle estimator which observes the random utilities of RUM directly. The Bradley-Terry (BT) model is the special case of the PL model where we only observe pairwise comparisons. For the BT model, [27] proposes RankCentrality algorithm based on the stationary distribution of a random walk over a suitably defined comparison graph and shows $\Omega(n\mathsf{poly}(\log n))$ randomly chosen pairwise comparisons are sufficient to accurately estimate the underlying parameters; one corollary of our result is a matching performance guarantee for the ML estimator under the BT model. More recently, [85] analyzed various algorithms including RankCentrality and the ML estimator under a general, not necessarily uniform, sampling scheme.

In a PL model with priors, MAP inference becomes computationally challeng-

ing. Instead, an efficient message-passing algorithm is proposed in [25] to approximate the MAP estimate. For a more general family of random utility models, Soufiani et al. in [86, 87] give a sufficient condition under which the likelihood function is concave, and propose a Monte-Carlo EM algorithm to compute the ML estimator for general RUMs. More recently in [28, 29], the generalized method of moments together with the rank-breaking is applied to estimate the parameters of the PL model and the random utility model when the data consists of full rankings.

7.3 Oracle Lower Bound

In this section, we derive an oracle lower bound for any estimator of θ^*. The lower bound is constructed by considering an oracle who reveals all the hidden scores in the PL model as side information and holds for the general Thurstone models.

Theorem 7.3.1. *Suppose σ_1^m are generated from the Thurstone model for some CDF F. For any estimator $\widehat{\theta}$,*

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_b} E[||\widehat{\theta} - \theta^*||_2^2] \geq \frac{1}{2I(\mu) + \frac{2\pi^2}{b^2(d_1+d_2)}} \sum_{i=2}^{n} \frac{1}{d_i} \geq \frac{1}{2I(\mu) + \frac{2\pi^2}{b^2(d_1+d_2)}} \frac{(n-1)^2}{mk},$$

where μ is the probability density function of F, i.e., $\mu = F'$ and $I(\mu) = \int \frac{(\mu'(x))^2}{\mu(x)} dx$; the second inequality follows from the Jensen's inequality. For the PL *model, which is a special case of the Thurstone models with F being the standard Gumbel distribution, $I(\mu) = 1$.*

Theorem 7.3.1 shows that the oracle lower bound scales as $\sum_{i=2}^{n} \frac{1}{d_i}$. We remark that the summation begins with $1/d_2$. This makes some sense, in view of the fact that the parameters θ_i^* need to sum to zero. For example, if d_1 is a moderate value and all the other d_i's are very large, then with the hidden scores as side information, we may be able to accurately estimate θ_i^* for $i \neq 1$ and therefore accurately estimate θ_1^*. The oracle lower bound also depends on the dynamic range b and is tight for $b = 0$, because a trivial estimator that always outputs the all-zero vector achieves the lower bound.

Comparison to previous work Theorem 7.3.1 implies that $mk = \Omega(n)$ is necessary for any item assignment scheme to reliably infer θ^*, i.e., ensuring $E[||\widehat{\theta} -$

$\theta^*||_2^2] = o(n)$. It provides the first converse result on inferring the parameter vector under the general Thurstone models to our knowledge. For the Bradley-Terry model, which is a special case of the PL model where all the partial rankings reduce to the pairwise comparisons, i.e., $k = 2$, it is shown in [27] that $m = \Omega(n)$ is necessary for the random item assignment scheme to achieve the reliable inference based on the information-theoretic argument. In contrast, our converse result is derived based on the Bayesian Cramér-Rao lower bound [88], applies to the general models with any item assignment, and is considerably tighter if d_i's are of different orders.

7.4 Cramér-Rao Lower Bound

In this section, we derive the Cramér-Rao lower bound for any unbiased estimator of θ^*.

Theorem 7.4.1. *Let $k_{\max} = \max_{j\in[m]} k_j$ and $\mathcal{U}$ denote the set of all unbiased estimators of θ^*, i.e., $\widehat{\theta} \in \mathcal{U}$ if and only if $\mathbb{E}[\widehat{\theta}|\theta^* = \theta] = \theta, \forall \theta \in \Theta_b$. If $b > 0$, then*

$$\begin{aligned}\inf_{\widehat{\theta}\in\mathcal{U}} \sup_{\theta^*\in\Theta_b} \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] &\geq \left(1 - \frac{1}{k_{\max}} \sum_{\ell=1}^{k_{\max}} \frac{1}{\ell}\right)^{-1} \sum_{i=2}^{n} \frac{1}{\lambda_i} \\ &\geq \left(1 - \frac{1}{k_{\max}} \sum_{\ell=1}^{k_{\max}} \frac{1}{\ell}\right)^{-1} \frac{(n-1)^2}{mk},\end{aligned}$$

where the second inequality follows from the Jensen's inequality.

The Cramér-Rao lower bound scales as $\sum_{i=2}^{n} \frac{1}{\lambda_i}$. When G is disconnected, i.e., all the items can be partitioned into two groups such that no user ever compares an item in one group with an item in the other group, $\lambda_2 = 0$ and the Cramér-Rao lower bound is infinity, which is valid (and of course tight) because there is no basis for gauging any item in one connected component with respect to any item in the other connected component and the accurate inference is impossible for any estimator. Although the Cramér-Rao lower bound only holds for any unbiased estimator, we suspect that a lower bound with the same scaling holds for any estimator, but we do not have a proof.

7.5 ML Upper Bound

In this section, we study the ML estimator based on the partial rankings. The ML estimator of θ^* is defined as $\widehat{\theta}_{\mathrm{ML}} \in \arg\max_{\theta \in \Theta_b} \mathcal{L}(\theta)$, where $\mathcal{L}(\theta)$ is the log likelihood function given by

$$\mathcal{L}(\theta) = \log \mathbb{P}_\theta[\sigma_1^m] = \sum_{j=1}^{m} \sum_{\ell=1}^{k_j-1} \left[\theta_{\sigma_j(\ell)} - \log\left(\exp(\theta_{\sigma_j(\ell)}) + \cdots + \exp(\theta_{\sigma_j(k_j)})\right)\right]. \tag{7.1}$$

As observed in [24], $\mathcal{L}(\theta)$ is concave in θ and thus the ML estimator can be efficiently computed either via the gradient descent method or the EM type algorithms.

The following theorem gives an upper bound on the error rates inversely dependent on λ_2. Intuitively, by the well-known Cheeger's inequality, if the spectral gap λ_2 becomes larger, then there are more edges across any bi-partition of G, meaning more pairwise comparisons are available between any bi-partition of movies, and therefore θ^* can be estimated more accurately.

Theorem 7.5.1. *Assume $\lambda_n \geq C \log n$ for a sufficiently large constant C in the case with $k > 2$. Then with high probability,*

$$\|\widehat{\theta}_{\mathrm{ML}} - \theta^*\|_2 \leq \begin{cases} 4(1+e^{2b})^2 \lambda_2^{-1} \sqrt{m \log n} & \text{If } k = 2, \\ \frac{8e^{4b}\sqrt{2mk \log n}}{\lambda_2 - 16e^{2b}\sqrt{\lambda_n \log n}} & \text{If } k > 2. \end{cases}$$

We compare the above upper bound with the Cramér-Rao lower bound given by Theorem 7.4.1. Notice that $\sum_{i=1}^n \lambda_i = mk$ and $\lambda_1 = 0$. Therefore, $\frac{mk}{\lambda_2^2} \geq \sum_{i=2}^n \frac{1}{\lambda_i}$ and the upper bound is always larger than the Cramér-Rao lower bound. When the comparison graph G is an expander and $mk = \Omega(n \log n)$, by the well-known Cheeger's inequality, $\lambda_2 \sim \lambda_n = \Omega(\log n)$, the upper bound is only larger than the Cramér-Rao lower bound by a logarithmic factor. In particular, with the random item assignment scheme, we show that $\lambda_2, \lambda_n \sim \frac{mk}{n}$ if $mk \geq C \log n$ and as a corollary of Theorem 7.5.1, $mk = \Omega(n \log n)$ is sufficient to ensure $\|\widehat{\theta}_{\mathrm{ML}} - \theta^*\|_2 = o(\sqrt{n})$, proving the random item assignment scheme with the ML estimation is minimax-optimal up to a $\log n$ factor.

Corollary 7.5.2. *Suppose S_1^m are chosen independently and uniformly at random among all possible subsets of $[n]$. Then there exists a positive constant $C > 0$*

such that if $m \geq Cn \log n$ *when* $k = 2$ *and* $mk \geq Ce^{2b} \log n$ *when* $k > 2$*, then with high probability*

$$\|\widehat{\theta}_{\mathrm{ML}} - \theta^*\|_2 \leq \begin{cases} 4(1+e^{2b})^2 \sqrt{\frac{n^2 \log n}{m}}\,, & \text{if } k = 2, \\ 32e^{4b} \sqrt{\frac{2n^2 \log n}{mk}}\,, & \text{if } k > 2. \end{cases}$$

Comparison to previous work Theorem 7.5.1 provides the first finite-sample error rates for inferring the parameter vector under the PL model to our knowledge. For the Bradley-Terry model, which is a special case of the PL model with $k = 2$, [27] derived the similar performance guarantee by analyzing the rank centrality algorithm and the ML estimator. More recently, [85] extended the results to the non-uniform sampling scheme of item pairs, but the performance guarantees obtained when specialized to the uniform sampling scheme require at least $m = \Omega(n^4 \log n)$ to ensure $\|\widehat{\theta} - \theta^*\|_2 = o(\sqrt{n})$, while our results only require $m = \Omega(n \log n)$.

7.6 Rank Breaking Upper Bound

In this section, we study two rank-breaking schemes which decompose partial rankings into pairwise comparisons. For a partial ranking σ over S, i.e., σ is a mapping from $[|S|]$ to S, let σ^{-1} denote the inverse mapping.

Definition 8. *Given a partial ranking* σ *over the subset* $S \subset [n]$ *of size* k*, the independence-preserving breaking scheme (*IB*) breaks* σ *into* $\lfloor k/2 \rfloor$ *non-intersecting pairwise comparisons of form* $\{i_t, i'_t, y_t\}_{t=1}^{\lfloor k/2 \rfloor}$ *such that* $\{i_s, i'_s\} \cap \{i_t, i'_t\} = \emptyset$ *for any* $s \neq t$ *and* $y_t = 1$ *if* $\sigma^{-1}(i_t) < \sigma^{-1}(i'_t)$ *and* 0 *otherwise. The random* IB *chooses* $\{i_t, i'_t\}_{t=1}^{\lfloor k/2 \rfloor}$ *uniformly at random among all possibilities.*

If σ is generated under the PL model, then the IB breaks σ into independent pairwise comparisons generated under the PL model. Hence, we can first break partial rankings σ_1^m into independent pairwise comparisons using the random IB and then apply the ML estimator on the generated pairwise comparisons with the constraint that $\theta \in \Theta_b$, denoted by $\widehat{\theta}_{\mathrm{IB}}$. Under the random assignment scheme, as a corollary of Theorem 7.5.1, $mk = \Omega(n \log n)$ is sufficient to ensure $\|\widehat{\theta}_{\mathrm{IB}} - \theta^*\|_2 = o(\sqrt{n})$, proving the random item assignment scheme with the random IB is minimax-optimal up to a $\log n$ factor in view of the oracle lower bound in Theorem 7.3.1.

Corollary 7.6.1. *Suppose S_1^m are chosen independently and uniformly at random among all possible subsets of $[n]$ with size k. There exists a positive constant $C > 0$ such that if $mk \geq Cn \log n$, then with high probability,*

$$\|\widehat{\theta}_{\mathrm{IB}} - \theta^*\|_2 \leq 4(1 + e^{2b})^2 \sqrt{\frac{2n^2 \log n}{mk}}.$$

Definition 9. *Given a partial ranking σ over the subset $S \subset [n]$ of size k, the full breaking scheme (*FB*) breaks σ into all $\binom{k}{2}$ possible pairwise comparisons of form $\{i_t, i_t', y_t\}_{t=1}^{\binom{k}{2}}$ such that $y_t = 1$ if $\sigma^{-1}(i_t) < \sigma^{-1}(i_t')$ and 0 otherwise.*

If σ is generated under the PL model, then the FB breaks σ into pairwise comparisons which are not independently generated under the PL model. We pretend the pairwise comparisons induced from the full breaking are all independent and maximize the weighted log likelihood function $\mathcal{L}(\theta)$ given by

$$\sum_{j=1}^{m} \frac{1}{2(k_j - 1)} \sum_{i,i' \in S_j} \left(\theta_i \mathbf{1}_{\left\{\sigma_j^{-1}(i) < \sigma_j^{-1}(i')\right\}} + \theta_{i'} \mathbf{1}_{\left\{\sigma_j^{-1}(i) > \sigma_j^{-1}(i')\right\}} - \log\left(e^{\theta_i} + e^{\theta_{i'}}\right) \right) \tag{7.2}$$

with the constraint that $\theta \in \Theta_b$. Let $\widehat{\theta}_{\mathrm{FB}}$ denote the maximizer. Notice that we put the weight $\frac{1}{k_j - 1}$ to adjust the contributions of the pairwise comparisons generated from the partial rankings over subsets with different sizes.

Theorem 7.6.2. *With high probability,*

$$\|\widehat{\theta}_{\mathrm{FB}} - \theta^*\|_2 \leq 2(1 + e^{2b})^2 \frac{\sqrt{mk \log n}}{\lambda_2}.$$

Furthermore, suppose S_1^m are chosen independently and uniformly at random among all possible subsets of $[n]$. There exists a positive constant $C > 0$ such that if $mk \geq Cn \log n$, then with high probability,

$$\|\widehat{\theta}_{\mathrm{FB}} - \theta^*\|_2 \leq 4(1 + e^{2b})^2 \sqrt{\frac{n^2 \log n}{mk}}.$$

Theorem 7.6.2 shows that the error rates of $\widehat{\theta}_{\mathrm{FB}}$ inversely depend on λ_2. When the comparison graph G is an expander, i.e., $\lambda_2 \sim \lambda_n$, the upper bound is only larger than the Cramér-Rao lower bound by a logarithmic factor. The similar observation holds for the ML estimator as shown in Theorem 7.5.1. With the

random item assignment scheme, Theorem 7.6.2 implies that the FB only needs $mk = \Omega(n \log n)$ to achieve the reliable inference, which is optimal up to a $\log n$ factor in view of the oracle lower bound in Theorem 7.3.1.

Comparison to previous work The rank breaking schemes considered in [28, 29] break the full rankings according to rank positions while our schemes break the partial rankings according to the item indices. The results in [28, 29] establish the consistency of the generalized method of moments under the rank breaking schemes when the data consists of full rankings. In contrast, Corollary 7.6.1 and Theorem 7.6.2 apply to the more general setting with partial rankings and provide the finite-sample error rates, proving the optimality of the random IB and FB with the random item assignment scheme.

7.7 Numerical Experiments

Suppose there are $n = 1024$ items and θ^* is uniformly distributed over $[-b, b]$. We first generate d full rankings over 1024 items according to the PL model with parameter θ^*. Then for each fixed $k \in \{512, 256, \ldots, 2\}$, we break every full ranking σ into n/k partial rankings over subsets of size k as follows: Let $\{S_j\}_{j=1}^{n/k}$ denote a partition of $[n]$ generated uniformly at random such that $S_j \cap S_{j'} = \emptyset$ for $j \neq j'$ and $|S_j| = k$ for all j; generate $\{\sigma_j\}_{j=1}^{n/k}$ such that σ_j is the partial ranking over set S_j consistent with σ. In this way, in total we get $m = dn/k$ k-way comparisons which are all independently generated from the PL model. We apply the minorization-maximization (MM) algorithm proposed in [24] to compute the ML estimator $\widehat{\theta}_{\mathrm{ML}}$ based on the k-way comparisons and the estimator $\widehat{\theta}_{\mathrm{FB}}$ based on the pairwise comparisons induced by the FB. The estimation error is measured by the rescaled mean square error (MSE) defined by $\log_2\left(\frac{mk}{n^2}\|\widehat{\theta} - \theta^*\|_2^2\right)$.

We run the simulation with $b = 2$ and $d = 16, 64$. The results are depicted in Fig. 7.1. We also plot the Cramér-Rao (CR) limit given by $\log_2\left(1 - \frac{1}{k}\sum_{l=1}^{k}\frac{1}{l}\right)^{-1}$ as per Theorem 7.4.1. The oracle lower bound in Theorem 7.3.1 implies that the rescaled MSE is at least 0. We can see that the rescaled MSE of the ML estimator $\widehat{\theta}_{\mathrm{ML}}$ is close to the CR limit and approaches the oracle lower bound as k becomes large, suggesting the ML estimator is minimax-optimal. Furthermore, there is an approximately constant gap between the rescaled MSE of $\widehat{\theta}_{\mathrm{FB}}$ and the CR limit, suggesting that the FB is minimax-optimal up to a constant factor.

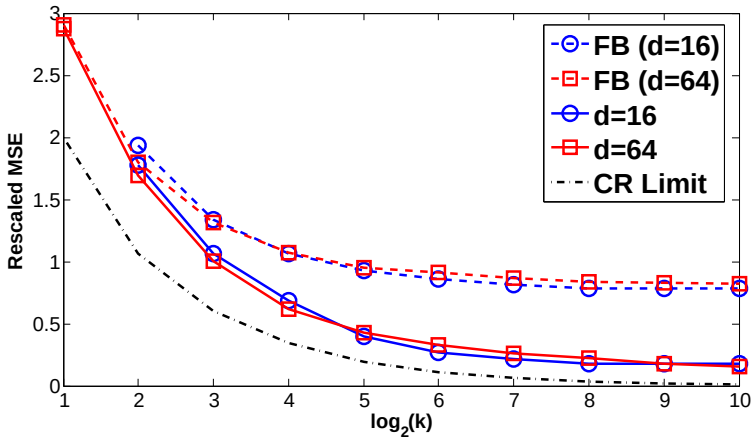

Figure 7.1: The error rate based on nd/k k-way comparisons with and without full breaking.

Finally, we point out that when $d = 16$ and $\log_2(k) = 1$, the MSE returned by the MM algorithm is infinity. Such singularity occurs for the following reason. Suppose we consider a directed comparison graph with nodes corresponding to items such that for each (i, j), there is a directed edge $(i \to j)$ if item i is ever ranked higher than j. If the graph is not strongly connected, i.e., if there exists a partition of the items into two groups A and B such that items in A are always ranked higher than items in B, then if all $\{\theta_i : i \in A\}$ are increased by a positive constant a, and all $\{\theta_i : i \in B\}$ are decreased by another positive constant a' such that all $\{\theta_i, i \in [n]\}$ still sum up to zero, the log likelihood (7.1) must increase; thus, the log likelihood has no maximizer over the parameter space Θ, and the MSE returned by the MM algorithm will diverge. Theoretically, if b is a constant and d exceeds the order of $\log n$, the directed comparison graph will be strongly connected with high probability and so such singularity does not occur in our numerical experiments when $d \geq 64$. In practice we can deal with this singularity issue in three ways: 1) find the strongly connected components and then run MM in each component to come up with an estimator of θ^* restricted to each component; 2) introduce a proper prior on the parameters and use Bayesian inference to come up with an estimator (see [25]); 3) add to the log likelihood objective function a regularization term based on $\|\theta\|_2$ and solve the regularized ML using the gradient descent algorithms (see [27]).

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This thesis studies the information and computational limits and efficient algorithms, for two typical statistical inference problems in networks: Finding communities within a network and inferring group preference from partial rankings.

Our derivation of information limits is based on Fano's inequality, Bayesian Cramér-Rao lower bound, and non-asymptotic analysis of the ML estimators. The techniques developed represent an important contribution to not only the study of community detection and rank aggregation, but also other statistical inference problems in networks. The computational limits for finding communities is established by reducing the cluster recovery problem from the well-known planted clique detection problem. The reduction proof deviates substantially from the classical reduction arguments in the worst-case computational complexity theory and draws upon a number of tools from statistics and discrete probability, such as the second moment method, decoupling argument and negative associations. The reduction scheme could be useful for studying the average-case complexity in other inference problems, a topic becomingly increasingly important in both computer science and statistics. The efficient algorithm proposed for finding communities is based on the semidefinite programming relaxation of the ML estimator. Our derivation of the performance limits of the SDP exploits the convex duality theory and spectral properties of random graphs. The techniques developed could be useful to analyze the SDP in other combinatorial optimization problems.

Going forward, there are several interesting research directions.

8.1 Computation Lower Bounds for Statistical Inference

My investigation of the fundamental limits for community detection unveils an interesting phase transition in the statistical and computational complexities, but

only touches a tip of the iceberg. Appearances of a hard regime are also observed in some other related inference problems such as detecting or recovering a single sparse principal component [57, 58], detecting or localizing a single sparse submatrix [54, 59, 60, 61], and detecting a single cluster [9, 56]. In contrast, in some special cases with two symmetric clusters of size $n/2$ or a single cluster of size proportional to n, we discover the SDP achieves the sharp information limit and there is no hard regime. A theory to completely characterize the performance limit of the SDP and to predict the existence of the hard regime is needed, and such theory will have profound impact in many disciplines.

8.2 Space Lower Bounds for Statistical Inference

In this thesis, we primarily focus on computational complexity rather than other types of complexity. However, the available space resource is a bottleneck in many systems. For example, in applications to big data analysis, data often comes in a stream fashion and we would like to design estimators which access the data in few passes (ideally, a single pass) over the input stream and use limited (ideally, sublinear in input size) memory space. Therefore, it is fundamentally important to characterize the estimation accuracy under space constraint. To be more specific, consider the planted clique detection problem (Definition 3). Assume the edges of the graph come in a stream and one is interested in distinguishing between an Erdős-Rényi random graph $\mathcal{G}(n, \frac{1}{2})$ and the planted clique model with a hidden clique of size K in a single pass. Notice that when $K = \omega(\sqrt{n})$, the simple algorithm based on thresholding the total number of edges only uses $O(\log n)$ space. Also, it is known that one can maintain a spectral sparsifier of the graph in $\widetilde{O}(n/\epsilon^2)$ space that approximates the original graph spectrum up to a multiplicative factor $(1+\epsilon)$ [89]. Hence, if $K = \Omega(\sqrt{n})$, one can detect the hidden clique by applying the spectral method on the spectral sparsifier, which only uses $\widetilde{O}(n)$ space. From these two observations, it is tempting to ask if one can detect the clique of size $\Theta(\sqrt{n})$ in sub-polynomial (say, poly-logarithmic) space. More broadly, what are the fundamental limits for community detection under a given space constraint? This streaming community detection problem has many applications such as detecting new events or topics, and has deep connections with communication complexity, spectral graph theory, and stochastic optimization.

APPENDIX A

PROOFS FOR FINDING COMMUNITIES

In this section, we give proofs for finding communities under the planted cluster model with $p > q$. Let $n_1 \triangleq rK$ and $n_2 \triangleq n - rK$ denote the numbers of non-outlier nodes and outlier nodes, respectively. Let J and I denote the all-one matrix and identify matrix, respectively. 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 Frobenius norm $\|X\|_{\mathrm{F}} = \sum_{i,j} |X_{i,j}|^2$; the ℓ_1 norm $\|X\|_1 = \sum_{i,j} |X_{ij}|$; and the ℓ_∞ norm $\|X\|_\infty = \max_{i,j} |X_{ij}|$. Let $\langle X, Y\rangle \triangleq \mathrm{Tr}(X^\top Y)$ denote the inner product between two matrices and then $\|X\|_{\mathrm{F}}^2 = \langle X, X\rangle$.

A.1 Proof of Theorem 3.1.1 and Corollary 3.1.2

We will make use of the following upper and lower bounds on the KL-divergence $D(u\|v)$ between two Bernoulli distributions with means $u \in [0,1]$ and $v \in [0,1]$:

$$\begin{aligned} D\left(u\|v\right) &\triangleq u\log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v} \\ &\overset{(a)}{\leq} u\frac{u-v}{v} + (1-u)\frac{v-u}{1-v} = \frac{(u-v)^2}{v(1-v)}, \end{aligned} \tag{A.1}$$

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

$$D\left(u\|v\right) = D\left(v\|v\right) + (u-v)D'\left(v\|v\right) + \frac{(u-v)^2}{2}D''\left(\xi\|v\right) \tag{A.2}$$

$$\overset{(b)}{\geq} \frac{(u-v)^2}{2(u\vee v)(1-u\wedge v)}, \tag{A.3}$$

where (b) follows because $D'(v\|v) = 0$ and

$$D''(\xi\|v) = \frac{1}{\xi(1-\xi)} \geq \frac{1}{(u \vee v)(1 - u \wedge v)}.$$

Theorem 3.1.1 is established through the following three lemmas, each of which provides a sufficient condition for having a large error probability.

