Exact recovery in the stochastic co-block model

Ziliang Samuel Zhong

Department of Mathematics

New York University Shanghai

Supervisor

Prof. Shuyang Ling

In partial fulfillment of the requirements for the degree of

Bachelor of Science in Mathematics

May 25, 2021

Abstract

After the best-known Gormans-Williamson relaxation, more and more SDP relaxations of NP-hard combinatorial optimization problems has been proposed by the research community. We have proposed an SDP based directed community detection algorithm and analyze its performance in the stochastic coblock models (directed stochastic block models). In this work, we will answer the fundamental question: under what condition, the SDP algorithm can exactly recover all the hidden communities? In 2016, Bandeira proposed a tight spectral approximation of Laplacian matrices, which provide us with powerful tools to analyze the matrices involved in the SDP.We prove that the SDP algorithm is able to achieve exact recovery of the planted community structure under conditions that match the information-theoretic limits.

Contents

Chapter 1

Introduction

Research is formalized
curiosity.

Zora Neale Hurston, 1942

1.1 Background

Community detection and clustering are central problems in machine learning and data science. Networks, as a ubiquitous structure in the nature, appear in diverse domains, including sociology, biology, neuroscience and computer science [GN02], [New03]. Among many properties of networks, one of the most significant features is community structure or clustering i.e., a subset of vertices in a huge network are strongly connected while the inter-community connectivity is relatively weak. There are two kinds of networks(graphs): directed (such as citation networks, food web and airline networks) and undirected (such as maps, social networks and computer networks). A huge amount of research has been done to solve the challenging community detection problem in undirected networks: when and how to infer the hidden community structure from the linkage among vertices in undirected networks [For10], [GN02], [LFR08], [New03]. It is worth mentioning that most real world interactions are directional so finding clusters in directed networks is also a challenging task with several important applications [MV13]. However, the problem of graph clustering has mainly been considered and studied for the case of undirected networks. Given its importance but lack of attention from the research community, studying the clustering algorithms and their performance in directed graphs becomes our original motivation of the thesis.

Network with random block structure is common in various domains including mathematics, computer science, physics, and statistics. Originally proposed in [Kno08] to study the social networks, the stochastic block model (SBM) is a classical example. Now it has become a benchmark model for comparing different community detection algorithms [LF09]. Its growing popularity is largely due to the fact that its structure is simple to describe, but at the same time it has interesting and involved phase transition properties which have only recently been discovered [ABH14], [Ban15], [AS15]. Recently, lots of research has been done to develop algorithms and methods to either recover or detect the hidden communities with emphasis on understanding the fundamental limits for community detection in connection with the undirected SBM [ABH14]. However, there is limited discussion about the recovery and detection in the directed SBM.

SBM (directed and undirected) is known to have desirable consistency and in some sense optimality properties, but the exactly estimating the parameters of SBM is in general NP-hard [KBG18]. Thus, finding an effective semidefinite programming (SDP) relaxation of the likelihood optimization problem becomes a way to overcome the computational challenge.

In this work, we will study the performance of semidefinite programming in community detection for the stochastic co-block model. We denote by $\mathcal{G}(n, p_n, q_n)$ the stochastic co-block model (ScBM) or the directed SBM with a total of n vertices and $\frac{n}{2}$ for each community; the adjacency matrix $A = (A_{ij})_{1\leq i,j\leq n}$ of this directed network is an asymmetric matrix which has zero diagonal and its (i, j)-entry an independent Bernoulli random variable:

$$\mathbb{P}(a_{ij} = 1) = \begin{cases} p_n & \text{if directed pairs } (i, j) \text{ are in the same community} \\ q_n & \text{if directed pairs } (i, j) \text{ are in different communities} \end{cases}$$

where $p_n > q_n$ $\forall n$. Note that the parameters p and q usually depend on n; for simplicity, we replace p_n and q_n by p and q if there is no confusion.

We will answer the following fundamental question: under what condition on

(n, p, q) is the semidefinite programming able to recover the underlying hidden communities exactly from directed networks generated from ScBM $\mathcal{G}(n, p, q)$?

1.2 Related works and our contribution

The four topics: community detection, semidefinite programming, stochastic block model and directed networks been studied over the years with various different objectives and guarantees, so naturally there is a wealth of research production in these areas. We will briefly review the topics and highlight the literatures that inspired our research.

Community detection for general networks is well studied and has found many applications. There has been significant recent literature concentration around the bipartiton (bisection) and the general k-partition problems (multisection) in random and semi-random models [Dec+11a], [ABH14], [AS15], [Ban15]. For undirected networks, spectral clustering [NJW01], [Lux07] based on Laplacian matrix [Chu97] plays an important role in undirected community detection tasks. The classical spectral clustering follows the well-known two step procedure: Laplacian eigenmap and clustering the eigenspace [NJW01], [Lux07]. Although spectral clustering is shown to be effective in many domains, its theoretical understanding is still relatively unclear. For directed networks, current clustering algorithms are mostly based on information-theoretic approaches and probabilistic methods. Generally, the existence of communities in networks represent structural patterns, which can be used to effectively compress the network (data). Rosvall and Bergstrom proposed a method (called Isomap) in [RB08] to identify communities in directed networks, by combining random walks and compression principles. Newman and Leicht [LN08] proposed an approach for community detection in directed networks based on mixture models for statistical inference. To highlight the difficulty of community detection in directed network, Fortunato stated that *"Developing methods of community detection for directed graphs is a hard task. For instance, a directed graph is characterized by asymmetrical matrices (adjacency matrix, Laplacian, etc.), so spectral analysis is much more complex. Only a few methods can be easily extended from the undirected to the directed case. Otherwise,* *the problem must be formulated from scratch"* [LFR08]. Therefore, finding effective ways to deal with asymmetry plays an important role in directed community detection.

The analysis of the stochastic block model originated from [Kno08] to study the interaction of social networks. Since then, a vast amount of follow-up research has been conducted to understand how to recover the hidden planted partition with efficient polynomial-time algorithms. There are two natural problems that arise in context of the SBM: exact recovery [ABH14], [AS15], [Ban15], where the aim is to recover the hidden partition completely; and detection [MNS18], where the aim is to recover the partition better than what a random guess would achieve. In this work, we will focus on exact recovery in directed SBM. For undirected SBM, [Abb17] has done an exhaustive literature review about the sharp threshold of detection and proved that for exact recovery in SBM with 2 communities: (1) in the regime $p = an^{-1}, q = bn^{-1}$, [Dec+11b] applied the cavity method to predict that a detection threshold exists for the community detection problem under stochastic block models. Later on, this detection threshold is confirmed by [Mas14] [MNS18] that the detection of community is possible if and only if $(a-b)^2 > 2(a+b)$; (2) in the regime $p = a\log(n)n^{-1}, q = b\log(n)n^{-1}$, [ABH14] proved that the sharp threshold for exact recovery is $\sqrt{a} - \sqrt{b} > \sqrt{2}$ by applying spectral clustering and SDP respectively to the SBM. For directed SBM, [RQY15] represented stochastic coblock model (directed SBM) based on the notion of co-clustering (i.e., the task in which both rows and columns of the adjacency matrix are clustered simultaneously). The new model is also accompanied by a new spectral clustering algorithm for directed networks based on the singular value decomposition of the Laplacian defined by the authors. They extend the spectral clustering presented in [Lux07] to an asymmetric context.

The use of SDP in combinatorial optimization originated from [Lov79] for the *Lovasz theta function*, which is the upper bound of the Shannon capacity of a graph. In 1995, Goemans and Williamson proposed the first graph approximation algorithm based on SDP [GW95]. The well-known Goemans-Williams relaxation gives the approximation ratio to the NP-hard max-cut problem. After that, many SDP relaxation of NP-hard combinatorial problems had been proposed by the research community. In fact, the algorithm we will analyze in ScBM is greatly inspired by [GW95] and the algorithm used by [Aga+15] in undirected SBM. In this work, we are interested in when is it the case the SDP algorithm can exactly recover all the hidden communities for inputs generated from ScBM. As a result of the analysis, we can deduce the sharp threshold for exact recovery in stochastic cblock models. Our contribution is that: (1) we proved the sharp threshold for exact recovery in ScBM; (2) we proposed an SDP algorithm that can achieve the threshold; (3) we proposed a directed spectral clustering algorithm that can achieve the threshold (we will not cover the details in this work).

1.3 Notations

For a matrix M, we denote the kth smallest eigenvalue by $\lambda_k(M)$, the largest eigenvalue by $\lambda_{\max}(M)$, and its operator norm by $\|M\|_{op}$.

$\mathbf{1}$ denotes the all-ones vector, whenever there is no risk of ambiguity for its dimension and J_n denotes the $n \times n$ all-one matrix.

For a scalar random variable Y, we will write its p-norm as $\|Y\|_p = \mathbb{E}\|Y\|^p$ and infinity norm as $\|Y\|_\infty = \{a : Y \leq a \textbf{ a.s.}\}$.

Given a directed graph $\mathcal{G}$ and its adjacency matrix A, $\deg_R(i)$ denotes the ith row sum of its adjacency matrix A, and $\deg_C(i)$ denotes the ith column sum of its adjacency matrix A. In the stochastic coblock model, in the context of co-clustering, $\deg_R^+(i)$ and $\deg_R^-(i)$ denote the in-cluster and out-cluster degree in row of node i; and $\deg_C^+(i)$ and $\deg_C^-(i)$ denote the in-cluster and out-cluster degree in column of node i.

We say an event $\mathcal{E}$ happens with high probability if

$$\mathbb{P}[\mathcal{E}] = 1 - n^{-\Omega(1)}$$

where n is an underlying parameter that is thought of going to infinity such as dimension of the matrix and the number of nodes in the graph.

Chapter 2

Premilinaries

2.1 Positive Semidefinite Matrices

In this section we will review the definition of positive semidefinite matrices and some of their properties. We refer interested readers to [HJ85] and [**GLS90**] for proof of the theorems and details about this topic.

Definition 2.1.1. *An $n \times n$ matrix A is is said to be positive semidefinite (PSD) if A is symmetric and if $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^n$. We will write $A \succcurlyeq 0$ if A is PSD.*

Theorem 2.1.1. *(Spectral theory for PSD matrices) Let A be an $n \times n$ PSD matrix, then the following hold:*

(1) The eigenvalues $\lambda_1 \leq \lambda_2 \leq, .., \leq \lambda_n$ are non-negative.

(2) A has orthonormal eigenvectors $v - 1, ..., v_n$ corresponding to λ_i for $i = 1, .., n$. Moreover, we have the decomposition $A = V\Lambda V^\top$, where $V = (v_1, ..., v_n)$ and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$.

(3) The rank of A is equal to the number of nonzero eigenvalues of A.

Definition 2.1.2. *Given symmetric matrices A, B, we define $\langle A, B\rangle := A \bullet B = \mathrm{tr}(A^\top B) = \sum_{i,j} A_{ij}B_{ij}$.*

Fact 1. $x \top Ax = \sum_{ij} x_i x_j A_{ij} = (xx^\top) \bullet A$.

Fact 2. *For any two $n \times n$ matrices A, B, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.*

Lemma 2.1.2. *For $A \succcurlyeq 0$, $A \bullet B \succcurlyeq 0$, $\forall B \succcurlyeq 0$.*

Lemma 2.1.3. *For PSD matrices A, B, $A \bullet B = 0$ iff $AB = 0$.*

2.2 Semidefinite Programming

Given this understanding of PSD matrices, we can now look at semidenite programs (SDPs), and the duality theory. For additional background information we refer readers to Chapter 4 and 5 of [BV04] and [HRW00].

2.2.1 Basic Definitions

In semidefinite programming, we are interested in optimizing a linear function of a symmetric matrix subject to linear constraints and a crucial additional constraint that the matrix be positive semidefinite. It could be viewed as an extension of linear programming and a particular case of conic programmings (restricted to the cone of positive semidefinite matrices). Let $\mathcal{S}_n$ denote the set of $n \times n$ symmetric matrices and $\mathcal{S}_n^+$ denote the set of $n \times n$ PSD matrices. The SDP of the standard form is defined to be: for $C, A_1, ..., A_m \in \mathcal{S}_n$, and $b \in \mathbb{R}^m$ be given,

$$(PSDP): \quad \begin{aligned} \min \quad & C \bullet X \\ \textbf{s.t} \quad & A_i \bullet X = b_i \quad 1 \leq i \leq m \\ & X \in \mathcal{S}_n^+ \end{aligned} \tag{2.1}$$

By Lagrangian dual, its dual form could be written as:

$$(DSDP): \quad \begin{aligned} \max \quad & b^\top y \\ \textbf{s.t} \quad & \sum_{i=1}^{m} y_i A_i + Z = C \\ & y \in \mathbb{R}^m, \quad Z \in \mathcal{S}_n^+ \end{aligned} \tag{2.2}$$

We shall refer to PSDP as the primal problem and to DSDP as the dual problem. In order to decide about possible infeasibility and unboundedness of the problems (PSDP) and (DSDP), let us consider the following definitions.

Definition 2.2.1. *A matrix $X \in \mathcal{S}_n^+$ is called a primal ray if $A_1 \bullet X = 0$ for $i = 1, ..., m$ and $C \bullet X < 0$.*

Definition 2.2.2. *A vector $y \in \mathbb{R}^m$ is called dual ray if $-\sum_{i=1}^m y_i A_i \succcurlyeq 0$ and $b^\top y > 0$.*

We then have the following elementary result:

Proposition 2.2.1. *The existence of a dual ray implies the infeasibility of (PSDP). Similarly, the existence of a primal ray implies the infeasibility of (DSDP).*

2.2.2 Duality Theory

By simply taking the difference between the objective function of PSDP and DSDP, we get the weak duality theorem:

Theorem 2.2.1. *(**SDP Weak Duality**) Let X and (y, Z) be feasible for PSDP and DSDP, respectively. Then we have $C \bullet X - b^\top y = X \bullet Z \geq 0$.*

As the generalization of Linear Programming, it is natural to try generalizing the Farkas lemma to the positive semidefinite cone $\mathcal{S}_n^+$. It turns out that such generalization is possible, but the strong duality theorem for SDP is slightly weaker than that for LP because some additional qualifications should be satisfied. Let's introduce the strong duality theory and KKT conditions.

Theorem 2.2.2. *(**SDP Strong Duality**) Consider the following primal-dual pair:*

$$
\begin{aligned}
(PSDP):& \quad v_p^* = \inf\{C \bullet X : A_i \bullet X = b_i, i = 1, ..., n, X \in \mathcal{S}_n^+\} \\
(DSDP):& \quad v_d^* = \inf\{b^\top y : \sum_{i=1}^m y_i A_i + Z = C, y \in \mathbb{R}^m, Z \in \mathcal{S}_n^+\}
\end{aligned}
$$

Then the following hold:

(1) If the DSDP is strictly feasible, (i.e there exists an $(\tilde{y}, \tilde{Z}) \in \mathbb{R}^n \times \mathcal{S}_n^+$, such that $\sum_{i=1}^m \tilde{y}_i A_i + \tilde{Z} = C, \tilde{y} \in \mathbb{R}^m$ and $\tilde{Z} \succcurlyeq 0$), then $v_p^* = v_d^*$. If in addition (DSDP) is bounded above, then the common optimal value is attained by some $X^* \in \{X \in \mathcal{S}_n^+ : A_i \bullet X = b_i, i = 1, ..., n\}$.

(2) If the PSDP is strictly feasible, (i.e there exists an $\tilde{X} \succcurlyeq 0$ such that $A_i \bullet X = b_i, i = 1, ..., n$), then $v_p^* = v_d^*$. If in addition (PSDP) is bounded above, then the common optimal value is attained by some $(y^*, Z^*) \in \{(y, Z) : \sum_{i=1}^m y_i A_i + Z = C\}$.