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

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

Moreover, (A.4) *is implied by*

$$K(p-q)^2 \leq \frac{1}{4} q(1-q) \log \frac{n}{K}. \tag{A.5}$$

Proof. We use an information theoretic argument via Fano's inequality. Recall that $\mathcal{Y}$ is the set of 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 cluster model. Lower-bounding the supremum by the average, we have

$$\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \geq \inf_{\widehat{Y}} \mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right]. \tag{A.6}$$

It suffices to bound $\mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right]$. Let $H(X)$ be 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 $\widehat{Y}$,

$$\mathbb{P}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right] \geq 1 - \frac{I(Y^*;A) + 1}{\log|\mathcal{Y}|}. \tag{A.7}$$

We first lower bound $\log|\mathcal{Y}|$. Simple counting gives that $|\mathcal{Y}| = \binom{n}{n_1}\frac{n_1!}{r!(K!)^r}$, where $n_1 \triangleq rK$. 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 $8 \leq K \leq \frac{n}{2}$ and

$n \geq 32$.

Next we upper bound $I(Y^*; A)$. Note that $H(A) \leq \binom{n}{2} H(A_{12})$ because A_{ij}'s are identically distributed by symmetry. Furthermore, A_{ij}'s are independent conditioning 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 bound $I(Y^*_{12}; A_{12})$ as follows. Simple 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 $\mathbb{P}(A_{12} = 1) = \delta p + (1-\delta)q = \bar{p}$. It follows that $I(Y^*_{12}; A_{12}) = \delta D(p\|\bar{p}) + (1-\delta) D(q\|\bar{p})$. Using the upper bound (A.1) on the KL divergence and condition (A.4), we obtain

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

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 (A.7) 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 $K \geq \frac{n}{2}$ and $n_1 \geq 32$. This proves the sufficiency of (A.4).

We turn to the second part of the lemma. Using the upper bound (A.1) on the KL divergence, we get

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

where (a) holds because $\frac{n^2\delta}{n_1} \leq K$ and

$$\bar{p}(1-\bar{p}) \geq \delta p(1-p) + (1-\delta)q(1-q) \geq (1-\delta)q(1-q)$$

thanks to the concavity of $x(1-x)$. Therefore, condition (A.5) implies condi-

tion (A.4).

□

Lemma 3. *Suppose* $128 \le K \le \frac{n}{2}$. *Then* $\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \frac{1}{2}$ *if*

$$K \max\left\{D(p\|q), D(q\|p)\right\} \le \frac{1}{24}\log(n-K). \tag{A.8}$$

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 $\bar{M}+1$, which is specified later. Let $\bar{\mathbb{P}}_{(Y^*,A)}$ denote the joint distribution of (Y^*, A) when Y^* is sampled from $\bar{\mathcal{Y}}$ uniformly at random and then A is generated from the planted cluster model based on Y^*. By Fano's inequality, we have

$$\inf_{\widehat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \inf_{\widehat{Y}} \bar{\mathbb{P}}_{(Y^*,A)}\left[\widehat{Y} \neq Y^*\right] \ge \inf_{\widehat{Y}} \left\{1 - \frac{I(Y^*;A)+1}{\log|\bar{\mathcal{Y}}|}\right\}. \tag{A.9}$$

We construct $\bar{\mathcal{Y}}$ as follows. Let Y_0 be the cluster matrix with clusters $\{C_l\}_{l=1}^r$ 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 node K and node $K+i$. Formally, for each $i \in [\bar{M}]$: (1) if node $(K+i)$ belongs to cluster C_l for some l, then Y_i has the first cluster given by $\{1, 2, \ldots, K-1, K+i\}$ and the l-th cluster given by $C_l \setminus \{K+i\} \cup \{K\}$, and all the other clusters identical to those of Y_0; (2) if node $(K+i)$ is an outlier node in Y_0, then Y_i has the first cluster given by $\{1, 2, \ldots, K-1, K+i\}$, and node K as an outlier node, and all the other clusters identical to those of 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 assumption (A.8) holds, then $I(Y;A) \le \frac{1}{4}\log(n-K) = \frac{1}{4}\log\left|\bar{\mathcal{Y}}\right|$. Since $\log(n-K) \ge \log(n/2) \ge 4$ if $n \ge 128$, it follows from (A.9) that the minimax error probability is at least $1/2$. □

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

$$Kp \le \frac{1}{8}\min\{\log(rK/2), K\}, \tag{A.10}$$

$$\underline{\textbf{or}} \quad K(1-q) \le \frac{1}{4}\log K. \tag{A.11}$$

Proof. First assume condition (A.10) holds. We call a node a *disconnected node* if it is not connected to any other node in its own clusters. Let E be the event that there exist two disconnected nodes from two different clusters. Suppose Y^* is uniformly distributed over $\mathcal{Y}$ and let $\rho := \mathbb{P}[E]$. We claim that $\mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \rho/2$. To see this, consider the ML estimate of Y^* given by $\widehat{Y}_{\text{ML}}(a) := \arg\max_y \mathbb{P}[A = a|Y^* = y]$ with tie broken uniformly at random. It is a standard fact that the ML estimator minimizes the error probability under the uniform prior, so for all $\widehat{Y}$ we have

$$\mathbb{P}\left[\widehat{Y} \neq Y^*\right] \ge \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\sum_{a\in\{0,1\}^{n\times n}} \mathbb{P}\left[\widehat{Y}_{\text{ML}}(a) \neq y\right]\mathbb{P}[A = a|Y^* = y]. \tag{A.12}$$

Let $\mathcal{A}_y \subseteq 0, 1^{n\times n}$ denote the set of adjacency matrices with at least two disconnected nodes 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 obtain by swapping the two rows and columns of y corresponding to the two disconnected nodes in a. It is easy to check that for each $a \in \mathcal{A}_y$, the likelihood satisfies $\mathbb{P}[A = a|Y^* = y] \le \mathbb{P}[A = a|Y^* = y'(a)]$ and therefore $\mathbb{P}[\widehat{Y}_{\text{ML}}(a) \neq y] \ge 1/2$. It follows from (A.12) that

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

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

Since the minimax error probability is lower bounded by the average error probability, it suffices to show $\rho \ge 1/2$. Without loss of generality, suppose r is even and the first $rK/2$ nodes $i \in [rK/2]$ form $r/2$ clusters. For each $i \in [rK/2]$, let ξ_i be the indicator random variable for node i being a disconnected node. Then $\rho_1 := \mathbb{P}\left[\sum_{i=1}^{rK/2}\xi_i \ge 1\right]$ is the probability that there exists at least one disconnected node among the first $rK/2$ nodes. We use a second moment argument [90] to lower-bound ρ_1. Observe that $\xi_1, \ldots, \xi_{rK/2}$ are (possibly dependent) Bernoulli

variables with mean $\mu = (1-p)^{K-1}$. For $i \neq j$, we have

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

Therefore, we have

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

On the other hand, by the assumptions (A.10) we have $p \leq 1/8$ and

$$\mu = (1-p)^{K-1} \overset{(a)}{\geq} e^{-2(K-1)p} \geq (rK/2)^{-1/4}, \tag{A.13}$$

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

$$\mathbb{P}\left[\left|\sum_{i=1}^{rK/2} \xi_i - rK\mu/2\right| \geq rK\mu/2\right] \leq \frac{\frac{1}{2}rK\mu + \frac{1}{4}(rK\mu)^2\frac{p}{1-p}}{r^2K^2\mu^2/4} \leq \frac{2}{rK\mu} + \frac{p}{1-p} \leq \frac{1}{4}, \tag{A.14}$$

where the last inequality holds due to (A.13) and $p \leq 1/8$. It follows that $\rho_1 \geq \frac{3}{4}$. If we let ρ_2 denote the probability that there exists a disconnected node among the next $rK/2$ nodes $rK/2+1, \ldots, rK$, then by symmetry $\rho_2 \geq \frac{3}{4}$. Therefore $\rho = \rho_1\rho_2 \geq 1/2$, proving the sufficiency of (A.10).

We next assume the condition (A.11) holds and bound 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 at least one betrayed node in each of cluster 1 and 2, and let $\mathbb{P}[E'] := \rho'$. Suppose Y^* is uniformly distributed over $\mathcal{Y}$; again we can show that $\mathbb{P}[\widehat{Y} \neq Y^*] \geq \rho'/2$ for any $\widehat{Y}$. Suppose cluster 1 is formed by nodes $i \in [K]$. For each $i \in [K]$, let ξ_i' be the indicator for node i being a betrayed node. Then $\rho_1' := \mathbb{P}\left[\sum_{i=1}^K \xi_i' > 0\right]$ is the probability having a betrayed node in cluster 1. We have

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

where (a) follows from (A.11) and $q^K = (1-(1-q))^K \geq \exp(-2(1-q)K)$ since $1-q \leq 1/2$. Let ρ_2' be the probability of having a betrayed node in cluster 2 and by symmetry $\rho_2' \geq 3/4$. We thus get $\rho' = \rho_1'\rho_2' \geq 1/2$, proving the sufficiency of (A.11). □

We are ready to prove Theorem 3.1.1 by combining the above three lemmas.

Proof of Theorem 3.1.1. Since $256 \leq 2K \leq n$, we have the following relations between the log terms:

$$\log(n-K) \geq \log(n/2) \geq \frac{1}{2}\log n,\ \log(rK/2) \geq \frac{1}{2}\log(rK). \tag{A.15}$$

Our goal is to show that if condition (3.1) or (3.2) holds, then the minimax error probability is large.

First assume (3.1) holds. By (A.3) we know condition (3.1) implies

$$K(p-q)^2 \leq \frac{1}{96}p(1-q)\left(\log(rK) \wedge K\right). \tag{A.16}$$

(i) If $p \leq 2q$, then (A.16) implies $K(p-q)^2 \leq \frac{1}{48}q(1-q)\log(rK)$; it follows from (A.1) and (A.15) that $KD(p\|q) \leq \frac{1}{48}\log(rK) \leq \frac{1}{24}\log(n-K)$ and thus Lemma 3 proves the conclusion. (ii) If $p > 2q$, (A.16) 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 4 proves the conclusion.

Next assume the condition (3.2) holds. By the lower-bound (A.3) on the KL divergence, we know (3.2) implies

$$K(p-q)^2 \leq \frac{1}{96}p(1-q)\log n. \tag{A.17}$$

(i) If $1-q \leq 2(1-p)$, then (A.17) implies that $K(p-q)^2 \leq \frac{1}{48}p(1-p)\log n$; it follows from (A.1) and (A.15) that $KD(q\|p) \leq \frac{1}{48}\log n \leq \frac{1}{24}\log(n-K)$ and thus Lemma 3 implies the conclusion. (ii) If $1-q > 2(1-p)$ then (A.17) implies

$$K(1-q) \leq \frac{1}{24}\log n \leq \frac{1}{12}\max\left\{\log\frac{n}{K}, \log K\right\}. \tag{A.18}$$

We divided the analysis into two subcases.

Case $(ii.1)$: $K \geq \log n$. It follows from (A.18) that $1-q \leq \frac{1}{24}$, i.e., $q \geq \frac{23}{24}$ and thus $(p-q)^2 \leq 2q(1-q)^2$. Therefore, (A.18) implies either the condition (A.5) in Lemma 2 or the condition (A.11) in Lemma 4, which proves the conclusion.

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

$$\frac{n^2(q-\bar{p})^2}{n_1\bar{p}(1-\bar{p})} = \frac{n^2\delta^2(p-q)^2}{n_1\bar{p}(1-\bar{p})} \leq \frac{2n^2\delta(p-q)^2}{n_1 p(1-q)} \overset{(a)}{\leq} 4KD(p\|q) \overset{(b)}{\leq} \frac{1}{24}\log\frac{n}{K}, \tag{A.19}$$

where we use (A.3) in (a) and (3.2) in (b). On the other hand, we have

$$\begin{aligned} D(p\|\bar{p}) &= p\log\frac{p}{\bar{p}} + (1-p)\log\frac{1-p}{1-\bar{p}} \leq p\log\frac{p}{q} + (1-p)\log\frac{10(1-p)}{9(1-q)} \\ &\leq D(p\|q) + (1-q)\log\frac{10}{9} \leq \frac{1}{6K}\log\frac{n}{K}, \end{aligned} \tag{A.20}$$

where the last inequality follows from (3.2) and (A.18). Equations (A.19) and (A.20) imply assumption (A.4) in Lemma 2, and therefore the conclusion follows. □

A.1.1 Proof of Corollary 3.1.2

The corollary is derived from Theorem 3.1.1 using the upper bound (A.1) on the KL divergence. In particular, Condition (3.3) in the corollary implies Condition (3.2) in Theorem 3.1.1 in view of (A.1). Similarly, Condition (3.4) implies Condition (3.1) because $D(q\|p) \leq \frac{p}{1-p}$ in view of (A.1) and $p \leq \frac{1}{193}$; Condition (3.5) implies Condition (3.2) because $D(p\|q) \leq p\log\frac{p}{q}$ by definition.

A.2 Proof of Theorem 3.2.1 and Corollary 3.2.2

For any feasible solution $Y \in \mathcal{Y}$ of (3.7), we define $\Delta(Y) \triangleq \langle A, Y^* - Y\rangle$ and $d(Y) \triangleq \langle Y^*, Y^* - Y\rangle$. To prove the theorem, it suffices to show that $\Delta(Y) > 0$ for all feasible Y with $Y \neq Y^*$. 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:

$$\Delta(Y) = \langle \mathbb{E}[A], Y^* - Y \rangle + \langle A - \mathbb{E}[A], Y^* - Y \rangle = (p-q)d(Y) + \langle A - \mathbb{E}[A], Y^* - Y \rangle, \tag{A.21}$$

where the second equality follows from $\sum_{i,j} Y_{ij} = \sum_{i,j} Y^*_{ij}$ and $\sum_{i,i} Y_{ii} = \sum_{i,i} Y^*_{i,i}$ by feasibility of Y. For the second term in (A.21), 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 $Y \in \mathcal{Y}$:

$$\mathbb{P}\left\{T_1(Y) \le -\frac{p-q}{4}d(Y)\right\} \le \exp\left(-\frac{1}{2}d(Y)D\left(\frac{p+q}{2}\middle\|p\right)\right)$$
$$\mathbb{P}\left\{T_2(Y) \ge \frac{p-q}{4}d(Y)\right\} \le \exp\left(-\frac{1}{2}d(Y)D\left(\frac{p+q}{2}\middle\|q\right)\right)$$

We need to control the perturbation uniformly over $Y \in \mathcal{Y}$. Define the equivalence class $[Y] = \{Y' \in \mathcal{Y} : Y'_{ij} = Y_{ij}, \forall (i,j) \text{ s.t. } Y^*_{ij} = 1\}$. Notice that all cluster matrices in the equivalence class $[Y]$ have the same value $T_1(Y)$. 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 any feasible $Y \ne Y^*$.

Lemma 5. *For each integer $t \in [K, rK^2]$, we have*

$$|\{Y \in \mathcal{Y} : d(Y) = t\}| \le \left(\frac{4t}{K}\right)^2 n^{16t/K},$$
$$|\{[Y] : d(Y) = t\}| \le \frac{4t}{K}(rK)^{8t/K}.$$

We also need the following lemma to upper bound $D\left(\frac{p+q}{2}\|q\right)$ and $D\left(\frac{p+q}{2}\|p\right)$ using $D\left(p\|q\right)$ and $D\left(q\|p\right)$, respectively.

Lemma 6.

$$D\left(\frac{p+q}{2}\Big\|q\right) \geq \frac{1}{36}D\left(p\|q\right) \tag{A.22}$$

$$D\left(\frac{p+q}{2}\Big\|p\right) \geq \frac{1}{36}D\left(q\|p\right) \tag{A.23}$$

We prove the lemmas in the next subsection. Using the union bound and Lemma 5 and Lemma 6, we obtain

$$\begin{aligned}
&\mathbb{P}\left\{\exists[Y] : Y \neq Y^*, T_1(Y) \leq -\frac{p-q}{4}d(Y)\right\} \\
\leq &\sum_{t=K}^{rK^2} \mathbb{P}\left\{\exists[Y] : d(Y) = t, T_1(Y) \leq -\frac{p-q}{4}t\right\} \\
\leq &\sum_{t=K}^{rK^2} |\{\exists[Y] : d(Y) = t\}| \mathbb{P}\left\{T_1(Y) \leq -\frac{p-q}{4}t\right\} \\
\leq &\sum_{t=K}^{rK^2} \frac{4t}{K}(rK)^{8t/K} \exp\left(-\frac{1}{72}tD(q\|p)\right) \\
\overset{(a)}{\leq} &4\sum_{t=K}^{rK^2} (rK)(rK)^{-4t/K} \leq 4(rK)^{-1},
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left\{\exists Y \in \mathcal{Y} : Y \neq Y^*, T_2(Y) \geq \frac{p-q}{4}d(Y)\right\} \\
\leq &\sum_{t=K}^{rK^2} \mathbb{P}\left\{\exists Y \in \mathcal{Y} : d(Y) = t, T_2(Y) \geq \frac{p-q}{4}t\right\} \\
\leq &\sum_{t=K}^{rK^2} |\{Y \in \mathcal{Y} : d(Y) = t\}| \cdot \mathbb{P}\left\{T_2(Y) \geq \frac{p-q}{4}t\right\} \\
\leq &\sum_{t=K}^{rK^2} \frac{16t^2}{K^2} n^{16t/K} \cdot \exp\left(-\frac{1}{72}tD(p\|q)\right) \overset{(a)}{\leq} 16n^{-1},
\end{aligned}$$

where (a) follows from the theorem assumption that $D(p\|q) \geq c_1 \log(n)/K$ for a

large constant c_1. Combining the above two bounds with (A.21), we obtain

$$\mathbb{P}\left\{\exists Y \in \mathcal{Y} : \Delta(Y) \leq 0\right\} \leq 4(rK)^{-1} + 16n^{-1} \tag{A.24}$$

and thus Y^* is the unique optimal solution with high probability. This proves the theorem.

A.2.1 Proof of Lemma 5

Let $C_1^*, \ldots, C_r^*$ denote the true clusters associated with Y^*. Let V denote the set of nodes. Recall that the nodes in V which do not belong to any clusters are called outlier nodes.

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 as follows:

1. Let $C_{r+1} := \{i : Y_{ij} = 0, \forall j\}$.
2. The nodes in $V \setminus C_{r+1}$ are further partitioned into r new clusters of size K, such that nodes i and i' are in the same cluster if and only if the i-th and i'-th rows of Y are identical. We now define an ordering $C_1, \ldots, C_r$ of these r new clusters in the following manner.
   (a) For each new cluster C, if there exists a $k \in [r]$ such that $|C \cap C_k^*| > K/2$, then we label this new cluster as C_k; this label is unique because the cluster size is K.
   (b) The remaining unlabeled clusters are labeled arbitrarily.

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

(A0) $(C_1, \ldots, C_r, C_{r+1})$ is a partition of V with $|C_k| = K$ for all $k \in [r]$;

(A1) For each $k \in [r]$, exactly one of the following is true: (1) $\alpha_{kk} > K/2$; (2) $\alpha_{kk'} \leq K/2$ for all $k' \in [r]$;

(A2) We have

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

here and henceforth, all the 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 identity:

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

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:

$$|\{Y \in \mathcal{Y} : d(Y) = t\}| \leq |\{(C_1, \ldots, C_{r+1}) : \text{it satisfies (A0)–(A2)}\}|. \quad \text{(A.25)}$$

It remains to upper-bound the right hand side of (A.25).

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

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

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

- If $\alpha_{11} \leq K/4$, then $m_1 \geq 3K/4$, and we must also have $\alpha_{1k'} \leq K/2$ for all

$1 \le k' \le r$ by Property (A1). Hence,

$$\begin{aligned}
&\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\alpha_{1k''} + \alpha_{1(r+1)}(\alpha_{1(r+1)} - 1) \\
\ge &\sum_{(k',k''):k'\neq k''} \mathbf{1}\{k' \neq 1\}\,\mathbf{1}\{k'' \neq 1\}\,\alpha_{1k'}\alpha_{1k''} + \alpha_{1(r+1)}(\alpha_{1(r+1)} - 1) \\
= &m_1^2 - \sum_{2\le k'\le r} \alpha_{1k'}\alpha_{1k'} - \alpha_{1(r+1)} \ge m_1^2 - \frac{1}{2}Km_1 \ge \frac{1}{4}m_1K.
\end{aligned}$$

Combining the above two cases, we conclude that we always have

$$\sum_{(k',k''):k'\neq k''} \alpha_{1k'}\alpha_{1k''} + \alpha_{1(r+1)}(\alpha_{1(r+1)} - 1) \ge \frac{1}{4}m_1K.$$

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

$$t = \sum_{k=1}^{r} \left\{ \alpha_{k(r+1)}(\alpha_{k(r+1)} - 1) + \sum_{(k',k''):k'\neq k''} \alpha_{kk'}\alpha_{kk''} \right\} \ge \frac{K}{4}\sum_{k=1}^{r} m_k.$$

In other words, we have $\sum_{k\in[r]} m_k \le 4t/K$, i.e., the total number of misclassified non-outlier nodes is upper bounded by $4t/K$. It implies that the total number of misclassified outlier nodes is also upper bounded by $4t/K$, because by the cluster size constraint in Property (A0), one misclassified outlier node must produce one misclassified non-outlier node.

We are ready to upper-bound the right hand side of (A.25). Fix a Y with $d(Y) = t$, let N_1 denote the total number of misclassified non-outlier nodes and N_2 denote the total number of misclassified outlier nodes. Since $N_1, N_2 \le 4t/K$, there are at most $(4t/K)^2$ different choices for the value of the pair (N_1, N_2). Moreover, for a fixed (N_1, N_2), there are at most $\binom{n_1}{N_1}\binom{n_2}{N_2} \le n^{8t/k}$ different ways to choose these misclassified nodes. Each misclassified non-outlier node can then be assigned to one of $r - 1 \le n$ different clusters or left as outlier, and each misclassified outlier node can be assigned to one of $r \le n$ different clusters. Hence, the right hand side of (A.25) is upper bounded by $\left(\frac{4t}{K}\right)^2 n^{16t/K}$. This proves the first part of the lemma.

To count the number of possible equivalence classes $[Y]$, we use a similar ar-

gument but only need to consider the misclassified *non-outlier* nodes. Note that N_1 can take at most $4t/K$ different values. For a fixed N_1, there are at most $\binom{rK}{N_1} \leq (rK)^{N_1}$ different ways to choose these misclassified non-outlier nodes. Each misclassified non-outlier node then can be assigned to one of $r-1$ different clusters or leave outlier. Therefore, the number of possible equivalence classes $[Y]$ with $d(Y) = t$ is upper bounded by $\frac{4t}{K}(rK)^{8t/K}$.

A.2.2 Proof of Lemma 6

We first prove (A.22). Note that if $u \geq v$, then

$$D\left(u\|v\right) = u\log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v} \leq u\log\frac{u}{v} \tag{A.26}$$

$$D\left(u\|v\right) \geq u\log\frac{u}{v} + (1-u)\log(1-u) \overset{(a)}{\geq} u\log\frac{u}{\mathrm{e}v}, \tag{A.27}$$

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

Case 1: $p \leq 8q$. In view of (A.1) and (A.3), $D\left(p\|q\right) \leq \frac{(p-q)^2}{q(1-q)}$ and $D\left(\frac{p+q}{2}\|q\right) \geq \frac{(p-q)^2}{4(p+q)(1-q)}$. Since $p \leq 8q$, it follows that $D\left(\frac{p+q}{2}\|q\right) \geq \frac{(p-q)^2}{36q(1-q)} \geq \frac{1}{36}D\left(p\|q\right)$.

Case 2: $p > 8q$. In view of (A.26) and (A.27), $D\left(p\|q\right) \leq p\log\frac{p}{q}$ and $D\left(\frac{p+q}{2}\|q\right) \geq \frac{p+q}{2}\log\frac{p+q}{2\mathrm{e}q}$. Since $p > 8q$ and $8 > \mathrm{e}^2$, it follows that $\log\frac{p}{q} > \frac{6}{5}\log(2\mathrm{e})$ and thus $D\left(\frac{p+q}{2}\|q\right) \geq \frac{p}{12}\log\frac{p}{q} \geq \frac{1}{12}D\left(p\|q\right)$.

We next prove (A.23). Similar to (A.26) and (A.27), if $u \leq v$, then

$$(1-u)\log\frac{1-u}{\mathrm{e}(1-v)} \leq D\left(u\|v\right) \leq (1-u)\log\frac{1-u}{1-v}. \tag{A.28}$$

We also divide the analysis into two cases:

Case 1: $1-q \leq 8(1-p)$. In view of (A.1) and (A.3), $D\left(q\|p\right) \leq \frac{(p-q)^2}{p(1-p)}$ and $D\left(\frac{p+q}{2}\|p\right) \geq \frac{(p-q)^2}{4p(2-p-q)}$. Since $1-q \leq 8(1-p)$, it follows that $D\left(\frac{p+q}{2}\|p\right) \geq \frac{(p-q)^2}{36p(1-p)} \geq \frac{1}{36}D\left(q\|p\right)$.

Case 2: $1-q < 8(1-p)$. In view of (A.28), $D\left(q\|p\right) \leq (1-q)\log\frac{1-q}{1-p}$ and $D\left(\frac{p+q}{2}\|p\right) \geq \left(1-\frac{p+q}{2}\right)\log\frac{2-p-q}{2\mathrm{e}(1-p)}$. Since $1-q < 8(1-p)$ and $8 > \mathrm{e}^2$, it follows that $\log\frac{1-q}{1-p} > \frac{6}{5}\log(2\mathrm{e})$ and thus $D\left(\frac{p+q}{2}\|p\right) \geq \frac{1}{12}(1-q)\log\frac{1-q}{1-p} \geq \frac{1}{12}D\left(q\|p\right)$.

A.2.3 Proof of Corollary 3.2.2

The corollary is derived from Theorem 3.2.1 using the lower bound (A.3) on the KL divergence. In particular, first assume $e^2 q \geq p$. Then $K(p-q)^2 \gtrsim q(1-q)\log n$ implies condition (3.9) in view of (A.3). Next assume $e^2 q < p$. 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 (A.3) and $p > e^2 q$. Therefore, $Kp \gtrsim \log(rK)$ implies $KD(q\|p) \gtrsim \log(rK)$.

A.3 Proof of Theorem 3.3.1

Our proof only relies on the standard concentration results for the adjacency matrix A (see Proposition A.3.1 below). Let $U \in \mathbb{R}^{n \times r}$ be the normalized characteristic matrix for the clusters, i.e.,

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

The true cluster matrix Y^* has the rank-r singular value decomposition given by $Y^* = KUU^\top$. Define the projections $\mathcal{P}_T(M) = UU^\top M + MUU^\top - UU^\top MUU^\top$ and $\mathcal{P}_{T^\perp}(M) = M - \mathcal{P}_T(M)$. Let $\nu \triangleq p-q$ and $\bar{A} \triangleq qJ + (p-q)Y^*$, where J is the all-ones matrix. The proof hinges on the following concentration property of the random matrix $A - \bar{A}$.

Proposition A.3.1. *Under the condition* (3.14)*, the following holds with probability at least* $1 - n^{-10}$*:*

$$\|A - \bar{A}\| \leq \frac{1}{8}\nu K, \tag{A.29}$$

$$\|\mathcal{P}_T(A - \bar{A})\|_\infty \leq \frac{1}{8}\nu. \tag{A.30}$$

We prove the proposition in Section A.3.1 to follow. In the rest of the proof we assume (A.29) and (A.30) 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

$$\begin{aligned}\langle Y^* - Y, A\rangle &= \langle \bar{A}, Y^* - Y\rangle + \langle A - \bar{A}, Y^* - Y\rangle \\ &\overset{(a)}{=} \nu\langle Y^*, Y^* - Y\rangle + \langle A - \bar{A}, Y^* - Y\rangle \\ &\overset{(b)}{=} \frac{\nu}{2}\|Y^* - Y\|_1 + \langle A - \bar{A}, Y^* - Y\rangle, \end{aligned} \tag{A.31}$$

where (a) holds in view of the definition of $\bar{A}$ and the fact that $\sum_{i,j} Y_{ij} = \sum_{i,j} Y^*_{ij}$; (b) holds because $Y_{ij} \in [0,1], \forall i,j$.

Let $W \triangleq \frac{8(A-\bar{A})}{\nu K}$. By (A.29) we have $\|\mathcal{P}_{T^\perp}(W)\| \leq \|W\| \leq 1$, so $UV^\top + \mathcal{P}_{T^\perp}(W)$ is a subgradient of $||X||_*$ at $X = Y^*$. It follows that

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

Since $\|Y^*\|_* \geq \|Y\|_*$, by rearranging terms and using the definition of W, we get

$$\left\langle A - \bar{A}, Y^* - Y\right\rangle = \frac{\nu K}{8}\langle W, Y^* - Y\rangle \geq \frac{\nu K}{8}\left\langle -UV^\top + \mathcal{P}_T(W), Y^* - Y\right\rangle. \tag{A.32}$$

Assembling (A.31) and (A.32), we obtain that for any feasible Y,

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

where the last inequality follows from the duality between ℓ_1 and ℓ_∞ norms. Using (A.30) and the fact that $\|UV^\top\|_\infty = \frac{1}{K}$, we get

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

which is positive for all $Y \neq Y^*$. This completes the proof.

A.3.1 Proof of Proposition A.3.1

We first prove (A.30). By definition of $\mathcal{P}_T$, we have

$$\begin{aligned}
&\|\mathcal{P}_T(A-\bar{A})\|_\infty \\
&\le \|UU^\top(A-\bar{A})\|_\infty + \|(A-\bar{A})UU^\top\|_\infty + \|UU^\top(A-\bar{A})UU^\top\|_\infty \\
&\le 3\max\left(\|UU^\top(A-\bar{A})\|_\infty, \|(A-\bar{A})VV^\top\|_\infty\right). \qquad \text{(A.33)}
\end{aligned}$$

Suppose node i is from cluster k. Then

$$(UU^\top(A-\bar{A}))_{ij} = \frac{1}{K}\sum_{l\in C_k^*}(A-\bar{A})_{lj} = \frac{1}{K}\sum_{l\in C_k^*}(A-\mathbb{E}A)_{lj} + \frac{1}{K}\sum_{l\in C_k^*}(\mathbb{E}A-\bar{A})_{lj}. \qquad \text{(A.34)}$$

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 (A.34) is the average of K such random variables; by Bernstein's inequality, with probability at least $1-n^{-13}$ and for some universal constant c_2,

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

where the last inequality follows because $Kp(1-q) > c_1\log n$ in view of the condition (3.14). By definition of $\bar{A}$, $\mathbb{E}[A]-\bar{A}$ is a diagonal matrix with diagonal entries equal to $-p$ or $-q$, so the second term of (A.34) has magnitude at most $1/K$. By the union bound over all (i,j) and substituting back to (A.33), we have with probability at least $1-2n^{-11}$,

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

where the last inequality follows from the condition (3.14). This proves (A.30) in the proposition.

We now turn to the proof of (A.29) in the proposition. Note that $\|A-\bar{A}\| \le \|A-\mathbb{E}[A]\| + \|\bar{A}-\mathbb{E}[A]\| \le \|A-\mathbb{E}[A]\| + 1$. Under the condition (3.14), $Kp(1-q) \ge c_1\log n$. The spectral norm term is bounded below.

Lemma 7. *If $Kp(1-q) \ge c_1\log n$, then there exists some universal constant c_4 such that $\|A-\mathbb{E}[A]\| \le c_4\sqrt{p(1-q)K\log n + q(1-q)n}$ with probability at least $1-n^{-10}$.*

We prove the lemma in Section A.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 (3.14).

A.3.2 Proof of Lemma 7

Let $R := \text{support}(Y^*)$ and $\mathcal{P}_R(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be the operator which sets the entries outside of R to zero. Let $B_1 = \mathcal{P}_R(A - \mathbb{E}[A])$ 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 [91] and using the assumption that $Kp(1-q) \geq c_1 \log n$ in the lemma, we get that there exists 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 [72] 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 [92] implies that $\|B_2\| \leq c_8\sigma\sqrt{n}$ with probability at least $1 - n^{-11}$ for some universal constant $c_8 \geq 3$. (See Lemma 8 in [93] for a similar derivation.) It follows 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.

A.4 Proof of Theorem 3.3.2

Observe that if any feasible solution Y has the same support as Y^*, then the constraint (3.13) implies that Y must be exactly equal to Y^*. Therefore, it suffices to show that Y^* is not an optimal solution.

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

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

Thus, $Kp < 8c_2^2$ which contradicts the assumption that $Kp > qn \geq c_1 \log n$. Therefore, $Kp > qn$ cannot hold. Hence, 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 A.3, and that $Y^* = KUU^\top$ is the SVD of Y^*. Consider the Lagrangian

$$\begin{aligned} L(Y;\lambda,\mu,F,G) \triangleq &- \langle A, Y\rangle + \lambda\left(\|Y\|_* - \|Y^*\|_*\right) + \eta\left(\langle J, Y\rangle - rK^2\right) \\ &- \langle F, Y\rangle + \langle G, Y - J\rangle, \end{aligned}$$

where the Lagrangian multipliers are $\lambda, \eta \in \mathbb{R}$ and $F, G \in \mathbb{R}^{n\times n}$. 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 $\|M - \mathcal{P}_T(M)\| \leq 1$. Let $H = F - G$; the KKT conditions

imply that there exist some numbers $\lambda \geq 0$, $\eta \in \mathbb{R}$ and matrices W, H obeying

$$A - \lambda \left(UU^\top + W\right) - \eta J + H = 0; \tag{A.35}$$

$$\mathcal{P}_T W = 0; \qquad \|W\| \leq 1; \tag{A.36}$$

$$H_{ij} \leq 0, \ \forall (i,j) \in \mathcal{R}; \qquad H_{ij} \geq 0, \ \forall (i,j) \in \mathcal{R}^c. \tag{A.37}$$

Now observe that $UU^\top W UU^\top = 0$ by (A.36). We left and right multiply (A.35) by $UU^\top$ to obtain

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

where for any $X \in \mathbb{R}^{n \times n}$, $\bar{X} := UU^\top X UU^\top$ 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. By the Bernstein inequality for each entry $\bar{A}_{ij}$, we have with probability at least $1 - 2n^{-11}$,

$$p - \frac{\lambda}{K} - \eta + \bar{H}_{ij} \geq -\frac{c_3\sqrt{p(1-p)\log n}}{K} - \frac{c_4 \log n}{2K^2} \overset{(a)}{\geq} -\frac{\epsilon_0}{8}, \quad \forall (i,j) \in \mathcal{R} \tag{A.38}$$

$$q - \eta + \bar{H}_{ij} \leq \frac{c_3\sqrt{q(1-q)\log n}}{K} + \frac{c_4 \log n}{2K^2} \overset{(b)}{\leq} \frac{\epsilon_0}{8}, \quad \forall (i,j) \in \mathcal{R}^c \tag{A.39}$$

for some universal constants $c_3, c_4 > 0$, where (a) and (b) follow from the assumption $K \geq c_1 \log n$ with the universal constant c_1 sufficiently large. In the rest of the proof, we assume (A.38) and (A.39) hold. Using (A.37), we get that

$$\begin{aligned} \eta &\geq q - \frac{c_3\sqrt{q(1-q)\log n}}{K} - \frac{c_4 \log n}{2K^2} \geq q - \frac{\epsilon_0}{8} \\ \eta &\leq p + \frac{c_3\sqrt{p(1-p)\log n}}{K} + \frac{c_4 \log n}{2K^2} - \frac{\lambda}{K} \leq p + \frac{\epsilon_0}{8} - \frac{\lambda}{K}. \end{aligned} \tag{A.40}$$

It follows that

$$\begin{aligned} \lambda &\leq K(p-q) + c_3(\sqrt{p(1-p)\log n} + \sqrt{q(1-q)\log n}) + \frac{c_4 \log n}{K} \\ &\leq 4 \max\left\{ K(p-q), c_3\sqrt{p(1-p)\log n}, c_3\sqrt{q(1-q)\log n}, \frac{c_4}{c_1} \right\}. \end{aligned} \tag{A.41}$$

On the other hand, (A.36) and (A.35) 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 (A.40), $\lambda \geq 0$ and the assumption $p \leq 1-\epsilon_0$, we obtain $\eta \leq 1-\frac{7}{8}\epsilon_0$ and therefore

$$\lambda^2 \geq \frac{49}{64n}\epsilon_0^2 \sum_{(i,j)\in\mathcal{R}^c} A_{ij}. \tag{A.42}$$

Note that $\sum_{(i,j)\in\mathcal{R}^c} A_{ij}$ equals two times 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 $qn \geq c_1 \log n$, 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 (A.42) that $\lambda^2 \geq \frac{1}{2}\epsilon_0^2 c_5 qn$. Combining with (A.41) 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 assumption sufficiently small such that $2c_2^2 < \frac{1}{32}\epsilon^2 c_5$, we have $K(p-q) > c_2\sqrt{2qn}$, which leads to the contradiction. This completes the proof of the theorem.

A.5 Proof of Theorem 4.2.1

Given G generated either under $\mathcal{G}(N,q)$ or $\mathcal{G}(N,2K,p,q)$, we obtain a sequence of N graphs $G_1,\ldots,G_N$ by each time picking a vertex (without replacement) in any arbitrary order and replacing it with a new vertex that connects to all other vertices independently at random with probability q. We run the given algorithm $\mathcal{A}$ on $G_1,\ldots,G_N$ and let $S_1,\ldots,S_N$ denote the outputs which are sets of K vertices. Let $E(S_i,S_i)$ denote the total number of edges in S_i and $\tau = q+(1-\epsilon)^2(p-q)/2$. Define a test $\phi : G \to \{0,1\}$ such that $\phi(G) = 1$ if $\max_{i\in[N]} E(S_i,S_i) > \tau\binom{K}{2}$; otherwise $\phi(G) = 0$. The construction of each G_i takes N time units; the running time of $\mathcal{A}$ on G_i is at most $T(N)$ time units; the computation of $E(S_i,S_i)$ takes at most K^2 time units. Therefore, the total running time of ϕ is at most $N^2 + NT(n) + NK^2$. Next we upper bound the Type-I+II error probabilities. Let C denote the positive universal constant whose value may change line by line.

If $G \sim \mathcal{G}(N, q)$, by the union bound and the Bernstein inequality,

$$\begin{aligned}\mathbb{P}_0\{\phi(G) = 1\} &\le \sum_{i=1}^{N} \mathbb{P}_0\left\{E(S_i, S_i) \ge \tau\binom{K}{2}\right\} \\ &\le N\left(-\frac{\binom{K}{2}^2(1-\epsilon)^4(p-q)^2/4}{2\binom{K}{2}q + \binom{K}{2}(1-\epsilon)^2(p-q)/3}\right) \\ &\le N\exp(-CK^2q).\end{aligned}$$

If $G \sim \mathcal{G}(N, 2K, p, q)$, let S denote the planted cluster. Then $|S| \sim \mathrm{Binom}(N, \frac{2K}{N})$ and by the Chernoff bound, $\mathbb{P}_1[|S| < K] \le \exp(-CK)$. If $|S| = K' \ge K$, then there must exist some $I \in [N]$ such that G_I is distributed exactly as the planted cluster model with a single cluster S^* of size K and $p = cq$; conditional on $I = i$ and the success of $\mathcal{A}$ on G_i, $|S_i \cap S^*| \ge (1-\epsilon)K$ and $S^* \sim \mathrm{Binom}(\binom{K}{2}, p)$. Therefore, by the Bernstein inequality

$$\begin{aligned}&\mathbb{P}_1\left\{E(S_i, S_i) < \frac{p+q}{2}\binom{K}{2}\middle| |S| = K', I = i\right\} \\ &\le \eta_N + \left(-\frac{\binom{K}{2}^2(1-\epsilon)^4(p-q)^2/4}{2\binom{K}{2}p + \binom{K}{2}(1-\epsilon)^2(p-q)/3}\right) \\ &\le \eta_N + \exp(-CK^2q).\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{P}_1\{\phi(G) = 0\} \\ &\le \mathbb{P}_1\{|S| < K\} + \sum_{K' \ge K}\sum_{i=1}^{N} \mathbb{P}_1\{|S| = K', I = i\}\mathbb{P}_1\{\phi(G) = 0 || S| = K', I = i\} \\ &\le \exp(-CK) + \eta_N + \exp(-CK^2q).\end{aligned}$$

A.6 Proof of Proposition 4.3.1

We first introduce several key auxiliary results used in the proof. The following lemma ensures that $P'_{\ell_s\ell_t}$ and $Q'_{\ell_s\ell_t}$ are well-defined under suitable conditions and that $P'_{\ell_s\ell_t}$ and P_{ℓ_s,ℓ_t} are close in total variation.

Lemma 8. *Suppose that $p = 2q$ and $16q\ell^2 \le 1$. Fix $\{\ell_t\}$ such that $\ell_t \le 2\ell$ for all*

$t \in [k]$. *Then for all* $1 \le s < t \le k$, $P'_{\ell_s\ell_t}$ *and* $Q'_{\ell_s\ell_t}$ *are probability measures and*

$$d_{\mathrm{TV}}(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}) \le 4(8q\ell^2)^{(m_0+1)}.$$

Proof. Fix an (s,t) such that $1 \le s < t \le k$. We first show that $P'_{\ell_s\ell_t}$ and $Q'_{\ell_s\ell_t}$ are well-defined. By definition, $\sum_{m=0}^{\ell_s\ell_t} P'_{\ell_s\ell_t}(m) = \sum_{m=0}^{\ell_s\ell_t} Q'_{\ell_s\ell_t}(m) = 1$ and it suffices to show positivity, i.e.,

$$P_{\ell_s\ell_t}(0) + a_{\ell_s\ell_t} \ge 0, \tag{A.43}$$

$$Q_{\ell_s\ell_t}(m) \ge \gamma P'_{\ell_s\ell_t}(m), \quad \forall 0 \le m \le m_0. \tag{A.44}$$

Recall that $P_{\ell_s\ell_t} \sim \mathrm{Binom}(\ell_s\ell_t, p)$ and $Q_{\ell_s\ell_t} \sim \mathrm{Binom}(\ell_s\ell_t, q)$. Hence, for $\forall 0 \le m \le \ell_s\ell_t$,

$$Q_{\ell_s\ell_t}(m) = \binom{\ell_s\ell_t}{m} q^m (1-q)^{\ell_s\ell_t - m},\ P_{\ell_s\ell_t}(m) = \binom{\ell_s\ell_t}{m} p^m (1-p)^{\ell_s\ell_t - m}.$$

It follows that

$$\frac{1}{\gamma} Q_{\ell_s\ell_t}(m) - P_{\ell_s\ell_t}(m) = \frac{1}{\gamma}\binom{\ell_s\ell_t}{m} q^m (1-2q)^{\ell_s\ell_t - m} \left[\left(\frac{1-q}{1-2q} \right)^{\ell_s\ell_t - m} - 2^m\gamma \right].$$

Recall that $m_0 = \lfloor \log_2(1/\gamma) \rfloor$ and thus $Q_{\ell_s\ell_t}(m) \ge \gamma P_{\ell_s\ell_t}(m)$ for all $m \le m_0$. Furthermore,

$$Q_{\ell_s\ell_t}(0) = (1-q)^{\ell_s\ell_t} \ge (1 - q\ell_s\ell_t) \ge 1 - 4q\ell^2 \ge \frac{3}{4} \ge \gamma \ge \gamma P'_{\ell_s\ell_t}(0),$$

and thus (A.44) holds. Recall that

$$a_{\ell_s\ell_t} = \sum_{m_0 < m \le \ell_s\ell_t} \left(P_{\ell_s\ell_t}(m) - \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) \right).$$

Since $2^{m_0+1}\gamma > 1$ and $8q\ell^2 \le 1/2$, it follows that

$$\frac{1}{\gamma} \sum_{m_0 < m \le \ell_s\ell_t} Q_{\ell_s\ell_t}(m) \le \frac{1}{\gamma} \sum_{m_0 < m \le \ell_s\ell_t} \binom{\ell_s\ell_t}{m} q^m \le \sum_{m > m_0} (2\ell_s\ell_t q)^m \le 2(8q\ell^2)^{(m_0+1)}, \tag{A.45}$$

and therefore $a_{\ell_s\ell_t} \geq -1/2$. Furthermore,

$$P_{\ell_s\ell_t}(0) = (1-p)^{\ell_s\ell_t} \geq 1 - p\ell_s\ell_t \geq 1 - 8q\ell^2 \geq 1/2,$$

and thus (A.43) holds.

Next we bound $d_{\mathrm{TV}}\left(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}\right)$. Notice that

$$\sum_{m_0<m\leq\ell_s\ell_t} P_{\ell_s\ell_t}(m) \leq \sum_{m_0<m\leq\ell_s\ell_t} \binom{\ell_s\ell_t}{m} p^m \leq \sum_{m>m_0} (\ell_s\ell_t p)^m \leq 2(8q\ell^2)^{(m_0+1)}. \tag{A.46}$$

Therefore, by the definition of the total variation distance and $a_{\ell_s\ell_t}$,

$$\begin{aligned} d_{\mathrm{TV}}(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}) &= \frac{1}{2}|a_{\ell_s\ell_t}| + \frac{1}{2}\sum_{m_0<m\leq\ell_s\ell_t} \left| P_{\ell_s\ell_t}(m) - \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) \right| \\ &\leq \sum_{m_0<m\leq\ell_s\ell_t} \left(P_{\ell_s\ell_t}(m) + \frac{1}{\gamma} Q_{\ell_s\ell_t}(m) \right) \leq 4(8q\ell^2)^{(m_0+1)}, \end{aligned}$$

where the last inequality follows from (A.45) and (A.46). □

The following lemma is useful for upper bounding the total variation distance between a truncated mixture of product distribution P_Y and a product distribution Q_Y.

Lemma 9. *Let $P_{Y|X}$ be a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$ and denote the marginal of Y by $P_Y = \mathbb{E}_{X\sim P_X}[P_{Y|X}]$. Let Q_Y be such that $P_{Y|X=x} \ll Q_Y$ for all x. Let E be a measurable subset of $\mathcal{X}$. Define $g : \mathcal{X}^2 \to \bar{\mathbb{R}}_+$ by*

$$g(x, \widetilde{x}) \triangleq \int \frac{\mathrm{d}P_{Y|X=x}\mathrm{d}P_{Y|X=\widetilde{x}}}{\mathrm{d}Q}.$$

Then

$$d_{\mathrm{TV}}(P_Y, Q_Y) \leq \frac{1}{2}P_X(E^{\mathrm{c}}) + \frac{1}{2}\sqrt{\mathbb{E}\left[g(X, \widetilde{X})\mathbf{1}_E(X)\mathbf{1}_E(\widetilde{X})\right] - 1 + 2P_X(E^{\mathrm{c}})}, \tag{A.47}$$

where $\widetilde{X}$ is an independent copy of $X \sim P_X$.