(3) Suppose that at least one of (PSDP) or (DSDP) is bounded and strictly feasible. Then, a primaldual feasible pair $(X; y, Z)$ is a pair of optimal solutions to the respective problems if and only if either one of the following holds:

*(a) (**Zero Duality Gap**) $CX = b^\top y$*

*(b) (**Complementary Slackness I**) $X \bullet Z = 0 \iff XZ = 0$.*

*(c) (**Complementary Slackness II**) There exists an $n \times n$ orthogonal matrix V ($V^\top V = I$) such that (i) $X = V\Lambda V^\top$, $Z = V\Omega V^\top$ for some $n \times n$ diagonal matrices Λ, Ω. (ii) $\Lambda\Omega = 0$. In, particular, we have $\mathrm{rank}(X) + \mathrm{rank}(Z) \leq n$.*

*(4) Suppose that both (PSDP) and (DSDP) are strictly feasible $(\star)$. Then we have $v_p^* = v_d^*$ and both values are attained. The the pair $(X; y, Z)$ is primal-dual optimal if and only if the **KKT conditions**: 1. $(*)$; 2. $3(a)$ or $3(b)$ hold.*

2.3 Perturbation theory

Suppose A is the adjacency matrix sampled from two-community symmetric stochastic coblock model $\mathcal{G}(n, p, q)$. Without loss of generality, we assume the first $n/2$ nodes form one community and the second half nodes form the other one. Since A is asymmetric, we usually consider the augmented matrix of A:

Definition 2.3.1. *(**Augmented Matrix**) The augmented matrix of an $n \times n$ matrix A is $A^* := \begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix}$, which is an $2n \times 2n$ symmetric matrix.*

If A has the singular value decomposition $A = U\Sigma V^\top$, then A^* has the eigen-decomposition:

$$\begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} V & V \\ U & -U \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} V^\top & U^\top \\ U^\top & -U^\top \end{pmatrix}$$

where $U = (u_1, ..., u_n)$ and $V = (v_1, ..., v_n)$ are both orthogonal matrices. $\Sigma = \text{diag}(\sigma_1, ..., \sigma_n)$. σ_i are called singular values, $\sigma_1 \leq ... \leq \sigma_n$; u_i are called left singular vectors; v_i are called right singular vectors.

Let $\overline{A} = \mathbb{E}A$, we have

$$\mathbb{E}A = \begin{pmatrix} pJ_{n/2} & qJ_{n/2} \\ qJ_{n/2} & pJ_{n/2} \end{pmatrix}$$

$p > q$, with two distinct eigen pairs: $(\frac{n(p+q)}{2}, \frac{1}{\sqrt{n}}(\mathbf{1}, \mathbf{1}))$ and $(\frac{n(p-q)}{2}, \frac{1}{\sqrt{n}}(\mathbf{1}, -\mathbf{1}))$. In the main results, we will use tools from pertubation theory, that is, seeing a matrix M as pertubed $\mathbb{E}M$: $M = \mathbb{E}M + (M - \mathbb{E}M)$. From this, we can establish the relationship between eigenvalues of M and eigenvalues of $\mathbb{E}M$. For eigenvalue perturbation, we will introduce two useful tools: the well-known min-max principle, which gives rise to the famous Weyls inequality.

Theorem 2.3.1. ***(Courant-Fischer-Weyl min-max/max-min principles)*** *Let A be an $n \times n$ Hermitian matrix with eigenvalues $\mu_1 \leq ... \leq \mu_n$. For any $d = 1, ..., n$, write ν_d for the d-dimensional subspace of $\mathbb{C}^n$. Then*

$$\lambda_t = \min_{V \in \nu_t} \max_{x \in V \setminus \{0\}} \frac{\langle x, Ax \rangle}{\langle x, x \rangle} == \min_{V \in \nu_{n-t+1}} \max_{x \in V \setminus \{0\}} \frac{\langle x, Ax \rangle}{\langle x, x \rangle}$$

Theorem 2.3.2. ***(Weyl)*** *Let A be an $n \times n$ Hermitian matrix with eigenvalues $\lambda_1 \leq ... \leq \lambda_n$. Let B be an $n \times n$ Hermitian matrix with eigenvalues $\mu_1 \leq ... \leq \mu_n$. Suppose the eigenvalues of $A + B$ are $\rho_1 \leq ... \leq \rho_n$, then for $i = 1, ..., n$,*

$$\lambda_i + \mu_1 \leq \rho_i \leq \lambda_i + \mu_n$$

In our main results, we will also use the conclusions from [Ban15] about the eigenvalue approximation of random Laplacian matrix.

Definition 2.3.2. ***(Laplacian)*** *Given a symmetric $n \times n$ matrix X, we define* *the Laplacian L_X of X as*

$$L_X := \mathcal{D} - X,$$

where D_X is the diagonal matrix whose diagonal entries are given by

$$(\mathcal{D})_{ii} = \sum_{j=1}^{n} X_{ij}.$$

We will refer to any symmetric matrix satisfying the condition $L\mathbf{1} = 0$ as a Laplacian matrix.

Theorem 2.3.3. *(**spectral approximation of random Laplacian**) (Theorem 2.1 in [Ban15]) Let L be an $n \times n$ symmetric random Laplacian ($L\mathbf{1} = 0$) with centered independent off-diagonal entries such that $\sum_{j\in[n]\setminus i} \mathbb{E}[L_{ij}^2]$ is equal for every i.*
Define σ and σ_∞ as

$$\sigma^2 = \sum_{j\in[n]\setminus i} \mathbb{E}[L_{ij}^2] \quad \sigma_\infty = \max_{j\neq i} \|L_{ij}\|_\infty^2.$$

If there exist $c > 0$ such that

$$\sigma \geq c\sqrt{\log(n)}\sigma_\infty,$$

then there exists c_1, C, β_1, all positive and depending only on c, such that

$$\lambda_{\max}(L) \leq (1 + \frac{C_1}{\sqrt{\log(n)}}) \max_i L_{ii}.$$

2.4 SDP relaxation of community detection in ScBM

We consider the directed stochastic block model with two communities $\mathcal{G}(n, p, q)$, $p > q$. We adapt the concept of co-clustering in [HJ85], that is, cluster the asymmetric adjacency matrix in rows and columns respectively. Intuitively, clustering the network means maximizing the degree discrepancy (i.e the difference between

$\deg^+(i)$ and $\deg^-(i)$) in rows and in columns respectively. We can write the following programming:

$$\begin{aligned} \max \quad & \mathrm{Tr}(u^\top A v) \\ \textbf{s.t} \quad & u = \{1, -1\}^n, \\ & v = \{1, -1\}^n \end{aligned} \tag{2.3}$$

In general, to find the row membership vector u and column membership vector v from this programming is NP-hard. The main obstacles are: (1) A is asymmetric, limited linear algebra tools could be used; (2) the problem is nonconvex; (3) we have prior knowledge about the model but no constraints about the model is in (2.3). To solve (3), we just have to penalize $u^\top \mathbf{1}$ and $v^\top \mathbf{1}$ in the objective function:

$$\begin{aligned} \max \quad & \mathrm{Tr}(u^\top A v) - \lambda(|u^\top \mathbf{1}| + |v^\top \mathbf{1}|) \\ \textbf{s.t} \quad & u = \{1, -1\}^n, \\ & v = \{1, -1\}^n \\ & \lambda > 0 \end{aligned} \tag{2.4}$$

Then, inspired by the well-known Geomans-Williams relaxation, we come up with the semidefinite programming to solve (1) and (2):

$$\begin{aligned} \max \quad & \langle A^*, X \rangle - \lambda \langle J_{2n}, X \rangle \\ \textbf{s.t} \quad & X_{ii} = 1 \\ & X \succcurlyeq 0 \\ & \lambda > 0 \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} X &= xx^\top \\ x &= (u, v)^\top \\ u &= \{1, -1\}^n \\ v &= \{1, -1\}^n \end{aligned}$$

(2.5) is a convex relaxation of (2.4). To recover the communities in the graph, we will maximize the difference between in-comminity degree and cross-community degree in rows and columns respectively and do not want u and v to be too close to the all-one vector or the all-negative-one vector. Choose $\lambda = \frac{1}{2}$, (2.5) becomes:

$$\begin{aligned} \max \quad & \mathrm{Tr}((2A^* - J_{2n})X) \\ \textbf{s.t} \quad & X_{ii} = 1 \\ & X \succcurlyeq 0 \end{aligned} \tag{2.6}$$