Proof. By definition of the total variation distance,

$$\begin{aligned} d_{\mathrm{TV}}(P_Y, Q_Y) &= \frac{1}{2}\|P_Y - Q_Y\|_1 \\ &\le \frac{1}{2}\|\mathbb{E}[P_{Y|X}] - \mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}]\|_1 + \frac{1}{2}\|\mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}] - Q_Y\|_1, \end{aligned}$$

where the first term is $\|\mathbb{E}[P_{Y|X}] - \mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}]\|_1 = \|\mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\notin E\}}]\|_1 = \mathbb{P}\{X \notin E\}$. The second term is controlled by

$$\begin{aligned} &\|\mathbb{E}[P_{Y|X}\mathbf{1}_{\{X\in E\}}] - Q_Y\|_1^2 \\ &= \left(\mathbb{E}_{Q_Y}\left[\left|\frac{\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]}{Q_Y} - 1\right|\right]\right)^2 \\ &\le \mathbb{E}_{Q_Y}\left[\left(\frac{\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]}{Q_Y} - 1\right)^2\right] \qquad \text{(A.48)} \\ &= \mathbb{E}_{Q_Y}\left[\left(\frac{\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]}{Q_Y}\right)^2\right] + 1 - 2\,\mathbb{E}[\mathbb{E}\left[P_{Y|X}\mathbf{1}_{\{X\in E\}}\right]] \qquad \text{(A.49)} \\ &= \mathbb{E}\left[g(X,\widetilde{X})\mathbf{1}_E(X)\mathbf{1}_E(\widetilde{X})\right] + 1 - 2\,\mathbb{P}\{X\in E\}, \qquad \text{(A.50)} \end{aligned}$$

where (A.48) is Cauchy-Schwartz inequality, (A.50) follows from Fubini theorem. This proves the desired (A.47). □

Note that $\{V_t : t \in [n]\}$ can be equivalently generated as follows: Throw balls indexed by $[N]$ into bins indexed by $[n]$ independently and uniformly at random; let V_t denote the set of balls in the t^{th} bin. We need the following negative association property [94, Definition 1].

Lemma 10. *Let $\{\widetilde{V}_t : t \in [n]\}$ be an independent copy of $\{V_t : t \in [n]\}$. Fix a subset $C \subset [n]$ and let $S = \cup_{t\in C} V_t$. Conditional on C and S, the full vector $\{|V_s \cap \widetilde{V}_t| : s, t \in C\}$ is negatively associated, i.e., for every two disjoint index sets $I, J \subset C \times C$,*

$$\begin{aligned} &\mathbb{E}[f(V_s \cap \widetilde{V}_t, (s,t) \in I) g(V_s \cap \widetilde{V}_t, (s,t) \in J)] \\ &\le \mathbb{E}[f(V_s \cap \widetilde{V}_t, (s,t) \in I)]\mathbb{E}[g(V_s \cap \widetilde{V}_t, (s,t) \in J)], \end{aligned}$$

for all functions $f : \mathbb{R}^{|I|} \to \mathbb{R}$ and $g :^{|J|}$ that are either both non-decreasing or both non-increasing in every argument.

Proof. Define the indicator random variables $Z_{m,s,t}$ for $m \in S, s, t \in C$ as

$$Z_{m,s,t} = \begin{cases} 1 & \text{if the } m^{\text{th}} \text{ ball is contained in the } s^{\text{th}} \text{ and the } t^{\text{th}} \text{ bins,} \\ 0 & \text{otherwise .} \end{cases}$$

By [94, Proposition 12], the full vector $\{Z_{m,s,t} : m \in S, s, t \in C\}$ is negatively associated. By definition, we have

$$|V_s \cap \widetilde{V}_t| = \sum_{m \in S} Z_{m,s,t},$$

which is a non-decreasing function of $\{Z_{m,s,t} : m \in S\}$. Moreover, for distinct pairs $(s,t) \neq (s',t')$, the sets $\{(m,s,t) : m \in S\}$ and $\{(m,s',t') : m \in S\}$ are disjoint. Applying [94, Proposition 8] yields the desired statement. □

The negative association property of $\{|V_s \cap \widetilde{V}_t| : s, t \in C\}$ allows us to bound the expectation of any non-decreasing function of $\{|V_s \cap \widetilde{V}_t| : s, t \in C\}$ conditional on C and S as if they were independent [94, Lemma 2], i.e., for any collection of non-decreasing functions $\{f_{s,t} : s, t \in [n]\}$,

$$\mathbb{E}\left[\prod_{s,t \in C} f_{s,t}(|V_s \cap \widetilde{V}_t|) \,\middle|\, C, S\right] \leq \prod_{s,t \in C} \mathbb{E}\left[f_{s,t}(|V_s \cap \widetilde{V}_t|) \,\middle|\, C, S\right]. \tag{A.51}$$

Lemma 11. *Suppose that* $X \sim \mathrm{Binom}(1.5K, \frac{1}{k^2})$ *and* $Y \sim \mathrm{Binom}(3\ell, \frac{e}{k})$ *with* $K = k\ell$ *and* $k \geq 6e\ell$. *Then for all* $1 \leq m \leq 2\ell - 1$,

$$\mathbb{P}[X = m] \leq \mathbb{P}[Y = m],$$

and $\mathbb{P}[X \geq 2\ell] \leq \mathbb{P}[Y = 2\ell]$.

Proof. In view of the fact that $(\frac{n}{m})^m \leq \binom{n}{m} \leq (\frac{en}{m})^m$, we have for $1 \leq m \leq 2\ell$,

$$\mathbb{P}[X = m] = \binom{1.5K}{m} \left(\frac{1}{k^2}\right)^m \left(1 - \frac{1}{k^2}\right)^{1.5K - m} \leq \left(\frac{1.5eK}{mk^2}\right)^m.$$

Therefore,

$$\mathbb{P}[X \geq 2\ell] \leq \sum_{m=2\ell}^{\infty} \left(\frac{1.5e\ell}{km}\right)^m \leq \sum_{m=2\ell}^{\infty} \left(\frac{3e}{4k}\right)^m \leq \frac{(0.75e/k)^{2\ell}}{1 - 0.75e/k}.$$

On the other hand, for $1 \le m \le 2\ell - 1$

$$\begin{aligned}\mathbb{P}[Y = m] &= \binom{3\ell}{m} \left(\frac{e}{k}\right)^m \left(1 - \frac{e}{k}\right)^{3\ell - m} \\ &\ge \left(\frac{3e\ell}{mk}\right)^m \left(1 - \frac{3e\ell}{k}\right) \\ &\ge 2^{m-1} \left(\frac{1.5e\ell}{mk}\right)^m \ge \mathbb{P}[X = m].\end{aligned}$$

Moreover, $\mathbb{P}[Y = 2\ell] \ge \mathbb{P}[X \ge 2\ell]$. □

Lemma 12. *Let $T \sim \mathrm{Binom}(\ell, \tau)$ and $\lambda > 0$. Assume that $\lambda\ell \le \frac{1}{16}$. Then*

$$\mathbb{E}[\exp(\lambda T(T-1))] \le \exp\left(16\lambda\ell^2\tau^2\right). \tag{A.52}$$

Proof. Let $(s_1, \ldots, s_\ell, t_1, \ldots, t_\ell) \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(\tau)$, $S = \sum_{i=1}^{\ell} s_i$ and $T = \sum_{i=1}^{\ell} t_i$. Next we use a decoupling argument to replace $T^2 - T$ by ST:

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda T(T-1)\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i \ne j} t_i t_j\right)\right] \\ &\le \mathbb{E}\left[\exp\left(4\lambda \sum_{i \ne j} s_i t_j\right)\right], \qquad \text{(A.53)} \\ &\le \mathbb{E}\left[\exp\left(4\lambda ST\right)\right],\end{aligned}$$

where (A.53) is a standard decoupling inequality (see, e.g., [95, Theorem 1]). Since $\lambda T \le \lambda\ell \le \frac{1}{16}$ and $\exp(x) - 1 \le \exp(a)x$ for all $x \in [0, a]$, the desired (A.52) follows from

$$\begin{aligned}\mathbb{E}\left[\exp\left(4\lambda ST\right)\right] &= \mathbb{E}\left[\left(1 + \tau(\exp(4\lambda T) - 1)\right)^\ell\right] \\ &\le \mathbb{E}\left[\left(1 + 8\tau\lambda T\right)^\ell\right] \\ &\le \mathbb{E}\left[\exp\left(8\tau\lambda\ell T\right)\right] \\ &= \left(1 + \tau\left(\exp\left(8\tau\lambda\ell\right) - 1\right)\right)^\ell \\ &\le \exp\left(16\tau^2\lambda\ell^2\right).\end{aligned}$$

□

Proof of Proposition 4.3.1. Let $[i, j]$ denote the unordered pair of i and j. For any set $I \subset [N]$, let $\mathcal{E}(I)$ denote the set of unordered pairs of distinct elements in I,

i.e., $\mathcal{E}(I) = \{[i,j] : i,j \in S, i \neq j\}$, and let $\mathcal{E}(I)^c = \mathcal{E}([N]) \setminus \mathcal{E}(I)$. For $s,t \in [n]$ with $s \neq t$, let $\widetilde{G}_{V_sV_t}$ denote the bipartite graph where the set of left (right) vertices is V_s (resp. V_t) and the set of edges is the set of edges in $\widetilde{G}$ from vertices in V_s to vertices in V_t. For $s \in [n]$, let $\widetilde{G}_{V_sV_s}$ denote the subgraph of $\widetilde{G}$ induced by V_s. Let $\widetilde{P}_{V_sV_t}$ denote the edge distribution of $\widetilde{G}_{V_sV_t}$ for $s,t \in [n]$.

First, we show that the null distributions are exactly matched by the reduction scheme. Lemma 8 implies that $P'_{\ell_s\ell_t}$ and $Q'_{\ell_s\ell_t}$ are well-defined probability measures, and by definition, $(1-\gamma)Q'_{\ell_s\ell_t}+\gamma P'_{\ell_s\ell_t} = Q_{\ell_s\ell_t} = \mathrm{Binom}(\ell_s\ell_t, q)$. Under the null hypothesis, $G \sim \mathcal{G}(n,\gamma)$ and therefore, according to our reduction scheme, $E(V_s,V_t) \sim \mathrm{Binom}(\ell_s\ell_t, q)$ for $s < t$ and $E(V_t,V_t) \sim \mathrm{Binom}(\binom{\ell_t}{2}, q)$. Since the vertices in V_s and V_t are connected uniformly at random such that the total number of edges is $E(V_s,V_t)$, it follows that $\widetilde{P}_{V_sV_t} = \prod_{(i,j)\in V_s\times V_t} \mathrm{Bern}(q)$ for $s < t$ and $\widetilde{P}_{V_sV_t} = \prod_{[i,j]\in\mathcal{E}(V_s)} \mathrm{Bern}(q)$ for $s = t$. Conditional on V_1^n, $\{E(V_s,V_t) : 1 \le s \le t \le n\}$ are independent and so are $\{\widetilde{G}_{V_sV_t} : 1 \le s \le t \le n\}$. Consequently, $P_{\widetilde{G}|H_0^C} = \mathbb{P}_0 = \prod_{[i,j]\in\mathcal{E}([N])} \mathrm{Bern}(q)$ and $\widetilde{G} \sim \mathcal{G}(N,q)$.

Next, we proceed to consider the alternative hypothesis, under which G is drawn from the planted clique model $\mathcal{G}(n,k,\gamma)$. Let $C \subset [n]$ denote the planted clique. Define $S = \cup_{t\in C} V_t$ and recall $K = k\ell$. Then $|S| \sim \mathrm{Binom}(N, K/N)$ and conditional on $|S|$, S is uniformly distributed over all possible subsets of size $|S|$ in $[N]$. By the symmetry of the vertices of G, the distribution of $\widetilde{A}$ conditional on C does not depend on C. Hence, without loss of generality, we shall assume that $C = [k]$ henceforth. The distribution of $\widetilde{A}$ can be written as a mixture distribution indexed by the random set S as

$$\widetilde{A} \sim \widetilde{\mathbb{P}}_1 \triangleq \mathbb{E}_S \left[\widetilde{P}_{SS} \times \prod_{[i,j]\in\mathcal{E}(S)^c} \mathrm{Bern}(q) \right].$$

By the definition of $\mathbb{P}_1$,

$$
\begin{aligned}
&d_{\mathrm{TV}}(\widetilde{\mathbb{P}}_1, \mathbb{P}_1) \\
&= d_{\mathrm{TV}}\left(\mathbb{E}_S\left[\widetilde{P}_{SS} \times \prod_{[i,j]\in\mathcal{E}(S)^c} \mathrm{Bern}(q)\right], \mathbb{E}_S\left[\prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p) \prod_{[i,j]\in\mathcal{E}(S)^c} \mathrm{Bern}(q)\right]\right) \\
&\le \mathbb{E}_S\left[d_{\mathrm{TV}}\left(\widetilde{P}_{SS} \times \prod_{[i,j]\in\mathcal{E}(S)^c} \mathrm{Bern}(q), \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p) \prod_{[i,j]\in\mathcal{E}(S)^c} \mathrm{Bern}(q)\right)\right] \\
&= \mathbb{E}_S\left[d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right)\right] \\
&\le \mathbb{E}_S\left[d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right)\mathbf{1}_{\{|S|\le 1.5K\}}\right] + \exp(-K/12), \qquad \text{(A.54)}
\end{aligned}
$$

where the first inequality follows from the convexity of $(P, Q) \mapsto d_{\mathrm{TV}}(P, Q)$, and the last inequality follows from applying the Chernoff bound to $|S|$. Fix an $S \subset [N]$ such that $|S| \le 1.5K$. Define $P_{V_tV_t} = \prod_{[i,j]\in\mathcal{E}(V_t)} \mathrm{Bern}(q)$ for $t \in [k]$ and $P_{V_sV_t} = \prod_{(i,j)\in V_s\times V_t} \mathrm{Bern}(p)$ for $1 \le s < t \le k$. By the triangle inequality,

$$
\begin{aligned}
&d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right) \\
&\le d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right) \qquad \text{(A.55)} \\
&\quad + d_{\mathrm{TV}}\left(\mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right], \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right). \qquad \text{(A.56)}
\end{aligned}
$$

To bound the term in (A.55), first note that conditional on S, $\{V_1^k\}$ can be generated as follows: Throw balls indexed by S into bins indexed by $[k]$ independently and uniformly at random; let V_t is the set of balls in the t^{th} bin. Define the event $E = \{V_1^k : |V_t| \le 2\ell, t \in [k]\}$. Since $|V_t| \sim \mathrm{Binom}(|S|, 1/k)$ is stochastically dominated by $\mathrm{Binom}(1.5K, 1/k)$ for each fixed $1 \le t \le k$, it follows from the

Chernoff bound and the union bound that $\mathbb{P}\{E^c\} \le k\exp(-\ell/18)$.

$$
\begin{aligned}
&d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right)\\
&\overset{(a)}{=} d_{\mathrm{TV}}\left(\mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t} \,\middle|\, S\right], \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right)\\
&\le \mathbb{E}_{V_1^k}\left[d_{\mathrm{TV}}\left(\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s\le t\le k} P_{V_sV_t}\right) \,\middle|\, S\right]\\
&\le \mathbb{E}_{V_1^k}\left[d_{\mathrm{TV}}\left(\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s\le t\le k} P_{V_sV_t}\right)\mathbf{1}_{\left\{V_1^k\in E\right\}} \,\middle|\, S\right] + k\exp(-\ell/18),
\end{aligned}
$$

where (a) holds because conditional on V_1^k, $\left\{\widetilde{A}_{V_sV_t} : s,t\in[k]\right\}$ are independent. Recall that $\ell_t = |V_t|$. For any fixed $V_1^k \in E$, we have

$$
\begin{aligned}
d_{\mathrm{TV}}\left(\prod_{1\le s\le t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s\le t\le k} P_{V_sV_t}\right) &\overset{(a)}{=} d_{\mathrm{TV}}\left(\prod_{1\le s<t\le k} \widetilde{P}_{V_sV_t}, \prod_{1\le s<t\le k} P_{V_sV_t}\right)\\
&\overset{(b)}{=} d_{\mathrm{TV}}\left(\prod_{1\le s<t\le k} P'_{\ell_s\ell_t}, \prod_{1\le s<t\le k} P_{\ell_s\ell_t}\right)\\
&\le d_{\mathrm{TV}}\left(\prod_{1\le s<t\le k} P'_{\ell_s\ell_t}, \prod_{1\le s<t\le k} P_{\ell_s\ell_t}\right)\\
&\le \sum_{1\le s<t\le k} d_{\mathrm{TV}}\left(P'_{\ell_s\ell_t}, P_{\ell_s\ell_t}\right) \overset{(c)}{\le} 2k^2(8q\ell^2)^{(m_0+1)},
\end{aligned}
$$

where (a) follows since $\widetilde{P}_{V_tV_t} = P_{V_tV_t}$ for all $t\in[k]$; (b) is because the number of edges $E(V_s,V_t)$ is a sufficient statistic for testing $\widetilde{P}_{V_sV_t}$ versus $P_{V_sV_t}$ on the submatrix $A_{V_sV_t}$ of the adjacency matrix; (c) follows from Lemma 8. Therefore,

$$
d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t} \,\middle|\, S\right]\right) \le 2k^2(8q\ell^2)^{(m_0+1)} + k\exp(-\ell/18). \tag{A.57}
$$

To bound the term in (A.56), applying Lemma 9 yields

$$d_{\mathrm{TV}}\left(\mathbb{E}_{V_1^k}\left[\prod_{1\le s\le t\le k} P_{V_sV_t}\;\middle|\; S\right], \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right)$$
$$\le \frac{1}{2}\mathbb{P}\left\{E^c\right\} + \frac{1}{2}\sqrt{\mathbb{E}_{V_1^k;\widetilde{V}_1^k}\left[g(V_1^k,\widetilde{V}_1^k)\mathbf{1}_{\left\{V_1^k\in E\right\}}\mathbf{1}_{\left\{\widetilde{V}_1^k\in E\right\}}\;\middle|\; S\right] - 1 + 2\mathbb{P}\left\{E^c\right\}}, \tag{A.58}$$

where

$$\begin{aligned} g(V_1^k,\widetilde{V}_1^k) &= \int \frac{\prod_{1\le s\le t\le k} P_{V_sV_t}\prod_{1\le s\le t\le k} P_{\widetilde{V}_s\widetilde{V}_t}}{\prod_{[i,j]\in\mathcal{E}(S)}\mathrm{Bern}(p)} \\ &= \prod_{s,t=1}^{k}\left(\frac{q^2}{p}+\frac{(1-q)^2}{1-p}\right)^{\binom{|V_s\cap\widetilde{V}_t|}{2}} \\ &= \prod_{s,t=1}^{k}\left(\frac{1-\frac{3}{2}q}{1-2q}\right)^{\binom{|V_s\cap\widetilde{V}_t|}{2}}. \end{aligned}$$

Let $X\sim\mathrm{Bin}(1.5K,\frac{1}{k^2})$ and $Y\sim\mathrm{Bin}(3\ell,e/k)$. It follows that

$$\begin{aligned} &\mathbb{E}_{V_1^k;\widetilde{V}_1^k}\left[\prod_{s,t=1}^{k}\left(\frac{1-\frac{3}{2}q}{1-2q}\right)^{\binom{|V_s\cap\widetilde{V}_t|}{2}}\prod_{s,t=1}^{k}\mathbf{1}_{\left\{|V_s|\le 2\ell,|\widetilde{V}_t|\le 2\ell\right\}}\;\middle|\; S\right] \\ &\overset{(a)}{\le} \mathbb{E}_{V_1^k;\widetilde{V}_1^k}\left[\prod_{s,t=1}^{k} e^{q\binom{|V_s\cap\widetilde{V}_t|\wedge 2\ell}{2}}\;\middle|\; S\right] \\ &\overset{(b)}{\le} \prod_{s,t=1}^{k}\mathbb{E}\left[e^{q\binom{|V_s\cap\widetilde{V}_t|\wedge 2\ell}{2}}\;\middle|\; S\right] \\ &\overset{(c)}{\le} \left(\mathbb{E}\left[e^{q\binom{X\wedge 2\ell}{2}}\right]\right)^{k^2} \overset{(d)}{\le} \mathbb{E}\left[e^{q\binom{Y}{2}}\right]^{k^2} \overset{(e)}{\le} \exp(72e^2q\ell^2), \end{aligned} \tag{A.59}$$

where (a) follows from $1+x\le e^x$ for all $x\ge 0$ and $q<1/4$; (b) follows from the negative association property of $\{|V_s\cap\widetilde{V}_t| : s,t\in[k]\}$ proved in Lemma 10 and (A.51), in view of the monotonicity of $x\mapsto e^{q\binom{x\wedge 2\ell}{2}}$ on $\mathbb{R}_+$; (c) follows because $|V_s\cap\widetilde{V}_t|$ is stochastically dominated by $\mathrm{Binom}(1.5K,1/k^2)$ for all $(s,t)\in[k]^2$; (d) follows from Lemma 11; (e) follows from Lemma 12 with $\lambda = q/2$ and

$q\ell \leq 1/8$. Therefore, by (A.58)

$$d_{\mathrm{TV}}\left(\widetilde{P}_{SS}, \prod_{[i,j]\in\mathcal{E}(S)} \mathrm{Bern}(p)\right) \leq 0.5ke^{-\frac{\ell}{18}} + 0.5\sqrt{e^{72e^2q\ell^2} - 1 + 2ke^{-\frac{\ell}{18}}}$$
$$\leq 0.5ke^{-\frac{\ell}{18}} + 0.5\sqrt{e^{72e^2q\ell^2} - 1} + \sqrt{0.5k}e^{-\frac{\ell}{36}}. \tag{A.60}$$

The proposition follows by combining (A.54), (A.55), (A.56), (A.57) and (A.60). □

A.7 Proof of Proposition 4.3.2

Proof. By assumption the test ϕ satisfies

$$\mathbb{P}_0\{\phi(G') = 1\} + \mathbb{P}_1\{\phi(G') = 0\} = \eta,$$

where G' is the graph in $\mathrm{PDS}(N, K, 2q, q)$ distributed according to either $\mathbb{P}_0$ or $\mathbb{P}_1$. Let G denote the graph in the $\mathrm{PC}(n, k, \gamma)$ and $\widetilde{G}$ denote the corresponding output of the randomized reduction scheme. Proposition 4.3.1 implies that $\widetilde{G} \sim \mathcal{G}(N, q)$ under H_0^{C}. Therefore $\mathbb{P}_{H_0^{\mathrm{C}}}\{\phi(\widetilde{G}) = 1\} = \mathbb{P}_0\{\phi(G') = 1\}$. Moreover,

$$|\mathbb{P}_{H_1^{\mathrm{C}}}\{\phi(\widetilde{G}) = 0\} - \mathbb{P}_1\{\phi(G') = 0\}| \leq d_{\mathrm{TV}}(P_{\widetilde{G}|H_1^C}, \mathbb{P}_1) \leq \xi.$$

It follows that

$$\mathbb{P}_{H_0^{\mathrm{C}}}\{\phi(\widetilde{G}) = 1\} + \mathbb{P}_{H_1^{\mathrm{C}}}\{\phi(\widetilde{G}) = 0\} \leq \eta + \xi.$$

□

A.8 Proof of Theorem 4.3.3

Proof. Fix $\alpha > 0$ and $0 < \beta < 1$ that satisfy (4.4). Then it is straightforward to verify that

$$\alpha < \beta < \min\left\{\frac{2 + m_0\delta}{4 + 2\delta}\alpha,\ \frac{1}{2} - \delta + \frac{1 + 2\delta}{4 + 2\delta}\alpha\right\} \tag{A.61}$$

holds for some $\delta > 0$. Let $\ell \in \mathbb{N}$ and $q_\ell = \ell^{-(2+\delta)}$. Define

$$n_\ell = \lfloor \ell^{\frac{2+\delta}{\alpha}-1} \rfloor,\ k_\ell = \lfloor \ell^{\frac{(2+\delta)\beta}{\alpha}-1} \rfloor,\ N_\ell = n_\ell \ell,\ K_\ell = k_\ell \ell. \tag{A.62}$$

Then

$$\lim_{\ell\to\infty} \frac{\log \frac{1}{q_\ell}}{\log N_\ell} = \frac{(2+\delta)}{(2+\delta)/\alpha - 1 + 1} = \alpha,\ \lim_{\ell\to\infty} \frac{\log K_\ell}{\log N_\ell} = \frac{(2+\delta)\beta/\alpha - 1 + 1}{(2+\delta)/\alpha - 1 + 1} = \beta. \tag{A.63}$$

Suppose that for the sake of contradiction there exists a small $\epsilon > 0$ and a sequence of randomized polynomial-time tests $\{\phi_\ell\}$ for $\mathrm{PDS}(N_\ell, K_\ell, 2q_\ell, q_\ell)$, such that

$$\mathbb{P}_0\{\phi_{N_\ell,K_\ell}(G') = 1\} + \mathbb{P}_1\{\phi_{N_\ell,K_\ell}(G') = 0\} \le 1/2 - \epsilon$$

holds for arbitrarily large ℓ, where G' is the graph in the $\mathrm{PDS}(N_\ell, K_\ell, 2q_\ell, q_\ell)$. Since $\beta > \alpha$, we have $k_\ell \ge \ell^{1+\delta}$. Therefore, $16 q_\ell \ell^2 \le 1$ and $k_\ell \ge 6e\ell$ for all sufficiently large ℓ. Applying Proposition 4.3.2, we conclude that $G \mapsto \phi(\widetilde{G})$ is a randomized polynomial-time test for $\mathrm{PC}(n_\ell, k_\ell, \gamma)$ whose Type-I+II error probability satisfies

$$\mathbb{P}_{H_0^{\mathrm{C}}}\{\phi_\ell(\widetilde{G}) = 1\} + \mathbb{P}_{H_1^{\mathrm{C}}}\{\phi_\ell(\widetilde{G}) = 0\} \le \frac{1}{2} - \epsilon + \xi, \tag{A.64}$$

where ξ is given by the right-hand side of (4.3). By the definition of q_ℓ, we have $q_\ell \ell^2 = \ell^{-\delta}$ and thus

$$k_\ell^2 (q_\ell \ell^2)^{m_0+1} \le \ell^{2((2+\delta)\beta/\alpha - 1) - (m_0+1)\delta} \le \ell^{-\delta},$$

where the last inequality follows from (A.61). Therefore $\xi \to 0$ as $\ell \to \infty$. Moreover, by the definition in (A.62),

$$\lim_{\ell\to\infty} \frac{\log k_\ell}{\log n_\ell} = \frac{(2+\delta)\beta/\alpha - 1}{(2+\delta)/\alpha - 1} \le \frac{1}{2} - \delta,$$

where the above inequality follows from (A.61). Therefore, (A.64) contradicts our assumption that Hypothesis 1 holds for γ. Finally, if Hypothesis 1 holds for any $\gamma > 0$, (4.5) follows from (4.4) by sending $\gamma \downarrow 0$. □

A.9 Proof of Theorem 5.1.2

The following lemma provides a deterministic sufficient condition for the success of SDP (5.5) in the case $a > b$.

Lemma 13. *Suppose there exist* $D^* = \mathsf{diag}\,\{d_i^*\}$ *and* $\lambda^* \in \mathbb{R}$ *such that* $S^* \triangleq D^* - A + \lambda^* J$ *satisfies* $S^* \succeq 0$, $\lambda_2(S^*) > 0$ *and*

$$S^*\sigma^* = 0. \tag{A.65}$$

Then $\widehat{Y}_{\mathrm{SDP}} = Y^*$ *is the unique solution to (5.5).*

Proof. The Lagrangian function is given by

$$L(Y, S, D, \lambda) = \langle A, Y\rangle + \langle S, Y\rangle - \langle D, Y - I\rangle - \lambda\langle J, Y\rangle,$$

where the Lagrangian multipliers are denoted by $S \succeq 0$, $D = \mathsf{diag}\,\{d_i\}$, and $\lambda \in \mathbb{R}$. Then for any Y satisfying the constraints in (5.5),

$$\begin{aligned}\langle A, Y\rangle &\overset{(a)}{\leq} L(Y, S^*, D^*, \lambda^*) = \langle D^*, I\rangle = \langle D^*, Y^*\rangle = \langle A + S^* - \lambda^* J, Y^*\rangle \\ &\overset{(b)}{=} \langle A, Y^*\rangle,\end{aligned}$$

where (a) holds because $\langle S^*, Y\rangle \geq 0$; (b) holds because $\langle Y^*, S^*\rangle = (\sigma^*)^\top S^*\sigma^* = 0$ by (A.65). Hence, Y^* is an optimal solution. It remains to establish its uniqueness. To this end, suppose $\widetilde{Y}$ is an optimal solution. Then,

$$\langle S^*, \widetilde{Y}\rangle = \langle D^* - A + \lambda^* J, \widetilde{Y}\rangle \overset{(a)}{=} \langle D^* - A, \widetilde{Y}\rangle \overset{(b)}{=} \langle D^* - A, Y^*\rangle = \langle S^*, Y^*\rangle = 0.$$

where (a) holds because $\langle J, \widetilde{Y}\rangle = 0$; (b) holds because $\langle A, \widetilde{Y}\rangle = \langle A, Y^*\rangle$ and $\widetilde{Y}_{ii} = Y^*_{ii} = 1$ for all $i \in [n]$. In view of (A.65), since $\widetilde{Y} \succeq 0$, $S^* \succeq 0$ with $\lambda_2(S^*) > 0$, $\widetilde{Y}$ must be a multiple of $Y^* = \sigma^*(\sigma^*)^\top$. Because $\widetilde{Y}_{ii} = 1$ for all $i \in [n]$, $\widetilde{Y} = Y^*$. □

Proof of Theorem 5.1.2. The theorem is proved first for $a > b$. Let $D^* = \mathsf{diag}\,\{d_i^*\}$ with

$$d_i^* = \sum_{j=1}^{n} A_{ij}\sigma_i^*\sigma_j^* \tag{A.66}$$

and choose any $\lambda^* \geq \frac{p+q}{2}$. It suffices to show that $S^* = D^* - A + \lambda^* J$ satisfies the conditions in Lemma 13 with high probability.

By definition, $d_i^* \sigma_i^* = \sum_j A_{ij} \sigma_j^*$ for all i, i.e., $D^* \sigma^* = A \sigma^*$. Since $J \sigma^* = 0$, (A.65) holds, that is, $S^* \sigma^* = 0$. It remains to verify that $S^* \succeq 0$ and $\lambda_2(S^*) > 0$ with probability converging to one, which amounts to showing that

$$\mathbb{P}\left\{ \inf_{x \perp\!\!\!\perp \sigma^*, \|x\|_2 = 1} x^\top S^* x > 0 \right\} \to 1. \tag{A.67}$$

Note that $\mathbb{E}[A] = \frac{p-q}{2} Y^* + \frac{p+q}{2} J - pI$ and $Y^* = \sigma^* (\sigma^*)^\top$. Thus for any x such that $x \perp\!\!\!\perp \sigma^*$ and $\|x\|_2 = 1$,

$$\begin{aligned} x^\top S^* x &= x^\top D^* x - x^\top \mathbb{E}[A]\, x + \lambda^* x^\top J x - x^\top (A - \mathbb{E}[A])\, x \\ &= x^\top D^* x - \frac{p-q}{2} x^\top Y^* x + \left(\lambda^* - \frac{p+q}{2} \right) x^\top J x + p - x^\top (A - \mathbb{E}[A])\, x \\ &\overset{(a)}{\geq} x^\top D^* x + p - x^\top (A - \mathbb{E}[A])\, x \geq \min_{i \in [n]} d_i^* + p - \|A - \mathbb{E}[A]\|. \end{aligned} \tag{A.68}$$

where (a) holds since $\lambda^* \geq \frac{p+q}{2}$ and $\langle x, \sigma^* \rangle = 0$. It follows from Theorem A.12.1 that $\|A - \mathbb{E}[A]\| \leq c' \sqrt{\log n}$ with high probability for a positive constant c' depending only on a. Moreover, note that each d_i is equal in distribution to $X - R$, where $X \sim \mathrm{Binom}(\frac{n}{2} - 1, \frac{a \log n}{n})$ and $R \sim \mathrm{Binom}(\frac{n}{2}, \frac{b \log n}{n})$ are independent. Hence, Lemma 15 implies that

$$\mathbb{P}\left\{ X - R \geq \frac{\log n}{\log \log n} \right\} \geq 1 - n^{-(\sqrt{a} - \sqrt{b})^2/2 + o(1)}.$$

Applying the union bound implies that $\min_{i \in [n]} d_i^* \geq \frac{\log n}{\log \log n}$ holds with probability at least $1 - n^{1 - (\sqrt{a} - \sqrt{b})^2/2 + o(1)}$. It follows from the assumption $(\sqrt{a} - \sqrt{b})^2 > 2$ and (A.68) that the desired (A.67) holds, completing the proof in the case $a > b$.

For the case $a < b$, we replace the $\arg\max$ by $\arg\min$ in the SDP (5.5), which is equivalent to substituting $-A$ for A in the original maximization problem, as well as the sufficient condition in Lemma 13. Set the dual variable d_i^* according to (A.66) with $-A$ replacing A and choose any $\lambda^* \geq -\frac{p+q}{2}$. Then (A.65) still holds and (A.68) changes to $x^\top S^* x \geq \min_{i \in [n]} d_i^* - p - \|A - \mathbb{E}[A]\|$, where $\min_{i \in [n]} d_i^* \geq \frac{\log n}{\log \log n}$ holds with probability at least $1 - n^{1 - (\sqrt{a} - \sqrt{b})^2/2 + o(1)}$ by Lemma 15 and the union bound. Therefore, in view of Theorem A.12.1 and the assumption $(\sqrt{a} - \sqrt{b})^2 > 2$, the desired (A.67) still holds, completing the proof for the case $a < b$. $\square$

A.10 Proof of Theorem 5.2.1

Lemma 14. *Suppose there exist* $D^* = \mathsf{diag}\{d_i^*\} \geq 0$, $B^* \in \mathcal{S}^n$ *with* $B^* \geq 0$, $\lambda^* \in \mathbb{R}$, *and* $\eta^* \in \mathbb{R}$ *such that* $S^* \triangleq D^* - B^* - A + \eta^* I + \lambda^* J$ *satisfies* $S^* \succeq 0$, $\lambda_2(S^*) > 0$, *and*

$$\begin{aligned} S^*\xi^* &= 0, \\ d_i^*(Z_{ii}^* - 1) &= 0, \quad \forall i, \\ B_{ij}^* Z_{ij}^* &= 0, \quad \forall i, j. \end{aligned} \tag{A.69}$$

Then $\widehat{Z}_{\mathrm{SDP}} = Z^*$ *is the unique solution to (5.9).*

Proof. The Lagrangian function is given by

$$\begin{aligned} L(Z, S, D, B, \lambda, \eta) =& \langle A, Z\rangle + \langle S, Z\rangle - \langle D, Z - I\rangle + \langle B, Z\rangle - \eta\left(\langle I, Z\rangle - K\right) \\ & - \lambda\left(\langle J, Z\rangle - K^2\right), \end{aligned}$$

where $S \succeq 0$, $D = \mathsf{diag}\{d_i\} \geq 0$, $B \in \mathcal{S}^n$ with $B \geq 0$, and $\lambda, \eta \in \mathbb{R}$ are the Lagrangian multipliers. Then, for any Z satisfying the constraints in (5.9), It follows that

$$\begin{aligned} \langle A, Z\rangle &\overset{(a)}{\leq} L(Z, S^*, D^*, B^*, \lambda^*, \eta^*) = \langle D^*, I\rangle + \eta^* K + \lambda^* K^2 \\ &\overset{(b)}{=} \langle D^*, Z^*\rangle + \eta^* K + \lambda^* K^2 \\ &= \langle A + B^* + S^* - \eta^* I - \lambda^* J, Z^*\rangle + \eta^* K + \lambda^* K^2 \overset{(c)}{=} \langle A, Z^*\rangle, \end{aligned}$$

where (a) follows because $\langle S^*, Z\rangle \geq 0$, $\langle D^*, Z - I\rangle \leq 0$, and $\langle B^*, Z\rangle \geq 0$; (b) holds due to $d_i^*(Z_{ii}^* - 1) = 0, \forall i$; (c) holds because $B_{ij}^* Z_{ij}^* = 0, \forall i, j$ and $\langle Z^*, S^*\rangle = (\xi^*)^\top S^* \xi^* = 0$. Hence, Z^* is an optimal solution. It remains to establish the uniqueness. To this end, suppose $\widetilde{Z}$ is another optimal solution. Then,

$$\begin{aligned} \langle S^*, \widetilde{Z}\rangle = \langle D^* - B^* - A + \eta^* I + \lambda^* J, \widetilde{Z}\rangle &\overset{(a)}{=} \langle D^* - B^* - A, \widetilde{Z}\rangle \overset{(b)}{\leq} \langle D^* - A, Z^*\rangle \\ = \langle S^*, Z^*\rangle = 0. & \end{aligned}$$

where (a) holds because $\langle I, \widetilde{Z}\rangle = K$ and $\langle J, \widetilde{Z}\rangle = K^2$; (b) holds because $\langle A, \widetilde{Z}\rangle = \langle A, Z^*\rangle$, $B^*, \widetilde{Z} \geq 0$, and $\langle D^*, \widetilde{Z}\rangle \leq \sum_{i \in C^*} d_i^* = \langle D^*, Z^*\rangle$ since $d_i^* \geq 0$

and $\widetilde{Z}_{ii} \leq 1$ for all $i \in [n]$. Since $\widetilde{Z} \succeq 0$ and $S^* \succeq 0$ with $\lambda_2(S^*) > 0$, $\widetilde{Z}$ needs to be a multiple of $Z^* = \xi^*(\xi^*)^\top$. Then $\widetilde{Z} = Z^*$ since $\mathrm{Tr}(\widetilde{Z}) = \mathrm{Tr}(Z^*) = K$. $\square$

Proof of Theorem 5.2.1. The theorem is proved first for $a > b$. Recall $\tau^* = \frac{a-b}{\log a - \log b}$ if $a, b > 0$ and $a \neq b$. Let $\tau^* = 0$ if $a = 0$ or $b = 0$. Choose $\lambda^* = \tau^* \log n/n$, $\eta^* = \|A - \mathbb{E}[A]\|$, $D^* = \mathsf{diag}\{d_i^*\}$ with

$$d_i^* = \begin{cases} \sum_{j \in C^*} A_{ij} - \eta^* - \lambda^* K & \text{if } i \in C^* \\ 0 & \text{otherwise} \end{cases}.$$

Define $b_i^* \triangleq \lambda^* - \frac{1}{K}\sum_{j \in C^*} A_{ij}$ for $i \notin C^*$. Let $B^* \in \mathcal{S}^n$ be given by

$$B_{ij}^* = b_i \mathbf{1}_{\{i \notin C^*, j \in C^*\}} + b_j \mathbf{1}_{\{i \in C^*, j \notin C^*\}}.$$

It suffices to show that (S^*, D^*, B^*) satisfies the conditions in Lemma 14 with probability tending to one.

By definition, we have $d_i^*(Z_{ii}^* - 1) = 0$ and $B_{ij}^* Z_{ij}^* = 0$ for all $i, j \in [n]$. Moreover, for all $i \in C^*$,

$$d_i^* \xi_i^* = d_i^* = \sum_j A_{ij}\xi_j^* - \eta^* - \lambda^* K = \sum_j A_{ij}\xi_j^* + \sum_j B_{ij}^* \xi_j^* - \eta^* - \lambda^* K,$$

where the last equality holds because $B_{ij}^* = 0$ if $(i, j) \in C^* \times C^*$; for all $i \notin C^*$,

$$\sum_j A_{ij}\xi_j^* + \sum_j B_{ij}^* \xi_j^* - \lambda^* K = \sum_{j \in C^*} A_{ij} + K b_i^* - \lambda^* K = 0,$$

where the last equality follows from our choice of b_i^*. Hence, $D^*\xi^* = A\xi^* + B^*\xi^* - \eta^*\xi^* - \lambda^* K\mathbf{1}$ and consequently $S^*\sigma^* = 0$.

We next show that $D^* \geq 0$, $B^* \geq 0$ with probability converging to 1. It follows from Theorem A.12.1 that $\eta^* \leq c'\sqrt{\log n}$ with probability tending to one for some positive constant c' depending only on a. Furthermore, let $X_i \triangleq \sum_{j \in C^*} A_{ij}$. Then $X_i \sim \mathrm{Binom}(K - 1, \frac{a \log n}{n})$ if $i \in C^*$ and $\mathrm{Binom}(K, \frac{b \log n}{n})$ otherwise. We divide the analysis into two separate cases. First consider the case $b = 0$, then $X_i = 0$ for all $i \notin C^*$. Since $\tau^* = 0$ in this case, $\min_{i \notin C^*} b_i^* = 0$ holds automatically. For any $i \in C^*$, applying Lemma 16 with $\tau = 0$ yields

$$\mathbb{P}\left\{X_i \geq \frac{\log n}{\log\log n}\right\} \geq 1 - \mathbb{P}\left\{X_i \leq \frac{\log n}{\log\log n}\right\} \geq 1 - n^{-\rho a + o(1)}.$$

Applying the union bound implies that $\mathbb{P}\{\min_{i\in C^*} X_i \geq \frac{\log n}{\log\log n}\} \geq 1-n^{1-\rho a+o(1)} \to 1$, because $\rho f(a,0) = \rho a > 1$ by the assumption (5.10). Since $\sqrt{\log n} = o(\frac{\log n}{\log\log n})$ and $\tau^* = 0$, it follows that with probability converging to 1, $\min_{i\in C^*} d_i^* \geq 0$ and we are done with the case $b = 0$. For $b > 0$, Lemma 16 implies that

$$\mathbb{P}\left\{X_i \geq \rho\tau^* \log n + \frac{\log n}{\log\log n}\right\} \geq 1 - n^{-\rho\left(a-\tau^*\log\frac{ea}{\tau^*}+o(1)\right)}, \quad \forall i \in C^*,$$
$$\mathbb{P}\left\{X_i \leq \rho\tau^* \log n\right\} \geq 1 - n^{-\rho\left(b-\tau^*\log\frac{eb}{\tau^*}+o(1)\right)}, \quad \forall i \notin C^*.$$

By definition, $f(a,b) = a - \tau^*\log\frac{ea}{\tau^*} = b - \tau^*\log\frac{eb}{\tau^*}$ in this case. Applying the union bound implies that with probability at least $1 - n^{1-\rho f(a,b)+o(1)}$,

$$\min_{i\in C^*} X_i \geq \rho\tau^* \log n + \frac{\log n}{\log\log n},$$
$$\max_{i\notin C^*} X_i \leq \rho\tau^* \log n.$$

Since $\sqrt{\log n} = o(\frac{\log n}{\log\log n})$ and $\rho f(a,b) > 1$ by the assumption (5.10), it follows that with probability converging to 1, $\min_{i\in C^*} d_i^* \geq 0$ and $\min_{i\notin C^*} b_i^* \geq 0$.

It remains to verify $S^* \succeq 0$ with $\lambda_2(S^*) > 0$ with probability converging to 1, i.e.,

$$\mathbb{P}\left\{\inf_{x\perp\!\!\!\perp\sigma^*, \|x\|_2=1} x^\top S^* x > 0\right\} \to 1. \tag{A.70}$$

Note that

$$\mathbb{E}\left[A\right] = (p-q)Z^* - p\begin{bmatrix} I_{K\times K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} - q\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{(n-K)\times(n-K)}\end{bmatrix} + qJ.$$

It follows that for any $x \perp\!\!\!\perp \sigma^*$ and $\|x\|_2 = 1$,

$$
\begin{aligned}
& x^\top S^* x \\
&= x^\top D^* x - x^\top B^* x + (\lambda^* - q) x^\top J x + p \sum_{i \in C^*} x_i^2 + q \sum_{i \notin C^*} x_i^2 + \eta^* - x^\top \left(A - \mathbb{E}\left[A\right] \right) x \\
&\overset{(a)}{=} \sum_{i \in C^*} \left(d_i^* + p \right) x_i^2 + (\lambda^* - q) x^\top J x + q \sum_{i \notin C^*} x_i^2 + \eta^* - x^\top \left(A - \mathbb{E}\left[A\right] \right) x \\
&\geq \left(\min_{i \in C^*} d_i^* + p \right) \sum_{i \in C^*} x_i^2 + (\lambda^* - q) x^\top J x + q \sum_{i \notin C^*} x_i^2 + \eta^* - \| A - \mathbb{E}\left[A\right] \| \\
&\overset{(b)}{\geq} \left(\min_{i \in C^*} d_i^* + p \right) \sum_{i \in C^*} x_i^2 + q \sum_{i \notin C^*} x_i^2 \\
&\overset{(c)}{\geq} \min \left\{ \min_{i \in C^*} d_i^* + p, q \right\}, \qquad \text{(A.71)}
\end{aligned}
$$

where (a) holds because $B_{ij}^* = 0$ for all $i, j \notin C^*$ and

$$
x^\top B^* x = 2 \sum_{i \notin C^*} \sum_{j \in C^*} x_i x_j B_{ij}^* = 2 \sum_{i \notin C^*} x_i b_i^* \sum_{j \in C^*} x_j = 0;
$$

(b) holds because $\eta^* = \| A - \mathbb{E}\left[A\right] \|$ and $\lambda^* = \tau^* \frac{\log n}{n} \geq q = b \frac{\log n}{n}$, since $\log \frac{a}{b} \leq \frac{a}{b} - 1$; (c) is due to $\|x\|_2^2 = 1$. Notice that we have shown $\min_{i \in C^*} d_i^* \geq 0$ with probability converging to 1. Therefore, the desired (A.70) holds in view of (A.71), completing the proof in the case $a > b$.

For the case $a < b$, it suffices to modify the above proof by replacing A with $-A$ in the SDP (5.9), Lemma 14, and the definitions of d_i^* and b_i^*, and choosing $\lambda^* = -\tau^* \log n/n - \log n/(K \log \log n)$, $\eta^* = \|A - \mathbb{E}[A]\| + 2q$. Then (A.69) and (A.70) still hold, and $D^* \geq 0$, $B^* \geq 0$ with probability converging to 1. Therefore the theorem follows by applying Lemma 14. □

A.11 Proof of Theorem 5.2.2

To lower bound the worst-case probability of error, consider the Bayesian setting where the planted cluster C^* is uniformly chosen among all K-subsets of $[n]$ with $K = \lfloor \rho n \rfloor$. If $a = b$, then the cluster is unidentifiable from the graph.

Next, we prove the theorem first for the case $a > b$. If $b = 0$, then perfect recovery is possible if and only if the subgraph formed by the nodes in cluster, which is

$\mathcal{G}(K, a\log n/n)$, contains no isolated node.[1] This occurs with high probability if $\rho a < 1$ [96]. Next we consider $a > b > 0$.

Since the prior distribution of C^* is uniform, the ML estimator minimizes the error probability among all estimators and thus we only need to find when the ML estimator fails. Let $e(i,S) \triangleq \sum_{j\in S} A_{ij}$ denote the number of edges between node i and nodes in $S \subset [n]$. Let F denote the event that $\min_{i\in C^*} e(i,C^*) < \max_{j\notin C^*} e(j,C^*)$, which implies the existence of $i \in C^*$ and $j \notin C^*$, such that the set $C^*\backslash\{i\}\cup\{j\}$ achieves a strictly higher likelihood than C^*. Hence $\mathbb{P}\{\text{ML fails}\} \geq \mathbb{P}\{F\}$. Next we bound $\mathbb{P}\{F\}$ from below.

By symmetry, we can condition on C^* being the first K nodes. Let T denote the set of first $\lfloor \frac{\rho n}{\log^2 n} \rfloor$ nodes. Then

$$\min_{i\in C^*} e(i,C^*) \leq \min_{i\in T} e(i,C^*) \leq \min_{i\in T} e(i,C^*\backslash T) + \max_{i\in T} e(i,T). \tag{A.72}$$

Let E_1, E_2, E_3 denote the event that $\max_{i\in T} e(i,T) < \frac{\log n}{\log\log n}$, $\min_{i\in T} e(i,C^*\backslash T) + \frac{\log n}{\log\log n} \leq \tau^*\rho\log n$ and $\max_{j\notin C^*} e(j,C^*) \geq \tau^*\rho\log n$, respectively. In view of (A.72), we have $F \supset E_1\cap E_2\cap E_3$ and hence it boils down to proving that $\mathbb{P}\{E_i\}\to 1$ for $i = 1,2,3$.