Note that

$$2A^* - J_{2n} = \begin{pmatrix} -J_n & B^\top \\ B & -J_n \end{pmatrix}$$

$$B_{ij} = \begin{cases} 1 & a_{ij} = 1 \\ -1 & a_{ij} = 0 \end{cases}$$

Then we will define the degree matrices for ScBM and the Laplacian matrix for ScBM:

$$(D_R^+)_{ii} := \begin{cases} \sum_{j=1}^{n/2} A_{ij} & i \in [1, n/2] \\ \sum_{j=n/2+1}^{n} A_{ij} & i \in [n/2+1, n] \end{cases}$$

$$(D_R^-)_{ii} := \begin{cases} \sum_{j=n/2+1}^{n} A_{ij} & i \in [1, n/2] \\ \sum_{j=1}^{n/2} A_{ij} & i \in [n/2+1, n] \end{cases}$$

$$(D_C^+)_{ii} := \begin{cases} \sum_{j=1}^{n/2} A^\top_{ij} & i \in [1, n/2] \\ \sum_{j=n/2+1}^{n} A^\top_{ij} & i \in [n/2+1, n] \end{cases}$$

$$(D_C^-)_{ii} := \begin{cases} \sum_{j=n/2+1}^{n} A^\top_{ij} & i \in [1, n/2] \\ \sum_{j=1}^{n/2} A^\top_{ij} & i \in [n/2+1, n] \end{cases}$$

Definition 2.4.1. *(**degree matrices**) We define the in-degree matrix and out-degree matrix for stochastic coblock model as:*

$$\mathcal{D}^+ := \begin{pmatrix} D_C^+ & 0 \\ 0 & D_R^+ \end{pmatrix} \quad \mathcal{D}^- := \begin{pmatrix} D_C^- & 0 \\ 0 & D_R^- \end{pmatrix}.$$

Definition 2.4.2. *Given a directed graph drawn from the stochastic coblock model with two communities, we define*

$$\Gamma_{SBM} := \mathcal{D}^{+} - \mathcal{D}^{-} - A^{*},$$

where A^* is the augmented adjacency matrix.

The dual SDP of (2.6) is:

$$\begin{aligned} \min \quad & \operatorname{Tr}(Y) \\ \textbf{s.t} \quad & Y \succcurlyeq 2A^{*} - J_{2n} \\ & Y \text{diagonal} \end{aligned} \tag{2.7}$$

2.5 Threshold of exact recovery

After having the SDP relaxation of the community detection problem, we are naturally curious about its performance in ScBM. The SDP algorithm maximizes the degree discrepancy between two communities. We will thus investigate when this estimator succeeds/fails in exactly recovering the planted partition. Recall that we will work in the regime:

$$p = \frac{a\log(n)}{n}, q = \frac{b\log(n)}{n}.$$

In Chapter 3, we will first illustrate when SDP fails to recover the planted communities, then prove that when $\sqrt{a} - \sqrt{b} > \sqrt{2}$, SDP solves exact recovery, finally we will discuss our current work about directed spectral clustering in ScBM.

Chapter 3

Main results

The main goal of this work is to show that the SDP algorithm we proposed in Section 2.4 exactly recovers hidden communities for the directed SBM $\mathcal{G}(n, p, q)$ where $p = \frac{a\log(n)}{n}$ and $q = \frac{b\log(n)}{n}$, $a > b$ when $\sqrt{a} - \sqrt{b} > \sqrt{2}$ and by combinatorial analysis, we will show that this threshold is sharp.

3.1 Converse

"Converse" refers to the impossibility part of a result, i.e., when exact recovery cannot be solved in this case. We proposed a condition under which exact recovery is unsolvable in the ScBM. We will use combinatorial analysis to prove the result.

Theorem 3.1.1. *Let G be a sample from $\mathcal{G}(n, p, q)$ where $p = \frac{a\log(n)}{n}$ and $q = \frac{b\log(n)}{n}$,$a > b$. if $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then exact recovery is not solvable.*

3.2 Achieving the threshold

3.2.1 SDP algorithm

We would like to study the performance of our SDP algorithm in ScBM and under what condition it can solve exact recovery.

Lemma 3.2.1. *Let* $\Lambda = 2\Gamma_{SBM} + I_{2n} + \begin{pmatrix} 2J_n & J_n - I_n \\ J_n - I_n & 2J_n \end{pmatrix}$*. If*

$$\begin{aligned} \Lambda &\succcurlyeq 0 \\ \textit{and } \lambda_2(\Lambda) &> 0 \end{aligned} \tag{3.1}$$

then $gg^\top$ *is the unique solution to the SDP* (2.1)*.*

Lemma 3.2.2. *Let* $n > 4$ *be even and let* G *be drawn from* $\mathcal{G}(n, p, q)$*,* $p > q$*. As long as*

$$\lambda_{\max}(-\Gamma_{SBM} + \mathbb{E}[\Gamma_{SBM}]) < n(p - q),$$

the semidefinite program for stochastic coblock model achieves exact recovery, meaning that $gg^\top$ *is its unique solution.*

We will use Theorem 2.1 in [Ban15] to estimate this largest eigenvalue. Then, we obtain the following theorem.

Theorem 3.2.3. *Let* $n \geq 4$ *be even and* G *drawn from* $\mathcal{G}(n, p, q)$*,* $p > q$*. As long as* $\frac{\log(n)}{\epsilon n} < q < p < \frac{1}{2}$*, for some constant* $\epsilon > 1$ *,then* $\exists \Delta > 0$ *such that, with high probability, the following holds: If,*

$$\min_{i \in [2n]} ((\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii}) \geq \frac{\Delta}{\sqrt{\log(n)}} \mathbb{E}[\deg_C^+(i) - \deg_C^-(i))] \tag{3.2}$$

then the semidefinite program (2.1) *achieves exact recovery.*

As a consequence of Theorem 3.2.3, we can prove that SDP achieves the sharp threshold for exact recovery in stochastic coblock model.

Lemma 3.2.4. *Let* G *be a directed random graph with* n *nodes drawn accordingly to the stochastic co-block model on two communities with edge probabilities* $p = \frac{a\log}{n}$*,* $q = \frac{b\log}{n}$*,* $a > b$ *are constants. Then for any constant* $\Delta > 0$*,*
(1) If

$$\sqrt{a} - \sqrt{b} > \sqrt{2}, \tag{3.3}$$

with high probability,

$$\min_{i\in[2n]}((\mathcal{D}^{+})_{ii}-(\mathcal{D}^{-})_{ii}) \geq \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg_C^{+}(i)-\deg_C^{-}(i))].$$

(2) On the other hand, if

$$\sqrt{a}-\sqrt{b}<\sqrt{2},$$

with high probability,

$$\min_{i\in[2n]}((\mathcal{D}^{+})_{ii}-(\mathcal{D}^{-})_{ii})<0$$

and exact recovery is impossible.

Together with theorem 3.2.3, Lemma 3.2.4 implies the following corollary.

Corollary 3.2.4.1. *Let G be a directed random graoh with n nodes drawn accordingly to the stochastic co-block model on two communities with edge probabilities $p=\frac{a\log}{n}$, $q=\frac{b\log}{n}$, $a>b$ are constants. As long as*

$$\sqrt{a}-\sqrt{b}>\sqrt{2},$$

the semidefinite program concides with the true partition with high probability.

3.2.2 Spectral algorithm

If we view the community detection task in ScBM as an unsupervised learning problem, we would like to find the maximum likelihood estimator of the parameters a and b. Recall the NP-hard programming:

$$\begin{aligned} \max \quad & \mathrm{Tr}(u^{\top}Av) \\ \textbf{s.t} \quad & u=\{1,-1\}^{n}, \\ & v=\{1,-1\}^{n} \end{aligned} \tag{3.4}$$

The SDP relaxation aims to relax the integral constraint to a convex set that contains all u and v, while the spectral relaxation is to relax the integral constraint to an Euclidean constraint on real valued vectors. This leads to looking for a maximizer of

$$\begin{aligned} \max \quad & x^\top A^* x \\ \textbf{s.t} \quad & \|x\|_2^2 = 2n \\ & x^\top \mathbf{1} = 0 \end{aligned} \tag{3.5}$$

Similar to the SDP relaxation, this program also maximizes the degree discrepancy in rows and columns respectively. Since $\mathbf{1}$ is close to an eigenvector of A^*, the constraint $x^\top \mathbf{1} = 0$ leads the maximization (3.5) to focus on the eigenspace orthogonal to the rst eigenvector, and thus to the eigenvector corresponding to the second largest eigenvalue (in absolute value). Thus it is reasonable to take the second largest eigenvector u_2 of A^* and round it to obtain an efficient relaxation of the MLE:

$$\hat{X}(i)_C = \begin{cases} \text{cluster 1 in columns} & \text{if } u_2(i) \geq 0 \\ \text{cluster 2 in columns} & \text{if } u_2(i) < 0 \end{cases}$$

$$\hat{X}(i)_R = \begin{cases} \text{cluster 1 in rows} & \text{if } u_2(i+n) \geq 0 \\ \text{cluster 1 in rows} & \text{if } u_2(i+n) < 0 \end{cases}$$