In view of the following Chernoff bound for binomial distributions [97, Theorem 4.4]: For $r \geq 1$ and $X \sim \mathrm{Binom}(n,p)$, $\mathbb{P}\{X \geq rnp\} \leq (\mathrm{e}/r)^{rnp}$, we have

$$\mathbb{P}\left\{e(i,T) \geq \frac{\log n}{\log\log n}\right\} \leq \left(\frac{\log^2 n}{a\mathrm{e}\log\log n}\right)^{-\log n/\log\log n} = n^{-2+o(1)}.$$

Applying the union bound yields

$$\mathbb{P}\{E_1\} \geq 1 - \sum_{i\in T}\mathbb{P}\left\{e(i,T) \geq \frac{\log n}{\log\log n}\right\} \geq 1 - n^{-1+o(1)}.$$

[1]To be more precise, if there is an isolated node in the cluster C^*, then the likelihood has at least $n-K$ maximizers, which, in turn, implies that the probability of exact recovery for any estimator is at most $\frac{1}{n-K}$.

Moreover,

$$\begin{aligned}\mathbb{P}\left\{E_2\right\} &\overset{(a)}{=} 1-\prod_{i\in T}\mathbb{P}\left\{e(i,C^*\backslash T)>\tau^*\rho\log n-\frac{\log n}{\log\log n}\right\}\\ &\overset{(b)}{=} 1-\left(1-n^{-\rho\left(a-\tau^*\log\frac{\mathrm{e}a}{\tau^*}+o(1)\right)}\right)^{|T|}\overset{(c)}{\geq} 1-\exp\left(-n^{1-\rho\left(a-\tau^*\log\frac{\mathrm{e}a}{\tau^*}\right)+o(1)}\right)\\ &\overset{(d)}{\to} 1,\end{aligned}$$

where (a) holds because $\{e(i,C^*\backslash T)\}_{i\in T}$ are mutually independent; (b) follows from Lemma 16; (c) is due to $1+x\leq e^x$ for all $x\in\mathbb{R}$; (d) follows from the assumption (5.11) that $\rho f(a,b)=\rho(a-\tau^*\log\frac{\mathrm{e}a}{\tau^*})<1$. Similarly,

$$\begin{aligned}\mathbb{P}\left\{E_3\right\} &= 1-\prod_{j\notin C^*}\mathbb{P}\left\{e(j,C^*)<\tau^*\rho\log n\right\}\\ &= 1-\left(1-n^{-\left(b-\tau^*\log\frac{\mathrm{e}b}{\tau^*}+o(1)\right)}\right)^{n-K}\geq 1-\exp\left(-n^{1-\rho\left(b-\tau^*\log\frac{\mathrm{e}b}{\tau^*}\right)+o(1)}\right)\\ &\to 1,\end{aligned}$$

completing the proof in the case $a>b>0$.

Finally, we prove the theorem for the case $a<b$. Consider the case $a=0$ first. For $j\notin C^*$, since $e(j,C^*)\sim\mathrm{Binom}(K,\frac{b\log n}{n})$, it follows that $\log\mathbb{P}\left\{e(j,C^*)=0\right\}=K\log\left(1-\frac{b\log n}{n}\right)=-\left(\rho b+o(1)\right)\log n$, and thus

$$\begin{aligned}\mathbb{P}\left\{\min_{j\notin C^*}e(j,C^*)=0\right\} &= 1-\prod_{j\notin C^*}\left(1-\mathbb{P}\left\{e(j,C^*)=0\right\}\right)\\ &= 1-\left(1-n^{-\rho b+o(1)}\right)^{n-K}\geq 1-\exp(-n^{1-\rho b+o(1)})\to 1,\end{aligned}$$

due to the assumption that $\rho f(0,b)=\rho b<1$. Then with probability tending to one, there exists an isolated node $j\notin C^*$, in which case the likelihood has at least K maximizers and the probability of exact recovery for any estimator is at most $\frac{1}{K}$. Next assume that $0<a<b$. By symmetry, we condition on C^* being the first K nodes. Let T denote the set of first $\lfloor\frac{\rho n}{\log^2 n}\rfloor$ nodes. Redefine F,E_2,E_3 as the event that $\max_{i\in C^*}e(i,C^*)>\min_{j\notin C^*}e(j,C^*)$, $\max_{i\in T}e(i,C^*\backslash T)>\tau^*\rho\log n$ and $\min_{j\notin C^*}e(j,C^*)\leq\tau^*\rho\log n$, respectively. Then by the same reasoning

$\mathbb{P}\{\text{ML fails}\} \geq \mathbb{P}\{F\} \geq \mathbb{P}\{E_2 \cap E_3\}$. Applying Lemma 16, we obtain

$$\begin{aligned}\mathbb{P}\{E_2\} &= 1 - \prod_{i \in T} \mathbb{P}\{e(i, C^* \backslash T) \leq \tau^* \rho \log n\} \\ &= 1 - \left(1 - n^{-\rho\left(a - \tau^* \log \frac{ea}{\tau^*} + o(1)\right)}\right)^{|T|} \geq 1 - \exp\left(-n^{1-\rho\left(a-\tau^* \log \frac{ea}{\tau^*}\right)+o(1)}\right) \\ &\to 1,\end{aligned}$$

and similarly,

$$\begin{aligned}\mathbb{P}\{E_3\} &= 1 - \prod_{j \notin C^*} \mathbb{P}\{e(j, C^*) > \tau^* \rho \log n\} \\ &= 1 - \left(1 - n^{-\left(b - \tau^* \log \frac{eb}{\tau^*} + o(1)\right)}\right)^{n-K} \geq 1 - \exp\left(-n^{1-\rho\left(b-\tau^* \log \frac{eb}{\tau^*}\right)+o(1)}\right) \\ &\to 1,\end{aligned}$$

completing the proof for the case $0 < a < b$.

A.12 Spectrum of Erdős-Rényi Random Graph

Let A denote the adjacency matrix of an Erdős-Rényi random graph G, where nodes i and j are connected independently with probability p_{ij}. Then $\mathbb{E}[A_{ij}] = p_{ij}$. Let $p = \max_{ij} p_{ij}$ and assume $p \geq c_0 \frac{\log n}{n}$ for any constant $c_0 > 0$. We aim to show that $\|A - \mathbb{E}[A]\|_2 \leq c' \sqrt{np}$ with high probability for some constant $c' > 0$. To this end, we establish the following more general result where the entries need not be binary-valued.

Theorem A.12.1. *Let A denote a symmetric and zero-diagonal random matrix, where the entries $\{A_{ij} : i < j\}$ are independent and $[0, 1]$-valued. Assume that $\mathbb{E}[A_{ij}] \leq p$, where $c_0 \log n/n \leq p \leq 1 - c_1$ for arbitrary constants $c_0 > 0$ and $c_1 > 0$. Then for any $c > 0$, there exists $c' > 0$ such that for any $n \geq 1$, $\mathbb{P}\{\|A - \mathbb{E}[A]\|_2 \leq c' \sqrt{np}\} \geq 1 - n^{-c}$.*

Let $\mathcal{G}(n, p)$ denote the Erdős-Rényi random graph model with the edge probability $p_{ij} = p$ for all i, j. Results similar to Theorem A.12.1 have been obtained in [98] for the special case of $\mathcal{G}(n, \frac{c_0 \log n}{n})$ for some *sufficiently large* c_0. In fact, Theorem A.12.1 can be proved by strengthening the combinatorial arguments in [98, Section 2.2]. Here we provide an alternative proof using results from random ma-

trices and concentration of measures and a seconder-order stochastic comparison argument from [99].

Furthermore, we note that the condition $p = \Omega(\log n/n)$ in Theorem A.12.1 is in fact necessary to ensure that $\|A - \mathbb{E}[A]\|_2 = \Omega_{\mathbb{P}}(\sqrt{np})$ (see Appendix A.12.1 for a proof). The condition $p \leq 1 - c_1$ can be dropped in the special case of $\mathcal{G}(n,p)$.

Proof. We first use the second-order stochastic comparison arguments from [99, Lemma 2]. Since $0 \leq \mathbb{E}[A_{ij}] \leq p$, we have $A_{ij} - \mathbb{E}[A_{ij}] \in [-p, 1]$ for all $i \neq j$ and hence $B_{ij} \triangleq (1-p)(A_{ij} - \mathbb{E}[A_{ij}]) \in [-p, 1-p]$. Let C denote the adjacency matrix of a graph generated from $\mathcal{G}(n,p)$. Then, for any i, j, B_{ij} is stochastically smaller than $C_{ij} - \mathbb{E}[C_{ij}]$ under the convex ordering, i.e., $\mathbb{E}[f(B_{ij})] \leq \mathbb{E}[f(C_{ij} - \mathbb{E}[C_{ij}])]$ for any convex function f on $[-p, 1-p]$.[2] Since the spectral norm is a convex function and the coordinate random variables are independent (up to symmetry), it follows that $\mathbb{E}[\|B\|] \leq \mathbb{E}[\|C - \mathbb{E}[C]\|]$ and thus

$$\mathbb{E}[\|A - \mathbb{E}[A]\|] = \frac{1}{1-p}\mathbb{E}[\|B\|] \leq \frac{1}{1-p}\mathbb{E}[\|C - \mathbb{E}[C]\|] \leq \frac{1}{c_1}\mathbb{E}[\|C - \mathbb{E}[C]\|]. \tag{A.73}$$

We next bound $\mathbb{E}[\|C - \mathbb{E}[C]\|]$. Let $E = (E_{ij})$ denote an $n \times n$ matrix with independent entries drawn from $\mu \triangleq \frac{p}{2}\delta_1 + \frac{p}{2}\delta_{-1} + (1-p)\delta_0$, which is the distribution of a Rademacher random variable multiplied with an independent Bernoulli with bias p. Define E' as $E'_{ii} = E_{ii}$ and $E'_{ij} = -E_{ji}$ for all $i \neq j$. Let C' be an independent copy of C. Let D be a zero-diagonal symmetric matrix whose entries are drawn from μ and D' be an independent copy of D. Let $M = (M_{ij})$ denote an $n \times n$ zero-diagonal symmetric matrix whose entries are Rademacher and independent from C and C'. We apply the usual symmetrization arguments:

$$\begin{aligned}
\mathbb{E}[\|C - \mathbb{E}[C]\|] = \ \mathbb{E}[\|C - \mathbb{E}[C']\|] &\overset{(a)}{\leq} \mathbb{E}[\|C - C'\|] \overset{(b)}{=} \mathbb{E}[\|(C - C') \circ M\|] \\
&\overset{(c)}{\leq} 2\mathbb{E}[\|C \circ M\|] = \ 2\mathbb{E}[\|D\|] = 2\mathbb{E}[\|D - \mathbb{E}[D']\|] \\
&\overset{(d)}{\leq} 2\mathbb{E}[\|D - D'\|] \overset{(e)}{=} 2\mathbb{E}[\|E - E'\|] \overset{(f)}{\leq} 4\,\mathbb{E}[\|E\|],
\end{aligned} \tag{A.74}$$

[2] This follows from $\frac{f(1-p)-f(b)}{1-p-b} \geq f(1-p) - f(-p)$ for any $-p \leq b < 1-p$, by the convexity of f.

where $(a), (d)$ follow from the Jensen's inequality; (b) follows because $C - C'$ has the same distribution as $(C - C') \circ M$, where $\circ$ denotes the element-wise product; $(c), (f)$ follow from the triangle inequality; (e) follows from the fact that $D - D'$ has the same distribution as $E - E'$. Then, we apply the result of Seginer [100] which characterized the expected spectral norm of i.i.d. random matrices within universal constant factors. Let $X_j \triangleq \sum_{i=1}^{n} E_{ij}^2$, which are independent $\mathrm{Binom}(n, p)$. Since μ is symmetric, [100, Theorem 1.1] and Jensen's inequality yield

$$\mathbb{E}[\|E\|] \leq \kappa \, \mathbb{E}\left[\left(\max_{j\in[n]} X_j\right)^{1/2}\right] \leq \kappa \left(\mathbb{E}\left[\max_{j\in[n]} X_j\right]\right)^{1/2} \tag{A.75}$$

for some universal constant κ. In view of the following Chernoff bound for the binomial distribution [97, Theorem 4.4]:

$$\mathbb{P}\left\{X_1 \geq t \log n\right\} \leq 2^{-t},$$

for all $t \geq 6np$, setting $t_0 = 6\max\{np/\log n, 1\}$ and applying the union bound, we have

$$\begin{aligned}\mathbb{E}\left[\max_{j\in[n]} X_j\right] &= \int_0^\infty \mathbb{P}\left\{\max_{j\in[n]} X_j \geq t\right\}\mathrm{d}t \leq \int_0^\infty (n\,\mathbb{P}\left\{X_1 \geq t\right\} \wedge 1)\mathrm{d}t \\ &\leq t_0 \log n + n\int_{t_0 \log n}^\infty 2^{-t}\mathrm{d}t \leq (t_0+1)\log n \leq 6(1+2/c_0)np,\end{aligned} \tag{A.76}$$

where the last inequality follows from $np \geq c_0 \log n$. Assembling (A.73) – (A.76), we obtain

$$\mathbb{E}[\|A - \mathbb{E}[A]\|] \leq c_2\sqrt{np}, \tag{A.77}$$

for some positive constant c_2 depending only on c_0, c_1. Since the entries of $A - \mathbb{E}[A]$ are valued in $[-1, 1]$, Talagrand's concentration inequality for 1-Lipschitz convex functions (see, e.g., [101, Theorem 2.1.13]) yields

$$\mathbb{P}\left\{\|A - \mathbb{E}[A]\| \geq \mathbb{E}[\|A - \mathbb{E}[A]\|] + t\right\} \leq c_3 \exp(-c_4 t^2)$$

for some absolute constants c_3, c_4, which implies that for any $c > 0$, there exists $c' > 0$ depending on c_0, c_1, such that $\mathbb{P}\left\{\|A - \mathbb{E}[A]\| \geq c'\sqrt{np}\right\} \leq n^{-c}$. $\square$

A.12.1 The Sharpness of the Condition for Theorem A.12.1

Consider the case where A is the adjacency matrix of $\mathcal{G}(n,p)$. We show that if $np = o(\log n)$, then

$$\mathbb{P}\left\{\|A - \mathbb{E}[A]\|_2 \geq c\sqrt{\frac{\log n}{\log\left(\log n/(np)\right)}}\right\} \to 1 \tag{A.78}$$

for some constant c and, consequently, $\|A - \mathbb{E}[A]\|_2 / \sqrt{np} \to \infty$ in probability. To this end, note that $\|A - \mathbb{E}[A]\|_2 \geq \max_{i\in[n]} \|(A - \mathbb{E}[A])e_i\|_2$, where $\{e_i\}$ denote the standard basis. Without loss of generality, assume n is even. Then by focusing on the upper-right part of A, we have $\|A - \mathbb{E}[A]\|_2 \geq \sqrt{\max_{i\in[n/2]} X_i} - \sqrt{n/2}p$, where X_i are independently distributed as $\mathrm{Binom}(n/2, p)$. Using the inequality $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$ for $k \geq 1$,

$$\mathbb{P}\{X_1 \geq k\} \geq \mathbb{P}\{X_1 = k\} = \binom{n/2}{k} p^k (1-p)^{n/2-k} \geq \left(\frac{np}{2k}\right)^k (1-p)^{n/2}.$$

Since $\log(1-x) \geq -2x$ for $x \in [0, 1/2]$, it follows that

$$-\log \mathbb{P}\{X_1 \geq k\} \leq k \log\left(\frac{2k}{np}\right) - \frac{n}{2}\log(1-p) \leq k\log\left(\frac{2k}{np}\right) + np, \tag{A.79}$$

Plugging $k^* \triangleq \lfloor \frac{\log n}{\log(\log n/(np))} \rfloor$ into (A.79) and noting $np = o(k^*)$ and $\log\log n = o\left(\log n/\log\log n\right)$, we get

$$\begin{aligned} -\log \mathbb{P}\{X_1 \geq k^*\} &\leq \log n + np - \frac{\log n}{\log\left(\log n/(np)\right)} \log\left(\frac{\log\left(\log n/(np)\right)}{2}\right) \\ &\leq \log n - \log\log n. \end{aligned} \tag{A.80}$$

By the independence of $\{X_i, i \in [n/2]\}$, we have

$$\begin{aligned} \mathbb{P}\left\{\max_{i\in[n/2]} X_i < k^*\right\} &= \prod_{i=1}^{n/2} \mathbb{P}\{X_i < k^*\} = \left(1 - \mathbb{P}\{X_1 \geq k\}\right)^{n/2} \\ &\leq \exp\left(-\frac{n}{2}\mathbb{P}\{X_1 \geq k^*\}\right) \leq \frac{1}{\sqrt{n}}, \end{aligned}$$

where the last inequality follows in view of (A.80).

A.13 Tail of the Binomial Distribution

Let $X \sim \mathrm{Binom}\left(m, \frac{a \log n}{n}\right)$ and $R \sim \mathrm{Binom}\left(m, \frac{b \log n}{n}\right)$ for $m \in \mathbb{N}$ and $a, b > 0$, where $m = \rho n + o(n)$ for some $\rho > 0$ as $n \to \infty$. We need the following tail bounds.

Lemma 15 ([52]). *Assume that $a > b$ and $k_n \in \mathbb{N}$ such that $k_n = (1+o(1))\frac{\log n}{\log \log n}$. Then*

$$\mathbb{P}\{X - R \le k_n\} \le n^{-\rho\left(\sqrt{a}-\sqrt{b}\right)^2 + o(1)}.$$

Lemma 16. *Let $k_n, k_n' \in [m]$ be such that $k_n = \tau\rho \log n + o(\log n)$ and $k_n' = \tau'\rho \log n + o(\log n)$ for some $0 \le \tau \le a$ and $\tau' \ge b$. Then*

$$\mathbb{P}\{X \le k_n\} = n^{-\rho\left(a - \tau \log \frac{ea}{\tau} + o(1)\right)} \tag{A.81}$$

$$\mathbb{P}\{R \ge k_n'\} = n^{-\rho\left(b - \tau' \log \frac{eb}{\tau'} + o(1)\right)}. \tag{A.82}$$

Proof. We use the following non-asymptotic bound on the binomial tail probability [102, Lemma 4.7.2]: For $U \sim \mathrm{Binom}(n, p)$,

$$(8k(1-\lambda))^{-1/2} \exp(-nd(\lambda\|p)) \le \mathbb{P}\{U \ge k\} \le \exp(-nd(\lambda\|p)), \tag{A.83}$$

where $\lambda = \frac{k}{n} \in (0,1)$ and $d(\lambda\|p) = \lambda \log \frac{\lambda}{p} + (1-\lambda)\log \frac{1-\lambda}{1-p}$ is the binary divergence function. Then (A.82) follows from (A.83) by noting that $d(\frac{k_n'}{m} \| \frac{b \log n}{n}) = (b - \tau' \log \frac{be}{\tau'} + o(1))\frac{\log n}{n}$.

To prove (A.81), we use the following bound on binomial coefficients [102, Lemma 4.7.1]:

$$\frac{\sqrt{\pi}}{2} \le \frac{\binom{n}{k}}{(2\pi n\lambda(1-\lambda))^{-1/2} \exp(nh(\lambda))} \le 1. \tag{A.84}$$

where $\lambda = \frac{k}{n} \in (0,1)$ and $h(\lambda) = -\lambda \log \lambda - (1-\lambda)\log(1-\lambda)$ is the binary entropy function. Note that the mode of X is at $\lfloor (m+1)p \rfloor = (a\rho + o(1)) \log n$, which is at least k_n for sufficiently large n. Therefore, $\mathbb{P}\{X = k\}$ is non-decreasing in k for $k \in [0, k_n]$ and hence

$$\mathbb{P}\{X = k_n\} \le \mathbb{P}\{X \le k_n\} \le k_n \mathbb{P}\{X = k_n\}, \tag{A.85}$$

where $\mathbb{P}\{X = k_n\} = \binom{m}{k_n} p_n^k (1-p)^{m-k_n}$ and $p = a \log n / n$. Applying (A.84) to

(A.85) yields

$$\mathbb{P}\{X \leq k_n\} = (\log n)^{O(1)} \exp(-nd(k_n/m\|p)),$$

which is the desired (A.81). □

A.14 Proof of Theorem 6.1.1

Let $X \circ Y$ denote the Hadamard (entrywise) product between two matrices X, Y. In addition to the ℓ_1 matrix norm $\|X\|_1 = \sum_{i,j} |X_{ij}|$, we also use the weighted ℓ_1 norm $\|X\|_{1,\Theta} = \sum_{ij} \Theta_{ij}|X_{ij}|$, where $\Theta = \theta\theta^\top$. Define $\Theta^{\frac{1}{2}}$ with (i,j)-th entry given by $\sqrt{\Theta_{ij}}$. Similarly define $\Theta^{-\frac{1}{2}}$. Let $U \in \mathbb{R}^{n\times r}$ be the weighted characteristic matrix for the clusters, i.e.,

$$U_{ik} = \begin{cases} \frac{\sqrt{\theta_i}}{\sqrt{\|\theta^{(k)}\|_1}} & \text{if node } i \in C_k^* \\ 0 & \text{otherwise,} \end{cases}$$

Let $\Sigma \in \mathbb{R}^{r\times r}$ be the diagonal matrix with $\Sigma_{kk} = \|\theta^{(k)}\|_1$ for $k \in [r]$. The weighted true cluster matrix $Y^* \circ \Theta^{\frac{1}{2}}$ has the rank-r singular value decomposition given by $Y^* \circ \Theta^{\frac{1}{2}} = U\Sigma U^\top$. Define the projections $\mathcal{P}_T(M) = UU^\top M + MUU^\top - UU^\top MUU^\top$ and $\mathcal{P}_{T^\perp}(M) = M - \mathcal{P}_T(M)$.