This algorithm is equivalent to finding the eigenvector corresponding to the largest eigenvalue of $A^* - \frac{(a-b)\log(n)}{2n}[\mathbf{1}_{2n}\mathbf{1}_{2n}^\top - \begin{pmatrix} \mathbf{1}_n \\ -\mathbf{1}_n \end{pmatrix} \begin{pmatrix} \mathbf{1}_n \\ -\mathbf{1}_n \end{pmatrix}^\top]$ and round it. The main challenge is that as the graph becomes sparser, the noise in the node degrees become more important, and this can disrupt the second largest eigenvector from concentrating on the communities. To analyze how sparse the graph could be for the algorithm to exactly recover the communities, we express the augmented adjacency matrix as a perturbation of its expected value,

$$A^* = \mathbb{E}A^* + (\mathbb{E}A^* - A^*)$$

The spectral method will recover the true communities if the noise $\mathbb{E}A^* - A^*$ does not disrupt the first two eigenvectors of A^* to be somewhat aligned with those of $\mathbb{E}A^*$. We will discuss this in our future manuscripts.

Chapter 4

Numerical explorations

4.1 phase trasition plot

We will illustrate that SDP algorithm achieves the theoretical threshold $\sqrt{a}-\sqrt{b} > \sqrt{2}$ in Figure 4.1.

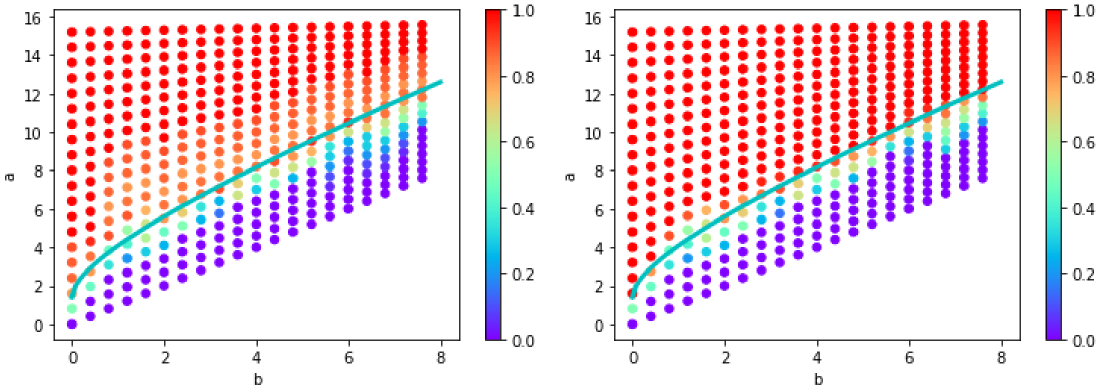

Figure 4.1 This plot shows that the empirical probability of success of the SDP based algorithm essentially matches the theoretical threshold of Lemma 3.2.4 in green. We fix $n = 300$ and $n = 500$ respectively and the number of trials to be 20. Then, at each trial and for xed $a > b$, we check how many times each method succeeds. Dividing by the number of trials, we obtain the empirical probability of success by generating the random matrix Λ corresponding to the correct communities $g = \{1, ..., 1, -1, ..., -1, 1, ..., 1, -1, ... - 1\}$ and check if Lemma 3.2.1 holds (which is the sufficient condition for SDP exact recovery). The green line indicates the theoretical threshold $\sqrt{a} - \sqrt{b} > \sqrt{2}$ for exact recovery.

Chapter 5

Conclusions and open probelms

This work mainly shows that the threshold for exact recovery in directed SBM is $\sqrt{a} - \sqrt{b} > \sqrt{2}$ and the our SDP algorithm can achieve this bound. Indeed, in our case of stochastic coblock model with two balanced communities $\mathcal{G}(n, p, q)$, the fundamental limit behavior is almost the same as that in the undirected case in [ABH14]. This might be because in our model $\mathbb{P}(a_{ij} = 1) = \mathbb{P}(a_{ji} = 1)$, which makes the behavior of rows and columns almost the same in expectation. Thus, the analysis of ScBM with inhomogenous connection probability ($\mathbb{P}(a_{ij} = 1) \neq \mathbb{P}(a_{ji} = 1)$) could be one of our future research direction. Our current work lies in the analysis of the directed spectral clustering algorithm in the model $\mathcal{G}(n, p, q)$ and most works have been done, which will soon be included in our next manuscript.

Although SBM is one of the most important benchmark model in community detection, it is still not fully understood, both directed and undirected. It is natural to expect that the results obtained in this work extend to a much more general family of network models. We list some open problems that could be our future research direction:

- (*Learning the general sparse SBM and ScBM*) Under what conditions can we learn the parameters in undirected $SBM(n, p, Q/n)$ and directed $SBM(n, p, Q/n)$. [ABH14]

- (*ScBM with inhomogenous connection probability*) Under what conditions can we learn the parameters in directed ScBM where $\mathbb{P}(a_{ij} = 1) \neq \mathbb{P}(a_{ji} = 1)$? Not only adjacency matrix, this model has also asymmetric expected adjacency matrix. Let $p_11 > p_2 > q_1 > q_2$, let G be a sample from the inhomogenous ScBM with two communities $ScBM(n, p_11, p_2, q_1, q_2)$. Let A

be the adjacency matrix ,then A_{ij} is a asymmetric Bernoulli random matrix with expectation

$$\mathbb{E}A = \begin{pmatrix} p_1 J_n & q_1 J_n \\ q_2 J_n & p_2 J_n \end{pmatrix}.$$

- (*Semi-supervised extensions*) How do the fundamental limits change in a semi-supervised setting, that is, the label of some node are exactly or probabilistically revealed before the clustering task ? [ABH14]

Chapter 6

Proofs

6.1 Proof for Theorem 3.1.1:

6.1.1 When does MLE fail?

We can find the upper bound of the threshold first, that is, the condition when MLE cannot exactly recover the clusters in the model. In this section we will prove that $ScBM(n, p = \frac{a\ln(n)}{n}, q = \frac{b\ln(n)}{n})$, $a > b$, if $|\sqrt{a} - \sqrt{b}| < \sqrt{2}$ then exact recovery is unsolvable (i.e.maximum likelihood estimator fails). Likelihood function: $L(x,y) = \prod_{i,j} P_{i,j}^{A_{i,j}}(1-P_{i,j})^{1-A_{i,j}}$. Let $E = \{\max\{L(\tilde{x},y), L(x,\tilde{y}), L(\tilde{x},\tilde{y}\} \geq L(x,y)\}$, which means there exists a pair of vertices (u,v) in different communites(in rows or colums), if their labels are flipped, the likelihood becomes larger. We want to show that if $|\sqrt{a} - \sqrt{b}| < \sqrt{2}$, then there is at least one bad vertex in rows or in columns, as defined below.

6.1.2 Degree and bad vertices

Definition 6.1.1. *We define the set of bad pairs of vertices(in rows, in colums or in rows and colums) by*

$$\mathcal{B}^R(G) := \{(u,v) : u \in C_1^R, v \in C_2^R, L(\tilde{x},y) > L(x,y)\}$$
$$\mathcal{B}^C(G) := \{(u,v) : u \in C_1^R, v \in C_2^R, L(x,\tilde{y}) > L(x,y)\}$$
$$\mathcal{B}^{R.C}(G) := \{(u,v) : u \in C_1, v \in C_2, L(\tilde{x},\tilde{y}) > L(x,y)\}$$

For a bad pair (u,v), the relationship between degrees can be infered from the relationship between likelihood functions:

$$
\begin{aligned}
&L(\tilde{x}, y) > L(x, y) \implies d_-^R(u) + d_-^R(v) > d_+^R(u) + d_+^R(v) \\
&L(x, \tilde{y}) > L(x, y) \implies d_-^C(u) + d_-^C(v) > d_+^C(u) + d_+^C(v) \\
&L(\tilde{x}, y) > L(x, y) \\
&\qquad \implies d_-^R(u) + d_-^R(v) + d_-^C(u) + d_-^C(v) > d_+^R(u) + d_+^R(v) + d_+^C(u) + d_+^C(v)
\end{aligned}
$$

Since if $L(\tilde{x}, y) > L(x, y)$, then $L(\tilde{x}, y) > L(x, y)$ or $L(x, \tilde{y}) > L(x, y)$. It is enough to only study the bad vertices in rows or in columns.