To establish the theorem, it suffices to show for any feasible solution Y with $Y \neq Y^*$, $\Delta(Y) \triangleq \langle Y^* - Y, A - \lambda\Theta\rangle > 0$. Note that $\mathbb{E}[A] \triangleq (qJ + (p-q)Y^* - pI) \circ \Theta$, where J is the all-ones matrix, and I is the identity matrix. Then, we can decompose $\Delta(Y)$ as

$$\Delta(Y) = \langle \mathbb{E}[A] - \lambda\Theta, Y^* - Y\rangle + \langle A - \mathbb{E}[A], Y^* - Y\rangle. \tag{A.86}$$

The first term in (A.21) can be written as

$$\begin{aligned} &\langle \mathbb{E}[A] - \lambda\Theta, Y^* - Y\rangle \\ &\overset{(a)}{=} (p-\lambda)\langle Y^* \circ \Theta, Y^* - Y\rangle + (\lambda - q)\langle (J - Y^*) \circ \Theta, Y - Y^*\rangle \\ &\overset{(b)}{\geq} \frac{p-q}{4}\|Y^* - Y\|_{1,\Theta}, \end{aligned} \tag{A.87}$$

where (a) holds in view of the definition of $\mathbb{E}[A]$ and the fact that $Y^*_{ii} = Y_{ii} = 1$

for all i; (b) holds because $\frac{1}{4}p + \frac{3}{4}q \leq \lambda \leq \frac{3}{4}p + \frac{1}{4}q$, $Y^*_{ij} \in \{0,1\}$ and $Y_{ij} \in [0,1]$ for all i, j. Next we control the second term in (A.86). Define the weighted noise matrix $W = (A - \mathbb{E}[A]) \circ \Theta^{-\frac{1}{2}}$. By matrix Bernstein inequality, with high probability,

$$\|W\| \leq c_2\sqrt{(p(1-p)K + q(1-q)n)\log n} + c_2\frac{\log n}{\theta_{\min}}.$$

Thus $UU^\top + \mathcal{P}_{T^\perp}\left(\frac{W}{\|W\|}\right)$ is a subgradient of $||X||_*$ at $X = Y^* \circ \Theta^{\frac{1}{2}}$. Note that if $X \succeq 0$, then $\|X\|_* = \mathrm{Tr}(X)$. Hence,

$$\begin{aligned} 0 \geq \mathrm{Tr}\left(Y \circ \Theta^{\frac{1}{2}}\right) - \mathrm{Tr}\left(Y^* \circ \Theta^{\frac{1}{2}}\right) &= \|Y \circ \Theta^{\frac{1}{2}}\|_* - \|Y^* \circ \Theta^{\frac{1}{2}}\|_* \\ &\geq \langle UU^\top + \mathcal{P}_{T^\perp}\left(\frac{W}{\|W\|}\right), (Y - Y^*) \circ \Theta^{\frac{1}{2}}\rangle. \end{aligned}$$

It follows that $\langle \mathcal{P}_{T^\perp}(W), (Y^* - Y) \circ \Theta^{\frac{1}{2}}\rangle \geq \|W\|\langle UU^\top, (Y - Y^*) \circ \Theta^{\frac{1}{2}}\rangle$. Therefore, the second term in (A.21) can be bounded as

$$\begin{aligned} \langle A - \mathbb{E}[A], Y^* - Y\rangle = \langle W, (Y^* - Y) \circ \Theta^{\frac{1}{2}}\rangle &\geq \langle \mathcal{P}_T(W) - \|W\|UU^\top, (Y^* - Y) \circ \Theta^{\frac{1}{2}}\rangle \\ &\geq -\left(\|W\|||UU^\top||_{\infty,\Theta^{-\frac{1}{2}}} + ||\mathcal{P}_T(W)||_{\infty,\Theta^{-\frac{1}{2}}}\right)\|Y^* - Y\|_{1,\Theta} \\ &\geq -\left(\frac{\|W\|}{\min_k \|\theta^{(k)}\|_1} + ||\mathcal{P}_T(W)||_{\infty,\Theta^{-\frac{1}{2}}}\right)\|Y^* - Y\|_{1,\Theta}, \end{aligned} \tag{A.88}$$

where the last inequality follows from the definition of U. Below we bound the term $||\mathcal{P}_T(W)||_{\infty,\Theta^{-\frac{1}{2}}}$. From the definition of $\mathcal{P}_T$,

$$||\mathcal{P}_T(W)||_{\infty,\Theta^{-\frac{1}{2}}} \leq ||UU^\top W||_{\infty,\Theta^{-\frac{1}{2}}} + ||WUU^\top||_{\infty,\Theta^{-\frac{1}{2}}} + ||UU^\top WUU^\top||_{\infty,\Theta^{-\frac{1}{2}}}.$$

We bound $||UU^\top W||_{\infty,\Theta^{-\frac{1}{2}}}$ below. To bound the term $(UU^\top W)_{ij}$, assume user i belongs to cluster k and recall C^*_k is the set of users in cluster k. Then

$$(UU^\top W)_{ij} = \frac{\sqrt{\theta_i}}{\|\theta^{(k)}\|_1}\sum_{i' \in \mathcal{C}_k}\sqrt{\theta_{i'}}W_{i'j},$$

which is the weighted average of independent random variables. Define the noise

variance $\sigma^2 \triangleq p(1-q)$. By Bernstein's inequality, with probability at least $1-n^{-3}$,

$$\left|\sum_{i' \in \mathcal{C}_k} \sqrt{\theta_{i'}} W_{i'j}\right| \leq \sqrt{6\sigma^2 \|\theta^{(k)}\|_1 \log n} + \frac{2\log n}{\sqrt{\theta_{\min}}} \leq c_2 \sigma \sqrt{\|\theta^{(k)}\|_1 \log n},$$

where the last inequality holds because by assumption (6.4), $\sigma^2 \min_k \|\theta^{(k)}\|_1 \theta_{\min} \gtrsim \log n$. Then with probability at least $1 - n^{-1}$,

$$||UU^\top W||_{\infty, \Theta^{-\frac{1}{2}}} \leq c_1 \sigma \sqrt{\frac{\log n}{\min_k \|\theta^{(k)}\|_1 \theta_{\min}}}.$$

Similarly we bound $||WUU^\top||_{\infty, \Theta^{-\frac{1}{2}}}$ and $||UU^\top WUU^\top||_{\infty, \Theta^{-\frac{1}{2}}}$. Therefore, with probability at least $1 - 3n^{-1}$,

$$||\mathcal{P}_T(W)||_{\infty, \Theta^{-\frac{1}{2}}} \leq 3c_2 \sigma \sqrt{\frac{\log n}{\min_k \|\theta^{(k)}\|_1 \theta_{\min}}}. \tag{A.89}$$

Substituting (A.89) into (A.88) and by assumption (6.4) and (6.5), we conclude that with high probability $\Delta(Y) > 0$ for any feasible $Y \neq Y^*$.

A.15 Proof of Lemma 1

Note that $d_i = \sum_j A_{ij}$. For all $j \neq i$, A_{ij} is mutually independent, and distributed as $\mathrm{Bern}(\theta_i \theta_j p)$ if $Y^*_{ij} = 1$ and $\mathrm{Bern}(\theta_i \theta_j q)$ if $Y^*_{ij} = 0$. Let

$$\bar{d}_i \triangleq \mathbb{E}[d_i] = \theta_i \left[(p-q)\|\theta^{(k(i))}\|_1 + q\|\theta\|_1 - p\theta_i\right]. \tag{A.90}$$

Then, by the law of large numbers, $d_i \to \bar{d}_i$ for all $i \in [n]$ and thus

$$\frac{d_i d_j}{\theta_i \theta_j \sum_{i'} d_{i'}} \to \frac{\bar{d}_i \bar{d}_j}{\theta_i \theta_j \sum_{i'} \bar{d}_{i'}}. \tag{A.91}$$

Plugging (A.90) into (A.91) and taking the limit $n \to \infty$, the lemma follows.

APPENDIX B

PROOFS FOR INFERRING PREFERENCES

We introduce some additional notations used in the proof. For a vector x, let $\|x\|_2$ denote the usual l_2 norm. Let $\mathbf{1}$ denote the all-one vector and $\mathbf{0}$ denote the all-zero vector with the appropriate dimension. Let $\mathcal{S}^n$ denote the set of $n \times n$ symmetric matrices with real-valued entries. For $X \in \mathcal{S}^n$, let $\lambda_1(X) \leq \lambda_2(X) \leq \cdots \leq \lambda_n(X)$ denote its eigenvalues sorted in increasing order. Let $\mathrm{Tr}(X) = \sum_{i=1}^n \lambda_i(X)$ denote its trace and $\|X\| = \max\{-\lambda_1(X), \lambda_n(X)\}$ denote its spectral norm. For two matrices $X, Y \in \mathcal{S}^n$, we write $X \leq Y$ if $Y - X$ is positive semi-definite, i.e., $\lambda_1(Y - X) \geq 0$. Recall that $\mathcal{L}(\theta)$ is the log likelihood function. The first-order partial derivative $\nabla_i \mathcal{L}(\theta)$ for any $i \in [n]$, is given by

$$\nabla_i \mathcal{L}(\theta) = \sum_{j: i \in S_j} \sum_{\ell=1}^{k_j - 1} \mathbf{1}_{\left\{\sigma_j^{-1}(i) \geq \ell\right\}} \left[\mathbf{1}_{\{\sigma_j(\ell) = i\}} - \frac{\exp(\theta_i)}{\exp(\theta_{\sigma_j(\ell)}) + \cdots + \exp(\theta_{\sigma_j(k_j)})} \right], \tag{B.1}$$

and the Hessian matrix $H(\theta) \in \mathcal{S}^n$ with $H_{ii'}(\theta) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i \partial \theta_{i'}}$ is given by

$$H(\theta) = -\frac{1}{2} \sum_{j=1}^m \sum_{i, i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \sum_{\ell=1}^{k_j - 1} \frac{\exp(\theta_i + \theta_{i'}) \mathbf{1}_{\left\{\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell\right\}}}{[\exp(\theta_{\sigma_j(\ell)}) + \cdots + \exp(\theta_{\sigma_j(k_j)})]^2}. \tag{B.2}$$

It follows from the definition that $-H(\theta)$ is positive semi-definite for any $\theta \in \mathbb{R}^n$. Define $L_j \in S^n$ as

$$L_j = \frac{1}{2(k_j - 1)} \sum_{i, i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top,$$

and then the Laplacian of the pairwise comparison graph G satisfies $L = \sum_{j=1}^m L_j$.

B.1 Proof of Theorem 7.3.1

We first introduce a key auxiliary result used in the proof. Let F be a fixed CDF (to be used in the Thurstone model), let $b > 0$ and suppose θ is a parameter to be estimated with $\theta \in [-b, b]$ from observation $U = (U_1, \ldots, U_d)$, where the U_i's are independent with the common CDF given by $F(c-\theta)$. The following proposition gives a lower bound on the average MSE for a fixed prior distribution based on Van Trees inequality [88].

Proposition B.1.1. *Let p_0 be a probability density on $[-1, 1]$ such that $p_0(1) = p_0(-1) = 0$ and define the prior density of Θ as $p(\theta) = \frac{1}{b} p_0(\frac{\theta}{b})$. Then for any estimator $T(U)$ of Θ,*

$$E[(\Theta - T(U))^2] \geq \frac{1}{d} \frac{1}{I(\mu) + I(p_0)/(b^2 d)},$$

where μ is the probability density function of F with $I(\mu) = \int \frac{(\mu'(x))^2}{\mu(x)} dx$ and $I(p_0) = \int_{-1}^{1} \frac{\left(p_0'(\theta)\right)^2}{p_0(\theta)} d\theta$.

Proof. It follows from the Van Trees inequality that

$$E[(\Theta - T(U))^2] \geq \frac{1}{\int I(\theta) p(\theta) d\theta + I(p)},$$

where the Fisher information $I(\theta) = dI(\mu)$ and

$$I(p) = \int_{-b}^{b} \frac{\left(p'(\theta)\right)^2}{p(\theta)} d\theta = \frac{1}{b^2} \int_{-1}^{1} \frac{\left(p_0'(\theta)\right)^2}{p_0(\theta)} d\theta = \frac{1}{b^2} I(p_0).$$

□

Proof of Theorem 7.3.1. Let $\widehat{\theta}$ be a given estimator. The minimax MSE for $\widehat{\theta}$ is greater than or equal to the average MSE for a given prior distribution on θ^*. Let $p_0(\theta) = \cos^2(\pi\theta/2)$, then $I(p_0) = \pi^2$. Define $p(\theta) = \frac{1}{b} p_0(\frac{\theta}{b})$. If n is even we use the following prior distribution. The prior distribution of θ_i^* for i odd is $p(\theta)$ and for i even, $\theta_i^* \equiv -\theta_{i-1}^*$. If n is odd use the same distribution for θ_1^* through θ_{n-1}^* and set $\theta_n^* \equiv 0$. Note that $\theta^* \in \Theta_b$ with probability one. For simplicity, we assume n is odd in the rest of this proof; the modification for n even is trivial. We use the genie argument, so that the observer can see the hidden utilities in the Thurstone model. The estimation of θ^* decouples into $\lfloor \frac{n}{2} \rfloor$ disjoint problems, so we can focus

on the estimation of θ_1 from the vector of random variables $U = (U_1, \ldots, U_{d_1})$ associated with item 1 and the vector of random variables $V = (V_1, \ldots, V_{d_2})$ associated with item 2. The distribution functions of the U_i's are all $F(c-\theta_1^*)$ and the distribution functions of the V_i's are all $F(c+\theta_1^*)$, and the U's and V's are all mutually independent given θ^*. Recall that μ is the probability density function of F, i.e., $\mu = F'$. The Fisher information for each of the $d_1 + d_2$ observations is $I(\mu)$, so that Proposition B.1.1 carries over to this situation with $d = d_1 + d_2$. Therefore, for any estimator $T(U, V)$ of Θ_1^* (the random version of θ_1^*),

$$E[(\Theta_1^* - T(U,V))^2] \geq \frac{1}{d_1 + d_2} \frac{1}{I(\mu) + \pi^2/(b^2(d_1+d_2))}.$$

By this reasoning, for any odd value of i with $1 \leq i < n$ we have

$$\begin{aligned} E[(\widehat{\theta}_i - \theta_i^*)^2] + E[(\widehat{\theta}_{i+1} - \theta_{i+1}^*)^2] &\geq \frac{2}{I(\mu) + \pi^2/(b^2(d_1+d_2))} \frac{1}{d_i + d_{i+1}} \\ &\geq \frac{1}{2I(\mu) + 2\pi^2/(b^2(d_1+d_2))} \left(\frac{1}{d_{i+1}} + \frac{1}{d_{i+2}} \right). \end{aligned}$$

Summing over all odd values of i in the range $1 \leq i < n$ yields the theorem. Furthermore, since $\sum_{i=1}^n d_i = mk$, by Jensen's inequality, $\sum_{i=2}^n \frac{1}{d_i} \geq \frac{(n-1)^2}{\sum_{i=2}^n d_i} \geq \frac{(n-1)^2}{mk}$. □

B.2 Proof of Theorem 7.4.1

The Fisher information matrix is defined as $I(\theta) = -\mathbb{E}_\theta[H(\theta)]$ and given by

$$I(\theta) = \frac{1}{2} \sum_{j=1}^m \sum_{i,i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \sum_{l=1}^{k_j - 1} \mathbb{P}_\theta[\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell] \frac{e^{\theta_i + \theta_{i'}}}{[e^{\theta_{\sigma_j(\ell)}} + \cdots + e^{\theta_{\sigma_j(k_j)}}]^2}.$$

Since $-H(\theta)$ is positive semi-definite, it follows that $I(\theta)$ is positive semi-definite. Moreover, $\lambda_1(I(\theta))$ is zero and the corresponding eigenvector is the normalized all-one vector. Fix any unbiased estimator $\widehat{\theta}$ of $\theta \in \Theta_b$. Since $\widehat{\theta} \in \mathcal{U}$, $\widehat{\theta} - \theta$ is orthogonal to $\mathbf{1}$. The Cramér-Rao lower bound then implies that $\mathbb{E}[\|\widehat{\theta} - \theta\|^2] \geq$

$\sum_{i=2}^{n} \frac{1}{\lambda_i(I(\theta))}$. Taking the supremum over both sides gives

$$\sup_{\theta} \mathbb{E}[\|\widehat{\theta} - \theta\|^2] \geq \sup_{\theta} \sum_{i=2}^{n} \frac{1}{\lambda_i(I(\theta))} \geq \sum_{i=2}^{n} \frac{1}{\lambda_i(I(0))}.$$

If θ equals the all-zero vector, then

$$\mathbb{P}[\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell] = \frac{(k_j - 2)(k_j - 3) \cdots (k_j - \ell)}{k_j(k_j - 1) \cdots (k_j - \ell + 2)} = \frac{(k_j - \ell + 1)(k_j - \ell)}{k_j(k_j - 1)}.$$

It follows from the definition that

$$\begin{aligned} I(0) &= \frac{1}{2} \sum_{j=1}^{m} \sum_{i,i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \sum_{l=1}^{k_j - 1} \frac{k_j - \ell}{k_j(k_j - 1)(k_j - \ell + 1)} \\ &\leq \left(1 - \frac{1}{k_{\max}} \sum_{\ell=1}^{k_{\max}} \frac{1}{\ell}\right) L. \end{aligned}$$

By Jensen's inequality,

$$\sum_{i=2}^{n} \frac{1}{\lambda_i} \geq \frac{(n-1)^2}{\sum_{i=2}^{n} \lambda_i} = \frac{(n-1)^2}{\mathrm{Tr}(L)} = \frac{(n-1)^2}{\sum_{i=1}^{n} d_i} = \frac{(n-1)^2}{mk}.$$

B.3 Proof of Theorem 7.5.1

The main idea of the proof is inspired from the proof of [27, Theorem 4]. We first introduce several key auxiliary results used in the proof. Observe that $\mathbb{E}_{\theta^*}[\nabla L(\theta^*)] = 0$. The following lemma upper bounds the deviation of $\nabla L(\theta^*)$ from its mean.

Lemma 17. *With probability at least* $1 - \frac{2e^2}{n}$,

$$\|\nabla \mathcal{L}(\theta^*)\|_2 \leq \sqrt{2mk \log n}. \tag{B.3}$$

Proof. The idea of the proof is to view $\nabla \mathcal{L}(\theta^*)$ as the final value of a discrete time vector-valued martingale with values in $\mathbb{R}^n$. Consider a user that ranks items $1, \ldots, k$. The PL model for the ranking can be generated in a series of $k - 1$ rounds. In the first round, the top rated item for the user is found. Suppose it is item I. This contributes the term $e_I - (p_1, p_2, \ldots, p_k, 0, 0, \ldots, 0)$ to $\nabla \mathcal{L}(\theta^*)$, where $p_i = P\{I = i\}$. This contribution is a mean zero random vector in $\mathbb{R}^n$ and

its norm is less than one. For notational convenience, suppose $I = k$. In the second round, item k is removed from the competition, and an item J is to be selected at random from among $\{1, \ldots, k-1\}$. If q_j denotes $P\{J = j\}$ for $1 \leq j \leq k-1$, then the contribution of the second round for the user to $\nabla \mathcal{L}(\theta^*)$ is the random vector $e_J - (q_1, q_2, \ldots, q_{k-1}, 0, 0, \ldots, 0)$, which has conditional mean zero (given I) and norm less than or equal to one. Considering all m users and $k_j - 1$ rounds for user j, we see that $\nabla \mathcal{L}(\theta^*)$ is the value of a discrete-time martingale at time $m(k-1)$ such that the martingale has initial value zero and increments with norm bounded by one. By the vector version of the Azuma-Hoeffding inequality found in [103, Theorem 1.8] we have

$$\mathbb{P}\{\|\nabla \mathcal{L}(\theta^*)\| \geq \delta\} \leq 2e^2 e^{-\frac{\delta^2}{2m(k-1)}},$$

which implies the result. □

Observe that $-H(\theta)$ is positive semi-definite with the smallest eigenvalue equal to zero. The following lemma lower bounds its second smallest eigenvalue.

Lemma 18. *Fix any $\theta \in \Theta_b$. Then*

$$\lambda_2\left(-H(\theta)\right) \geq \begin{cases} \frac{e^{2b}}{(1+e^{2b})^2}\lambda_2 & \text{If } k = 2, \\ \frac{1}{4e^{4b}}\left(\lambda_2 - 16e^{2b}\sqrt{\lambda_n \log n}\right) & \text{If } k > 2, \end{cases} \tag{B.4}$$

where the inequality holds with probability at least $1 - n^{-1}$ in the case with $k > 2$.

Proof. **Case** $k_j = 2, \forall j \in [m]$: The Hessian matrix simplifies as

$$H(\theta) = -\frac{1}{2}\sum_{j=1}^{m}\sum_{i,i' \in S_j}(e_i - e_{i'})(e_i - e_{i'})^\top \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_{i'})}\frac{\exp(\theta_{i'})}{\exp(\theta_i) + \exp(\theta_{i'})}.$$

Observe that $H(\theta)$ is deterministic given S_1^m. Since $|\theta_i| \leq b, \forall i \in [n]$,

$$\frac{\exp(\theta_i)\exp(\theta_{i'})}{\left[\exp(\theta_i) + \exp(\theta_{i'})\right]^2} \geq \frac{e^{2b}}{(1+e^{2b})^2}.$$

It follows that $-H(\theta) \geq \frac{e^{2b}}{(1+e^{2b})^2}L$ and the theorem follows.
Case $k_j > 2$ **for some** $j \in [m]$: We first introduce a key auxiliary result used in the proof.

Claim B.3.1. *Given $\theta \in \mathbb{R}^r$, let $A = diag(p) - pp^T$, where p is the column probability vector with*
$p_i = e^{\theta_i}/(e^{\theta_1} + \cdots + e^{\theta_r})$ for each i. If $|\theta_i| \leq b$, for $1 \leq i \leq r$, then $\lambda_2(A) \geq \frac{1}{re^{2b}}$. Equivalently, $e^{2b}A \geq B$ where $B = \frac{1}{r}diag(\mathbf{1}) - \frac{1}{r^2}\mathbf{1}\mathbf{1}^\top$.

Proof. Fix θ satisfying the conditions of the lemma. It is easy to see that for each i, $p_i \geq \frac{1}{re^{2b}}$. The matrix A is positive semidefinite, and its smallest eigenvalue is zero, with the corresponding eigenvector $\mathbf{1}$. So $\lambda_2(A) = \min_\alpha \alpha^T A \alpha$ subject to the constraints $\alpha^T \mathbf{1} = 0$ and $\|\alpha\|^2 = 1$. For α satisfying the constraints,

$$\begin{aligned}
\alpha^T A \alpha &= \sum_i \alpha_i^2 p_i - \left(\sum_j \alpha_j p_j\right)^2 = \sum_i \left(\alpha_i - \sum_j \alpha_j p_j\right)^2 p_i \\
&= \min_c \sum_{i=1}^r (\alpha_i - c)^2 p_i \geq \min_c \sum_{i=1}^r (\alpha_i - c)^2 \frac{1}{re^{2b}} \\
&= \sum_{i=1}^r \alpha_i^2 \frac{1}{re^{2b}} = \frac{1}{re^{2b}}.
\end{aligned}$$

The proof of the first part of the lemma is complete. We remark that the bound of the lemma is nearly tight for the case $\theta_1 = \ldots = \theta_{r-1} = b$ and $\theta_r = -b$, for which $\lambda_2(A) = \frac{e^{2b}r}{((r-1)e^{2b}+1)^2}$. The final equivalence mentioned in the lemma follows from the facts $\lambda_1(e^{2b}A) = \lambda_1(B) = 0$ with common corresponding eigenvector $\mathbf{1}$, and $\lambda_i(e^{2b}A) \geq \frac{1}{r} = \lambda_i(B)$ for $2 \leq i \leq r$. □

The Hessian matrix $H(\theta)$ depends on σ_1^m and therefore is random given S_1^m. For a given user j, and ℓ with $1 \leq \ell \leq k_j - 1$, let $S^{(j,\ell)}$ denote the set of items contending for the ℓ^{th} position in the ranking of user j after higher ranking items have been selected: $S^{(j,\ell)} = \{i : \sigma_j^{-1}(i) \geq \ell\}$, let $\mathbf{1}^{(j,\ell)}$ denote the indicator vector for the set $S^{(j,\ell)}$, and let $p^{(j,\ell)}$ denote the corresponding probability column vector for the selection:

$$p_i^{(j,\ell)} = P(\sigma_j(\ell) = i | \sigma_j(1), \ldots, \sigma_j(\ell-1)) = \frac{\mathbf{1}_i^{(j,\ell)} e^{\theta_i}}{\sum_{i' \in S_{j,\ell}} e^{\theta_{i'}}}.$$

The Hessian can be written as $H(\theta) = \sum_{j=1}^m \sum_{\ell=1}^{k_j - 1} H^{(j,\ell)}$ where

$$-H^{(j,\ell)} = \frac{1}{2} \sum_{i,i' \in S^{(j,\ell)}} (e_i - e_{i'})(e_i - e_{i'})^\top p_i^{(j,\ell)} p_{i'}^{(j,\ell)} = \text{diag}(p^{(j,\ell)}) - p^{(j,\ell)}(p^{(j,\ell)})^\top.$$

By Claim B.3.1 applied to the restriction of $-H^{(j,\ell)}$ to $S^{(j,\ell)} \times S^{(j,\ell)}$,

$$\begin{aligned} -e^{2b}H^{(j,\ell)} &\geq \frac{1}{k_j - \ell + 1}\mathrm{diag}(\mathbf{1}^{(j,\ell)}) - \frac{1}{(k_j-\ell+1)^2}\mathbf{1}^{(j,\ell)}(\mathbf{1}^{(j,\ell)})^\top \\ &= \frac{1}{2(k_j-\ell+1)^2}\sum_{i,i' \in S^{(j,\ell)}} (e_i - e_{i'})(e_i - e_{i'})^\top. \end{aligned} \tag{B.5}$$

Summing over j and ℓ in (B.5) and noting that $k_j - \ell + 1 \leq k_j$ for all j, ℓ yields

$$-e^{2b}H(\theta) \geq \frac{1}{2}\sum_{j=1}^{m}\sum_{i,i' \in S_j}(e_i - e_{i'})(e_i - e_{i'})^\top \frac{1}{k_j^2}\sum_{\ell=1}^{k_j-1}\mathbf{1}_{\left\{\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell\right\}} := \widetilde{L}. \tag{B.6}$$