Definition 6.1.2. *We define the set of bad vertices in rows*

$$\mathcal{B}_i^R(G) = \{u : u \in C_i^R, d_+^R(u) \leq d_-^R(u) - 1\}, i = 1, 2$$

Lemma 6.1.1. *If $\mathcal{B}_1^R(G)$ is non-empty with probability $\frac{1}{2} + \Omega(1)$, the $\mathcal{B}^R(G)$ is non-empty with non-vanishing probability.*

Proof. If $u \in C_1^R$ and $v \in C_2^R$ such that $d_+^R(u) \leq d_-^R(u) - 1$ and $d_+^R(v) \leq d_-^R(v) - 1$, then $d_-^R(u) + d_-^R(v) > d_+^R(u) + d_+^R(v)$.
Therefore,

$$
\begin{aligned}
\mathbb{P}(\exists (u, v) \in \mathcal{B}^R(G)) &\geq \mathbb{P}(\exists u \in \mathcal{B}_1^R(G), \exists v \in \mathcal{B}_2^R(G)) \\
&\geq 2\mathbb{P}(\exists u \in \mathcal{B}_1^R(G)) - 1
\end{aligned}
$$

□

Note that $\mathbb{P}(\exists u \in \mathcal{B}_1^R(G)) = n\mathbb{P}(d_-^R(u) > d_+^R(u)) = n\mathbb{P}(Bin(n/2, q) > Bin(n/2, p)) = n^{1-(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}})^2+o(1)}$.

Lemma 6.1.2. *If $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then*

$$\mathbb{P}(\exists u \in \mathcal{B}_1^R(G)) = 1 - o(1)$$

Proof. We have

$$\mathbb{P}(\exists u \in \mathcal{B}_1^R(G)) = 1 - \mathbb{P}(\forall u \in C_1^R, u \notin \mathcal{B}_1^R(G))$$

Note that the events $\{u \notin \mathcal{B}_1^R(G)\}_{u \in C_1^R}$ are pairewise independent because the graph is directed and there is no possibility of two nodes sharing an edge. This implies that $\mathbb{P}(\exists u \in \mathcal{B}_1^R(G)) = 1 - o(1)$.

□

6.2 Proof for Lemma 3.2.1:

Proof. We can suppose W.L.O.G that $g = (1, ..., 1, -1, ..., -1, 1, ..., 1, -1, ..., -1)$. Firstly, by KKT conditions we obtain a sucient condition for $gg^\top$ to be a solution to SDP (2.2). This will give us $\Lambda \succcurlyeq 0$. $gg^\top$ is guaranteed to be an optimal solution to SDP (2.2) under the following conditions:

- $gg^\top$ is a feasible solution for the primal problem
- There exists a matrix Y feasible for the dual problem such that $\mathrm{Tr}((2A^* - J_{2n})gg^\top) = \mathrm{Tr}(Y)$.

The first point being trivially verified, it remains to find such a Y (known as a dual certificate). Let $C = 2A^* - J_{2n}$.

$$\begin{aligned}(Cgg^T)_{ii} &= \text{correct edges} + \text{correct non edges} - \text{incorrect edges} - \text{incorrect non edges}\\ &= (D_C^+)_{ii} + (\frac{n}{2} - (D_C^-)_{ii}) - (\frac{n}{2} - 1 - (D_C^+)_{ii}) - (D_C^-)_{ii}\\ &= 2((D_C^+)_{ii} - (D_C^-)_{ii}) + 1\end{aligned}$$

for $i \in [n+1, 2n]$, i.e in rows, let $j = i - n$:

$$(Cgg^T)_{ii} = 2((D_R^+)_{jj} - (D_R^-)_{jj}) + 1$$

Hence: $\mathrm{Tr}(Cgg^T) = \mathrm{Tr}(2(D_C^+ - D_C^-) + I_n) + \mathrm{Tr}(2(D_R^+ - D_R^-) + I_n)$. Thus

$$Y = \begin{bmatrix} 2(D_C^+ - D_C^-) + I_n & \mathbf{0} \\ \mathbf{0} & 2(D_R^+ - D_R^-) + I_n \end{bmatrix},$$

which verifies $\mathrm{Tr}((2A^* - J_{2n})gg^\top) = \mathrm{Tr}(Y)$ and, thus dened, is diagonal. Let Let $\Lambda = Y - C = 2\Gamma_{SBM} + I_{2n} + \begin{pmatrix} 0 & J_n - I_n \\ J_n - I_n & 0 \end{pmatrix} + 2\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix}$. As long as $\Lambda \succcurlyeq 0$, $gg^\top$ is an optimal solution to SDP (2.2).

$\lambda_2(\Lambda) > 0$ ensures that $gg^\top$ is the unique solution to SDP (2.2). Suppose X^* is another optimal solution to SDP (2.2), Then $\mathrm{Tr}(X'\Lambda) = 0$ by complementary slackness. By assumption, the second smallest eigenvalue of Λ is non-zero. Combining this with complementary slackness, the fact that $X' \succcurlyeq 0$ and $\Lambda \succcurlyeq 0$, We obtain that $X' = kgg^\top$. Since $X'_{ii} = 1$, $X' = gg^\top$ by contradiction. □

6.3 Proof for Lemma 3.2.2:

Proof. Note that

$$\begin{aligned}
&\mathbb{E}[\Lambda] = \mathbb{E}[2\Gamma_{SBM} + I_{2n} + \begin{pmatrix} 0 & J_n - I_n \\ J_n - I_n & 0 \end{pmatrix} + 2\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix}] \\
&= 2(\frac{n}{2}(p-q)I_{2n} - (\frac{p+q}{2}\begin{pmatrix} 0 & J_n \\ J_n & 0 \end{pmatrix} + \frac{p-q}{2}gg^\top)) \\
&+ \begin{pmatrix} 0 & J_n \\ J_n & 0 \end{pmatrix} + I_{2n} - \begin{pmatrix} 0 & I_n \\ I_n & 0 \end{pmatrix} + (p-q)\begin{pmatrix} g'g'^\top & 0 \\ 0 & g'g'^\top \end{pmatrix} + 2\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix} \\
&= n(p-q)(I_{2n} - \frac{\begin{pmatrix} 0 & g'g'^\top \\ g'g'^\top & 0 \end{pmatrix}}{n}) + (1-(p+q))\begin{pmatrix} 0 & J_n \\ J_n & 0 \end{pmatrix} + I_{2n} - \begin{pmatrix} 0 & I_n \\ I_n & 0 \end{pmatrix} \\
&+ 2\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix}
\end{aligned}$$

Suppose $p < \frac{1}{2}$, $\lambda_2 = n(p-q)$, corresponding to eigenvector perpendicular to $(g', g')^\top$ and $(\mathbf{1}, \mathbf{1})^\top$.
Let $\Lambda = 2\Gamma_{SBM} + I_{2n} + \begin{pmatrix} 0 & J_n - I_n \\ J_n - I_n & 0 \end{pmatrix} + 2\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix}$. By Weyl's inequalities,

$$\begin{aligned}
\lambda_2 > \lambda_{\max}(\mathbb{E}[\Lambda] - \Lambda) &= \|\mathbb{E}[\Lambda] - \Lambda\|_{op} \\
&\geq |\sigma_2(\mathbb{E}[\Lambda]) - \sigma_2(\Lambda)| \\
&= |\lambda_2(\mathbb{E}[\Lambda]) - \lambda_2(\Lambda)| \\
&\geq \lambda_2(\mathbb{E}[\Lambda]) - \lambda_2(\Lambda) \\
\implies \lambda_2(\Lambda) &> \lambda_2(\mathbb{E}[\Lambda]) - \lambda_2 = 0
\end{aligned}$$

which implies that $gg^\top$ is the unique solution to the semidefinite programming (2.1). □

6.4 Proof for Theorem 3.2.3:

Proof. The idea is to apply Theorem 2.3.3. One obstacle is that Γ_{SBM} is not a Laplacian since $\Gamma_{SBM}\mathbf{1} \neq \mathbf{0}$. Let g denote the vector that labels the nodes in columns and rows respectively, W.L.O.G, $g = \{\mathbf{1_{n/2}}, -\mathbf{1_{n/2}}, \mathbf{1_{n/2}}, -\mathbf{1_{n/2}}\}$. We define

$$\Gamma'_{SBM} = \text{diag}(g)\Gamma_{SBM}\text{diag}(g)$$

Note that Γ'_{SBM} is a laplacian and both the eigenvalues and diagonal elements of $\mathbb{E}[\Gamma'_{SBM}] - \Gamma'_{SBM}$ are the same as those of $\mathbb{E}[\Gamma_{SBM}] - \Gamma_{SBM}$. Note that the off-diagonal entries of $\Gamma'_{SBM} = -A_{ij}g_ig_j$. We apply Theorem 2.1 to $L = \mathbb{E}[\Gamma'_{SBM}] -$

Γ'_{SBM},

$$\sigma^2 = \sum_{j\in[2n]\setminus\{i\}} \mathbb{E}[L_{ij}^2] = (\frac{n}{2}-1)p(1-p) + \frac{n}{2}q(1-q)$$

$$\geq \frac{n}{2}\cdot\frac{1}{4}(p+q) > \frac{n}{8}\frac{2\log(n)}{\epsilon n} \geq \frac{\log(n)}{4\epsilon}(1-q)^2 = \frac{\log(n)}{4\epsilon}\max_{i\neq j}\|L_{ij}\|_\infty^2.$$

Hence, there exists a constant Δ' such that, with high probability,

$$\lambda_{\max}(\mathbb{E}[\Gamma'_{SBM}] - \Gamma'_{SBM}) \leq (1+\frac{\Delta'}{\sqrt{\log(n)}})\max_{i\in[2n]}[\mathbb{E}[(\Gamma'_{SBM})_{ii}] - (\Gamma'_{SBM})_{ii}]$$

which is equivalent to

$$\lambda_{\max}(\mathbb{E}[\Gamma_{SBM}] - \Gamma_{SBM}) \leq (1+\frac{\Delta'}{\sqrt{\log(n)}})\max_{i\in[2n]}[\mathbb{E}[(\Gamma_{SBM})_{ii}] - (\Gamma_{SBM})_{ii}].$$

Note that

$$\min_{i\in[2n]}((\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii}) \geq \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg_C^+(i) - \deg_C^-(i))]$$

$$\iff \max_{i\in[2n]}(\mathbb{E}[(\Gamma'_{SBM})_{ii}] - (\Gamma'_{SBM})_{ii}) \leq (1-\frac{\Delta}{\sqrt{\log(n)}})(\frac{n}{2}(p-q)-p)$$

Therefore,

$$\lambda_{\max}(\mathbb{E}[\Gamma'_{SBM}] - \Gamma'_{SBM}) \leq (1+\frac{\Delta'}{\sqrt{\log(n)}})(1-\frac{\Delta}{\sqrt{\log(n)}})(\frac{n}{2}(p-q)-p)$$

For each Δ', there exists $\Delta' > 0$, such that,

$$(1+\frac{\Delta'}{\sqrt{\log(n)}})(1-\frac{\Delta}{\sqrt{\log(n)}}) < 1 \tag{6.1}$$

Then

$$\lambda_{\max}(\mathbb{E}[\Gamma'_{SBM}] - \Gamma'_{SBM}) < (\frac{n}{2}(p-q))$$

which garuentees the exact recovery of the semidefinite programming. □

In order to prove Lemma 3.2.4, we will use some conclusions from [ABH14] and [Ban15].

Definition 6.4.1. *(Definition 3 in [ABH14]) Let m be a natural number, $p, q \in [0,1]$, and $\delta \in \mathbb{R}$, we define*

$$T(m,p,q,\delta) = \mathbb{P}[\sum_{i=1}^{m}(Z_i - W_i) \geq \delta],$$

where $W_1, ..., W_m$ are i.i.d Bernoulli(p) and $Z_1, ..., Z_m$ are i.i.d Bernoulli(q) independent of $W_1, ..., W_m$.

Lemma 6.4.1. *Recall definition 6.4.1. Let a, b, Δ' be constants. Then*

$$T(\frac{n}{2}, \frac{a\log(n)}{n}, \frac{b\log(n)}{n}, -\Delta'\sqrt{\log(n)}) \leq \exp[-(\frac{a+b}{2} - \sqrt{ab} - \delta(n))\log(n)]$$

with $\delta(n) \to 0$.

Proof. This is obtained by straightforward adaptions to the proof of Lemma 8 in [ABH14]. □

Proof. (of **Lemma** 3.2.4) Let a and b satisfy $\sqrt{a} - \sqrt{b} > \sqrt{2}$. Given $\Delta > 0$, we want to show that, with high probability

$$\min_{i\in[2n]}((\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii}) \geq \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg^+_C(i) - \deg^-_C(i))] = \frac{\Delta}{\sqrt{\log(n)}}\frac{n}{2}(p-q).$$

For fixed i throughout the rest of the proof. Clearly,

$$(\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii} = (\sum_{i=1}^{\frac{n}{2}-1} W_i) - (\sum_{i=1}^{\frac{n}{2}} Z_i) = \sum_{i=1}^{\frac{n}{2}-1}(W_i - Z_i) + Z_{\frac{n}{2}}$$

where $W_1, ..., W_m$ are i.i.d Bernoulli(p) and $Z_1, ..., Z_m$ are i.i.d Bernoulli(q) independent of $W_1, ..., W_m$. Thus

$$\begin{aligned}
&\mathbb{P}((\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii} < \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg^+_C(i) - \deg^-_C(i))]) \\
&= \mathbb{P}(\sum_{i=1}^{\frac{n}{2}-1}(W_i - Z_i) + Z_{\frac{n}{2}} < \Delta\sqrt{\log(n)}(\frac{a-b}{2})) \\
&= \mathbb{P}(\sum_{i=1}^{\frac{n}{2}-1}(Z_i - W_i) - Z_{\frac{n}{2}} > -\Delta\sqrt{\log(n)}(\frac{a-b}{2})) \\
&\leq \mathbb{P}(\sum_{i=1}^{\frac{n}{2}-1}(Z_i - W_i) > -\Delta\sqrt{\log(n)}(\frac{a-b}{2}))
\end{aligned}$$

Let $\Delta' = (\frac{a-b}{2})\Delta$, By Lemma 6.4.1, for fixed i,

$$\begin{aligned}
&\mathbb{P}((\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii} < \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg^+_C(i) - \deg^-_C(i))]) \\
&\leq T(\frac{n}{2}, \frac{a\log(n)}{n}, \frac{b\log(n)}{n}, -\Delta'\sqrt{\log(n)}) \\
&\leq \exp[-(\frac{a+b}{2} - \sqrt{ab} - \delta(n))\log(n)]
\end{aligned}$$

where $\delta(n) \to 0$. By union bound,

$$\begin{aligned}
&\mathbb{P}(\min_{i\in[2n]}[(\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii}] < \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg_C^+(i) - \deg_C^-(i))]) \\
&\leq 2n\exp[-(\frac{a+b}{2} - \sqrt{ab} - \delta(n))\log(n)] \\
&= 2\exp[-(\frac{a+b}{2} - \sqrt{ab} - 1 - \delta(n))\log(n)]
\end{aligned}$$

As long as $\frac{a+b}{2} - \sqrt{ab} > 1$, $\mathbb{P}(\min_{i\in[2n]}[(\mathcal{D}^+)_{ii} - (\mathcal{D}^-)_{ii}] > \frac{\Delta}{\sqrt{\log(n)}}\mathbb{E}[\deg_C^+(i) - \deg_C^-(i))]) = 1 - 2n^{-\epsilon}$, $\epsilon > 0$. □

References

[Wey12] Hermann Weyl. “Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)”. In: *Mathematische Annalen* 71.4 (1912), pp. 441–479. DOI: 10.1007/BF01456804. URL: https://doi.org/10.1007/BF01456804.

[ER59] P. Erdös and A. Rényi. “On Random Graphs I”. In: *Publicationes Mathematicae Debrecen* 6 (1959), pp. 290–297.