Observe that

$$\begin{aligned} \sum_{\ell=1}^{k_j-1}\mathbb{P}_\theta\left[\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell\right] &= 1 + \sum_{i'' \in S_j}\mathbf{1}_{\{i'' \neq i, i'\}}\frac{e^{\theta_{i''}}}{e^{\theta_i} + e^{\theta_{i'}} + e^{\theta_{i''}}} \\ &\geq 1 + \frac{k_j - 2}{2e^{2b}+1} \geq \frac{k_j+1}{3e^{2b}}. \end{aligned}$$

Recall that L is the Laplacian of G and $L = \sum_{j=1}^m L_j$. It follows that

$$\begin{aligned} \mathbb{E}_\theta[\widetilde{L}] &= \frac{1}{2}\sum_{j=1}^{m}\sum_{i,i' \in S_j}(e_i - e_{i'})(e_i - e_{i'})^\top \frac{1}{k_j^2}\sum_{\ell=1}^{k_j-1}\mathbb{P}_\theta[\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell] \\ &\geq \frac{1}{2}\sum_{j=1}^{m}\sum_{i,i' \in S_j}(e_i - e_{i'})(e_i - e_{i'})^\top \frac{k_j+1}{3e^{2b}k_j^2} \\ &\geq \frac{1}{2}\sum_{j=1}^{m}\sum_{i,i' \in S_j}(e_i - e_{i'})(e_i - e_{i'})^\top \frac{1}{4e^{2b}(k_j-1)} = \frac{1}{4e^{2b}}L. \end{aligned} \tag{B.7}$$

Define $a_{ii'} = \frac{1}{k_j^2}\sum_{\ell=1}^{k_j-1}\left(\mathbf{1}_{\left\{\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell\right\}} - \mathbb{P}_\theta[\sigma_j^{-1}(i), \sigma_j^{-1}(i') \geq \ell]\right)$. Then

$$\widetilde{L} - \mathbb{E}_\theta[\widetilde{L}] = \frac{1}{2}\sum_{j=1}^{m}\left(\sum_{i,i' \in S_j} a_{ii'}(e_i - e_{i'})(e_i - e_{i'})^\top\right) := \sum_{j=1}^{m} Y_j.$$

Observe that $|a_{ii'}| \leq \frac{1}{k_j}$ and therefore $-\frac{(k_j-1)}{k_j}L_j \leq Y_j \leq \frac{(k_j-1)}{k_j}L_j$. Furthermore, $\|L_j\| = \frac{k_j}{k_j-1}$ and thus $\|Y_j\| \leq 1$. Moreover, $Y_j^2 = \sum_{i,i',i'' \in S_j} a_{ii'}a_{ii''}(e_i - e_{i'})(e_i -$

$e_{i''})^\top$. It follows that for any vector $x \in \mathbb{R}^n$,

$$x^\top Y_j^2 x = \sum_{i,i',i'' \in S_j} a_{ii'} a_{ii''} (x_i - x_{i'})(x_i - x_{i''}) \le \frac{1}{k_j^2} \sum_{i,i',i'' \in S_j} |x_i - x_{i'}||x_i - x_{i''}|$$
$$= \frac{1}{k_j^2} \sum_{i \in S_j} \left(\sum_{i' \in S_j} |x_i - x_{i'}| \right)^2 \le \frac{1}{k_j} \sum_{i,i' \in S_j} (x_i - x_{i'})^2 = 2x^\top L_j x,$$

where the last inequality follows from the Cauchy-Swartz inequality. Therefore, $Y_j^2 \le 2L_j$. It follows that $\sum_{j=1}^m \mathbb{E}_\theta[Y_j^2] \le 2L$ and thus $\| \sum_{j=1}^m \mathbb{E}_\theta[Y_j^2] \| \le 2\lambda_n$. By the matrix Bernstein inequality [91], with probability at least $1 - n^{-1}$,

$$\|\widetilde{L} - \mathbb{E}_\theta[\widetilde{L}]\| \le 2\sqrt{\lambda_n \log n} + \frac{2}{3} \log n.$$

By the assumption that $\lambda_n \ge C \log n$ for some sufficiently large constant C, $\|\widetilde{L} - \mathbb{E}_\theta[\widetilde{L}]\| \le 4\sqrt{\lambda_n \log n}$. It follows from (B.6) and (B.7) that

$$\lambda_2(-H(\theta)) \ge \frac{1}{e^{2b}} \lambda_2(\widetilde{L}) \ge \frac{1}{e^{2b}} \left(\frac{1}{4e^{2b}} \lambda_2 - 4\sqrt{\lambda_n \log n} \right).$$

□

Proof of Theorem 7.5.1. Define $\Delta = \widehat{\theta}_{\mathrm{ML}} - \theta^*$. It follows from the definition that Δ is orthogonal to the all-one vector. By the definition of the ML estimator, $\mathcal{L}(\widehat{\theta}_{\mathrm{ML}}) \ge \mathcal{L}(\theta^*)$ and thus

$$\mathcal{L}(\widehat{\theta}_{\mathrm{ML}}) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \ge -\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \ge -\|\nabla \mathcal{L}(\theta^*)\|_2 \|\Delta\|_2, \tag{B.8}$$

where the last inequality holds due to the Cauchy-Schwartz inequality. By the Taylor expansion, there exists a $\theta = a\widehat{\theta}_{\mathrm{ML}} + (1-a)\theta^*$ for some $a \in [0,1]$ such that

$$\mathcal{L}(\widehat{\theta}_{\mathrm{ML}}) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle = \frac{1}{2} \Delta^\top H(\theta) \Delta \le -\frac{1}{2} \lambda_2(-H(\theta)) \|\Delta\|_2^2, \tag{B.9}$$

where the last inequality holds because the Hessian matrix $-H(\theta)$ is positive semi-definite with $H(\theta)\mathbf{1} = \mathbf{0}$ and $\Delta^\top \mathbf{1} = 0$. Combining (B.8) and (B.9),

$$\|\Delta\|_2 \le 2\|\nabla \mathcal{L}(\theta^*)\|_2 / \lambda_2(-H(\theta)). \tag{B.10}$$

Note that $\theta \in \Theta_b$ by definition. The theorem follows by Lemma 17 and Lemma 18.

B.4 Proof of Corollary 7.5.2

Recall that $L = \sum_{j=1}^m L_j$. Observe that $\mathbb{E}[L_j] = \frac{k_j}{n-1}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)$. Define $Z_j = L_j - \mathbb{E}[L_j]$. Then $Z_1, \ldots, Z_m$ are independent symmetric random matrices with zero mean. Note that

$$\|Z_j\| \le \|L_j\| + \|\mathbb{E}[L_j]\| \le \frac{k_j}{k_j - 1} + \frac{k_j}{n-1} \le 4.$$

Moreover,

$$\mathbb{E}[Z_j^2] = \frac{k_j^2}{(k_j-1)(n-1)}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) - \frac{k_j^2}{(n-1)^2}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right).$$

Therefore, $\|\sum_{j=1}^m \mathbb{E}[Z_j^2]\| \le \frac{2mk}{n-1}$. By the matrix Bernstein inequality [91], with probability at least $1 - n^{-1}$,

$$\|L - \mathbb{E}[L]\| \le 2\sqrt{\frac{mk\log n}{n-1}} + \frac{8}{3}\log n \le 4\sqrt{\frac{mk\log n}{n-1}} \le \frac{mk}{2(n-1)},$$

where the last two inequalities follow from the assumption that $mk \ge C\log n$ for some sufficiently large constant C. Since $\mathbb{E}[L] = \frac{mk}{n-1}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)$, the smallest eigenvalue of $\mathbb{E}[L]$ is zero and all the other eigenvalues equal $\frac{mk}{n-1}$. It follows that

$$|\lambda_i - \frac{mk}{n-1}| \le \|L - \mathbb{E}[L]\| \le \frac{mk}{2(n-1)},\ 2 \le i \le n,$$

and thus $\lambda_2 \ge \frac{mk}{2(n-1)}$ and $\lambda_n \le \frac{3mk}{2(n-1)}$. By the assumption that $mk \ge Ce^{2b}\log n$ for some sufficiently large constant C, $\lambda_2 - 16e^{2b}\sqrt{\lambda_n \log n} \ge \frac{mk}{4n}$. Then the corollary follow from Theorem 7.5.1. □

B.5 Proof of Corollary 7.6.1

Without loss of generality, assume k_j is even for all $j \in [m]$. After the random IB, there are $mk/2$ independent pairwise comparisons and let L denote the Laplacian of the comparison graph after the breaking. Recall that $L = \sum_{j=1}^m L_j$. With

random IB, we have $\mathbb{E}[L_j] = \frac{k_j}{n-1}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)$. Define $Z_j = L_j - \mathbb{E}[L_j]$. Then $Z_1, \ldots, Z_m$ are independent symmetric random matrices with zero mean. Moreover,

$$\|Z_j\| \le \|L_j\| + \|\mathbb{E}[L_j]\| \le 2 + \frac{k_j}{n-1} \le 4,$$

and

$$\mathbb{E}[Z_j^2] = \frac{2k_j}{n-1}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) - \frac{k_j^2}{(n-1)^2}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right).$$

Therefore, $\|\sum_{j=1}^m \mathbb{E}[Z_j^2]\| \le \frac{2mk}{n-1}$. Following the same argument for proving Corollary 7.5.2, we can show that $\lambda_2(L_{\mathrm{IB}}) \ge \frac{mk}{2(n-1)}$ and the corollary follows by Theorem 7.5.1 with $k = 2$.

B.6 Proof of Theorem 7.6.2

It follows from the definition of $\mathcal{L}(\theta)$ given by (7.2) that

$$\nabla_i \mathcal{L}(\theta^*) = \sum_{j: i \in S_j} \frac{1}{k_j - 1} \sum_{i' \in S_j : i' \neq i} \left[\mathbf{1}_{\{\sigma_j^{-1}(i) < \sigma_j^{-1}(i')\}} - \frac{\exp(\theta_i^*)}{\exp(\theta_i^*) + \exp(\theta_{i'}^*)}\right], \tag{B.11}$$

which is a sum of d_i independent random variables with mean zero and bounded by 1. By Hoeffding's inequality, $|\nabla_i \mathcal{L}(\theta^*)| \le \sqrt{d_i \log n}$ with probability at least $1 - 2n^{-2}$. By union bound, $\|\nabla \mathcal{L}(\theta^*)\|_2 \le \sqrt{mk \log n}$ with probability at least $1 - 2n^{-1}$. The Hessian matrix is given by

$$H(\theta) = -\sum_{j=1}^m \frac{1}{2(k_j - 1)} \sum_{i, i' \in S_j} (e_i - e_{i'})(e_i - e_{i'})^\top \frac{\exp(\theta_i + \theta_{i'})}{[\exp(\theta_i) + \exp(\theta_{i'})]^2}.$$

If $|\theta_i| \le b, \forall i \in [n]$, $\frac{\exp(\theta_i + \theta_{i'})}{[\exp(\theta_i) + \exp(\theta_{i'})]^2} \ge \frac{e^{2b}}{(1+e^{2b})^2}$. It follows that $-H(\theta) \ge \frac{e^{2b}}{(1+e^{2b})^2} L$ for $\theta \in \Theta_b$ and the theorem follows from (B.10).

REFERENCES

[1] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?" *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5163–5181, Aug. 2011.

[2] M. R. Garey, D. S. Johnson, and L. Stockmeyer, "Some simplified NP-complete graph problems," *Theoret. Comput. Sci.*, vol. 1, no. 3, pp. 237–267, 1976.

[3] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[4] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[5] G. Simons and Y. Yao, "Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons," *Ann. Statist.*, vol. 27, no. 3, pp. 1041–1060, 1999.

[6] J. C. Duchi, L. Mackey, and M. I. Jordan, "On the consistency of ranking algorithms," in *Proceedings of the ICML Conference*, Haifa, Israel, June 2010.

[7] N. Alon, M. Krivelevich, and B. Sudakov, "Finding a large hidden clique in a random graph," *Random Structures and Algorithms*, vol. 13, no. 3-4, pp. 457–466, 1998.

[8] N. Alon and N. Kahale, "A spectral technique for coloring random 3-colorable graphs," *SIAM Journal on Computing*, vol. 26, no. 6, pp. 1733–1748, 1997.

[9] E. Arias-Castro and N. Verzelen, "Community detection in random networks," arXiv:1302.7099, 2013.

[10] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Struct. Algorithms*, vol. 18, no. 2, pp. 116–140, Mar 2001.

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

[12] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004.

[13] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *Ann. Statist.*, vol. 39, no. 4, pp. 1878–1915, 2011.

[14] R. R. Nadakuditi and M. E. J. Newman, "Graph spectra and the detectability of community structure in networks," *Physical Review Letters*, vol. 108, no. 18, pp. 188–701, 2012.

[15] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Phys. Rev. E*, vol. 84:066106, 2011.

[16] E. Mossel, J. Neeman, and A. Sly, "A proof of the block model threshold conjecture," arXiv:1311.4115, 2013.

[17] S. Heimlicher, M. Lelarge, and L. Massoulié, "Community detection in the labelled stochastic block model," arXiv: 1209.2910, Nov. 2012.

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

[19] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade, "A tensor spectral approach to learning mixed membership community models," *J. Mach. Learn. Res.*, vol. 15, pp. 2239–2312, June 2014.

[20] 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*, vol. 106, no. 50, pp. 21 068–21 073, 2009.

[21] A. Amini, A. Chen, P. J. Bickel, and E. Levina, "Pseudo-likelihood methods for community detection in large sparse networks," *Ann. Statist.*, vol. 41, no. 4, pp. 2097–2122, 2013.

[22] B. Bollobás and A. Scott, "Max cut for random graphs with a planted partition," *Combinatorics, Probability and Computing*, vol. 13, no. 4-5, pp. 451–474, 2004.

[23] D. R. Luce, *Individual Choice Behavior*. New York: Wiley, 1959.

[24] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *Ann. Statist.*, vol. 32, no. 1, pp. 384–406, 02 2004.

[25] J. Guiver and E. Snelson, "Bayesian inference for Plackett-Luce ranking models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, pp. 377–384.

[26] J. A. Lozano and E. Irurozki, "Probabilistic modeling on rankings," Available at http://www.sc.ehu.es/ccwbayes/members/ekhine/tutorial_ranking/info.html, 2012.

[27] S. Negahban, S. Oh, and D. Shah, "Rank centrality: Ranking from pairwise comparisons," arXiv:1209.1688, 2012.

[28] H. A. Soufiani, W. Chen, D. C. Parkes, and L. Xia, "Generalized method-of-moments for rank aggregation," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2706–2714.

[29] H. Azari Soufiani, D. Parkes, and L. Xia, "Computing parametric ranking models via rank-breaking," in *Proceedings of the International Conference on Machine Learning*, 2014.

[30] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1955.

[31] E. Mossel, J. Neeman, and A. Sly, "Stochastic block models and reconstruction," arXiv:1202.1499, 2012.

[32] Y. Chen, S. Sanghavi, and H. Xu, "Clustering sparse graphs," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2213–2221.

[33] 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*, 2013, pp. 995–1003.

[34] T. Cai and X. Li, "Robust and computationally feasible community detection in the presence of arbitrary outlier nodes," arXiv:1404.6000, 2014.

[35] K. Chaudhuri, F. C. Graham, and A. Tsiatas, "Spectral clustering of graphs with general degrees in the extended planted partition model." *J. Mach. Learn. Res.*, vol. 23, pp. 35.1–35.23, 2012.

[36] F. McSherry, "Spectral partitioning of random graphs," in *42nd IEEE Symposium on Foundations of Computer Science*, Oct. 2001, pp. 529 – 537.

[37] U. Feige and R. Krauthgamer, "Finding and certifying a large hidden clique in a semirandom graph," *Random Struct. Algorithms*, vol. 16, no. 2, pp. 195–208, Mar. 2000.

[38] U. Feige and D. Ron, "Finding hidden cliques in linear time," in *In 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA10), Discrete Math. Theor. Comput. Sci. Proc., AM*, 2010, pp. 189–203.

[39] Y. Dekel, O. Gurel-Gurevich, and Y. Peres, "Finding hidden cliques in linear time with high probability," arxiv:1010.2997, 2010.

[40] B. Ames and S. Vavasis, "Nuclear norm minimization for the planted clique and biclique problems," *Mathematical Programming*, pp. 1–21, 2011.

[41] Y. Deshpande and A. Montanari, "Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time," arxiv:1304.7047, 2012.

[42] L. Kučera, "Expected complexity of graph partitioning problems," *Discrete Appl. Math.*, vol. 57, no. 2-3, pp. 193–212, Feb. 1995.

[43] E. Hazan and R. Krauthgamer, "How hard is it to approximate the best Nash equilibrium?" *SIAM Journal on Computing*, vol. 40, no. 1, pp. 79–91, 2011.

[44] A. Juels and M. Peinado, "Hiding cliques for cryptographic security," *Designs, Codes and Cryptography*, vol. 20, no. 3, pp. 269–280, 2000.

[45] 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*. ACM, 2007, pp. 496–505.

[46] 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 Symposium on Theory of Computing*, 2013, pp. 655–664.

[47] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," *J. Mach. Learn. Res.*, vol. 30, pp. 1046–1066 (electronic), 2013.

[48] Z. Ma and Y. Wu, "Computational barriers in minimax submatrix detection," arXiv:1309.5914, 2013.

[49] B. Applebaum, B. Barak, and A. Wigderson, "Public-key cryptography from different assumptions," in *Proceedings of the 42nd ACM Symposium on Theory of Computing*. ACM, 2010, pp. 171–180.

[50] L. Massoulié, "Community detection thresholds and the weak Ramanujan property," in *STOC 2014: 46th Annual Symposium on the Theory of Computing*, New York, United States, June 2014, pp. 1–10.

[51] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for binary symmetric block models," arXiv:1407.1591, 2014.

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

[53] 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*, 2011.

[54] 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*, 2011.

[55] V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. E1181–E1190, 2013.

[56] N. Verzelen and E. Arias-Castro, "Community detection in sparse random networks," arXiv:1308.2955, 2013.

[57] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *Ann. Statist.*, vol. 41, no. 1, pp. 1780–1815, 2013.

[58] R. Krauthgamer, B. Nadler, and D. Vilenchik, "Do semidefinite relaxations really solve sparse PCA?" arXiv:1306.3690, 2013.

[59] 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.

[60] C. Butucea and Y. I. Ingster, "Detection of a sparse submatrix of a high-dimensional noisy matrix," *Bernoulli*, vol. 19, no. 5B, pp. 2652–2688, 11 2013.

[61] Z. Ma and Y. Wu, "Computational barriers in minimax submatrix detection," arXiv:1309.5914, 2013.

[62] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007. [Online]. Available: http://dx.doi.org/10.1007/s11222-007-9033-z

[63] R. Kannan and S. Vempala, "Spectral algorithms," *Found. Trends Theor. Comput. Sci.*, vol. 4, Mar 2009.

[64] J. Jin, "Fast community detection by SCORE," *Ann. Statist.*, vol. 43, no. 1, pp. 57–89, Feb. 2015.

[65] B. Ames and S. Vavasis, "Convex optimization for the planted k-disjoint-clique problem," arXiv:1008.2814, 2011.

[66] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, New York, NY, USA, 2005, pp. 36–43.

[67] G. R. Grimmett and C. J. H. McDiarmid, "On colouring random graphs," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 77, no. 02, 1975, pp. 313–324.

[68] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu, "Clustering partially observed graphs via convex optimization," *J. Mach. Learn. Res.*, vol. 15, pp. 2213–2238, June 2014.

[69] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1, pp. 89–113, 2004.

[70] A. Jalali and N. Srebro, "Clustering using max-norm constrained optimization," arXiv:1202.5598, 2012.

[71] B. P. W. Ames, "Guaranteed clustering and biclustering via semidefinite programming," *Mathematical Programming*, pp. 1–37, 2013.

[72] S. Chattergee, "Matrix estimation by universal singular value thresholding," arXiv: 1212.1247, 2012.

[73] 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 (ICASSP)*, 2014.

[74] 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 Symposium on Theory of Computing*, 2013, pp. 655–664.

[75] A. Coja-Oghlan, "Graph partitioning via adaptive spectral techniques," *Comb. Probab. Comput.*, vol. 19, no. 2, pp. 227–284.

[76] A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige, and A. Vijayaraghavan, "Detecting high log-densities: An $o(n^{1/4})$ approximation for densest k-subgraph," in *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, ser. STOC '10, 2010, pp. 201–210.

[77] M. Jerrum, "Large cliques elude the metropolis process," *Random Structures & Algorithms*, vol. 3, no. 4, pp. 347–359, 1992. [Online]. Available: http://dx.doi.org/10.1002/rsa.3240030402

[78] B. Applebaum, B. Barak, and A. Wigderson, "Public-key cryptography from different assumptions," in *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, ser. STOC '10, 2010, http://www.cs.princeton.edu/~boaz/Papers/ncpkcFull1.pdf. pp. 171–180.

[79] B. Hajek, Y. Wu, and J. Xu, "Computational lower bounds for community detection on random graphs," arXiv:1406.6625, 2014.

[80] N. Alon, S. Arora, R. Manokaran, D. Moshkovitz, and O. Weinstein., "Inapproximabilty of densest κ-subgraph from average case hardness," 2011, available at https://www.nada.kth.se/ rajsekar/papers/dks.pdf.

[81] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. ACM*, vol. 42, no. 6, pp. 1115–1145, Nov. 1995.

[82] T. Qin, X. Geng, and T. yan Liu, "A new probabilistic model for rank aggregation," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 1948–1956.

[83] S. Jagabathula and D. Shah, "Inferring rankings under constrained sensing." in *NIPS*, vol. 2008, 2008.

[84] M. Braverman and E. Mossel, "Sorting from noisy information," arXiv:0910.1191, 2009.

[85] A. Rajkumar and S. Agarwal, "A statistical convergence perspective of algorithms for rank aggregation from pairwise data," in *Proceedings of the International Conference on Machine Learning*, 2014.

[86] A. S. Hossein, D. C. Parkes, and L. Xia, "Random utility theory for social choice," in *Proceeedings of the 25th Annual Conference on Neural Information Processing Systems*, 2012.

[87] H. A. Soufiani, D. C. Parkes, and L. Xia, "Preference elicitation for general random utility models," arXiv:1309.6864, 2013.

[88] R. D. Gill and B. Y. Levit, "Applications of the van Trees inequality: a Bayesian Cramér-Rao bound," *Bernoulli*, vol. 1, no. 1-2, pp. 59–79, 03 1995.

[89] D. Spielman and S. Teng, "Spectral sparsification of graphs," *SIAM Journal on Computing*, vol. 40, no. 4, pp. 981–1025, 2011.

[90] R. Durrett, *Random Graph Dynamics*. New York, NY: Cambridge University Press, 2007.

[91] J. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[92] L. Massoulié and D. Tomozei, "Distributed user profiling via spectral methods," *Stochastic Systems*, vol. 4, pp. 1–43, 2014.

[93] V. Vu, "A simple SVD algorithm for finding hidden partitions," arXiv:1404.3918, 2014.

[94] D. Dubhashi and D. Ranjan, "Balls and bins: A study in negative dependence," *Random Structures and Algorithms*, vol. 13, no. 2, pp. 99–124, 1998.

[95] R. Vershynin, "A simple decoupling inequality in probability theory," 2011, preprint, available at http://arxiv.org/abs/1111.4452.

[96] P. Erdös and A. Rényi, "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[97] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. New York, NY, USA: Cambridge University Press, 2005.

[98] U. Feige and E. Ofek, "Spectral techniques applied to sparse random graphs," *Random Struct. Algorithms*, vol. 27, no. 2, pp. 251–275, Sept. 2005.

[99] L. Massoulié and D. Tomozei, "Distributed user profiling via spectral methods," *Stochastic Systems*, vol. 4, pp. 1–43, 2014.

[100] Y. Seginer, "The expected norm of random matrices," *Combinatorics, Probability and Computing*, vol. 9, no. 2, pp. 149–166, 2000.

[101] T. Tao, *Topics in random matrix theory*. Providence, RI, USA: American Mathematical Society, 2012.

[102] R. B. Ash, *Information Theory*. New York, NY: Dover Publications Inc., 1965.

[103] T. P. Hayes, "A large-deviation inequality for vector-valued martingales," Available at http://www.cs.unm.edu/~hayes/papers/VectorAzuma/VectorAzuma20050726.pdf, 2005.