[Har72] J. A. Hartigan. “Direct Clustering of a Data Matrix”. In: *Journal of the American Statistical Association* 67.337 (1972), pp. 123–129. DOI: 10.1080/01621459.1972.10481214. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1972.10481214. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481214.

[Lov79] L. Lovasz. “On the Shannon capacity of a graph”. In: *IEEE Transactions on Information Theory* 25.1 (1979), pp. 1–7. DOI: 10.1109/TIT.1979.1055985.

[HJ85] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. DOI: 10.1017/CBO9780511810817.

[WW87] Yuchung J. Wang and George Y. Wong. “Stochastic Blockmodels for Directed Graphs”. In: *Journal of the American Statistical Association* 82.397 (1987), pp. 8–19. ISSN: 01621459. URL: http://www.jstor.org/stable/2289119.

[See90] D. Seese. “Groetschel, M., L. Lovasz, A. Schrijver: Geometric Algorithms and Combinatorial Optimization. (Algorithms and Combinatorics. Eds.: R. L. Graham, B. Korte, L. Lovasz. Vol. 2), Springer-Verlag 1988, XII, 362 pp., 23 Figs., DM 148,-. ISBN 978-3-642-78240-4”. In: *Biometrical Journal* 32.8 (1990), pp. 930–930. DOI: https://doi.org/10.1002/bimj.4710320805. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.4710320805. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710320805.

[GW95] Michel X. Goemans and David P. Williamson. “Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming”. In: *J. ACM* 42.6 (Nov. 1995), 1115?1145. ISSN: 0004-5411. DOI: 10.1145/227683.227684. URL: https://doi.org/10.1145/227683.227684.

[Chu97] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[HRW00] C. Helmberg, F. Rendl, and R. Weismantel. “A Semidefinite Programming Approach to the Quadratic Knapsack Problem”. In: *Journal of Combinatorial Optimization* 4.2 (2000), pp. 197–215. DOI: 10.1023/A:1009898604624. URL: https://doi.org/10.1023/A:1009898604624.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. “On Spectral Clustering: Analysis and an Algorithm”. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. NIPS’01. Vancouver, British Columbia, Canada: MIT Press, 2001, pp. 849–856.

[GN02] M. Girvan and M. E. J. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826. ISSN: 0027-8424. DOI: 10.1073/pnas.122653799. eprint: https://www.pnas.org/content/99/12/7821.full.pdf. URL: https://www.pnas.org/content/99/12/7821.

[New03] M. E. J. Newman. “The Structure and Function of Complex Networks”. In: *SIAM Review* 45.2 (Jan. 2003), 167?256. ISSN: 1095-7200. DOI: 10.1137/s003614450342480. URL: http://dx.doi.org/10.1137/S003614450342480.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. USA: Cambridge University Press, 2004. ISBN: 0521833787.

[Lux07] Ulrike von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4 (2007), pp. 395–416. DOI: 10.1007/s11222-007-9033-z. URL: https://doi.org/10.1007/s11222-007-9033-z.

[Kno08] Andreas Knoblauch. “Closed-form Expressions for the Moments of the Binomial Probability Distribution”. In: *SIAM Journal of Applied Mathematics* 69 (Jan. 2008), pp. 197–204. DOI: 10.1137/070700024.

[LFR08] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. “Benchmark graphs for testing community detection algorithms”. In: *Physical Review E* 78.4 (Oct. 2008). ISSN: 1550-2376. DOI: 10.1103/physreve.78.046110. URL: http://dx.doi.org/10.1103/PhysRevE.78.046110.

[LN08] E. A. Leicht and M. E. J. Newman. “Community Structure in Directed Networks”. In: *Phys. Rev. Lett.* 100 (11 Mar. 2008), p. 118703. DOI: 10.1103/PhysRevLett.100.118703. URL: https://link.aps.org/doi/10.1103/PhysRevLett.100.118703.

[RB08] Martin Rosvall and Carl T. Bergstrom. “Maps of random walks on complex networks reveal community structure”. In: *Proceedings of the National Academy of Sciences* 105.4 (2008), pp. 1118–1123. ISSN: 0027-8424. DOI: 10.1073/pnas.0706851105. eprint: https://www.pnas.org/content/105/4/1118.full.pdf. URL: https://www.pnas.org/content/105/4/1118.

[LF09] Andrea Lancichinetti and Santo Fortunato. "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities". In: *Phys. Rev. E* 80 (1 July 2009), p. 016118. DOI: 10.1103/PhysRevE.80.016118. URL: https://link.aps.org/doi/10.1103/PhysRevE.80.016118.

[For10] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3-5 (Feb. 2010), 75?174. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002. URL: http://dx.doi.org/10.1016/j.physrep.2009.11.002.

[Dec+11a] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Phys. Rev. E* 84 (6 Dec. 2011), p. 066106. DOI: 10.1103/PhysRevE.84.066106. URL: https://link.aps.org/doi/10.1103/PhysRevE.84.066106.

[Dec+11b] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Physical Review E* 84.6 (Dec. 2011). ISSN: 1550-2376. DOI: 10.1103/physreve.84.066106. URL: http://dx.doi.org/10.1103/PhysRevE.84.066106.

[MV13] Fragkiskos D. Malliaros and Michalis Vazirgiannis. "Clustering and Community Detection in Directed Networks: A Survey". In: *CoRR* abs/1308.0971 (2013). arXiv: 1308.0971. URL: http://arxiv.org/abs/1308.0971.

[ABH14] Emmanuel Abbe, A. S. Bandeira, and Georgina Hall. "Exact Recovery in the Stochastic Block Model". In: *CoRR* abs/1405.3267 (2014). arXiv: 1405.3267. URL: http://arxiv.org/abs/1405.3267.

[Mas14] Laurent Massoulié. "Community Detection Thresholds and the Weak Ramanujan Property". In: *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*. STOC '14. New York, New York: Association for Computing Machinery, 2014, 694?703. ISBN: 9781450327107. DOI: 10.1145/2591796.2591857. URL: https://doi.org/10.1145/2591796.2591857.

[AS15] Emmanuel Abbe and Colin Sandon. *Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms*. 2015. arXiv: 1503.00609 [math.PR].

[Aga+15] Naman Agarwal et al. "Multisection in the Stochastic Block Model using Semidefinite Programming". In: *CoRR* abs/1507.02323 (2015). arXiv: 1507.02323. URL: http://arxiv.org/abs/1507.02323.

[Ban15] Afonso S. Bandeira. *Random Laplacian matrices and convex relaxations*. 2015. arXiv: 1504.03987 [math.PR].

[RQY15] Karl Rohe, Tai Qin, and Bin Yu. *Co-clustering for directed graphs: the Stochastic co-Blockmodel and spectral algorithm Di-Sim*. 2015. arXiv: 1204.2296 [stat.ML].

[AL16] Arash A. Amini and Elizaveta Levina. *On semidefinite relaxations for the block model*. 2016. arXiv: 1406.5647 [cs.LG].

[Abb17] Emmanuel Abbe. *Community detection and stochastic block models: recent developments*. 2017. arXiv: 1703.10146 [math.PR].

[KBG18] Chiheon Kim, Afonso S. Bandeira, and Michel X. Goemans. *Stochastic Block Model for Hypergraphs: Statistical limits and a semidefinite programming approach*. 2018. arXiv: 1807.02884 [math.PR].

[MNS18] Elchanan Mossel, Joe Neeman, and Allan Sly. "A Proof of the Block Model Threshold Conjecture". In: *Combinatorica* 38.3 (2018), pp. 665–708. DOI: 10.1007/s00493-016-3238-8. URL: https://doi.org/10.1007/s00493-016-3238-8.

[DLS20] Shaofeng Deng, Shuyang Ling, and Thomas Strohmer. *Strong Consistency, Graph Laplacians, and the Stochastic Block Model*. 2020. arXiv: 2004.09780 [stat.ML].

Acknowledgements

I would like to thank my advisor, Prof. Shuyang Ling, for his ongoing guidance and support, his frequent feedback at every stage of this project, as well as his commitment to introspection, and to reflecting upon and exploring meaningful issues in mathematics in data science. This thesis would not have come to fruition without the help of him.

In addition, I would also like to thank my family for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my girlfriend Carol who provided good-to-excellent care as well as happy distractions to rest my mind outside of my research.