STATISTICAL MODELING AND ESTIMATION OF COMMUNITY STRUCTURES IN NETWORKS

by

Song Wang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 09/22/2017

The dissertation is approved by the following members of the Final Oral Committee:

Karl Rohe, Associate Professor, Statistics
Garvesh Raskutti, Assistant Professor, Statistics
Kam-Wah Tsui, Professor, Statistics
Xiaojin Zhu, Professor, Computer Sciences
Daniel Sussman, Assistant Professor, Statistics

To my father Yujun Wang, mother Chun Xiang and brother Xing Wang.

Contents

List of Tables

List of Figures

Abstract

Networks and graphs represent the complex relationships among people or objects and allow us to study the patterns of these relationships. Many networks, especially social networks, display the community structures, where certain sets of nodes are tightly connected among themselves while loosely connected with the others. Identifying these communities, often called community detection, is an important but difficult task.

In this thesis, we have made solid contributions to community detection literature by proposing new random graph models with desired properties, offering consistency analysis of existing algorithms, and devising new tractable algorithms for community detection. We present them as below.

In Chapter 1, we propose a new random graph model by incorporating the Stochastic Block Model (SBM) with the Random Intersection Graph (RIG) to allow for high transitivity and rich local structures, which are missed by the popular SBM and its variants. Furthermore, we re-study Spectral Clustering, a popular community detection algorithm, and prove that Spectral Clustering can still estimate the community consistently, but with a slower error convergence rate.

In Chapter 2, we model the scaled eigenvectors of the adjacency matrix using a Gaussian Mixture Model (GMM) with covariance structure given by a Central Limit Theorem and propose approximate Expectation-Maximization algorithms to estimate the parameters and memberships under SBM and its variant, Degree-Corrected SBM. In the simulation, our algorithms show clear improvements over many existing algorithms.

In Chapter 3, we propose a two-stage model which starts with a network under the SBM and adds new edges randomly with probability proportional to the number of common neighbors. This model creates edge dependence and generates diverse local structures. We also prove the consistency result for Spectral Clustering similar to that in Chapter 1.

In Chapter 4, we propose a new algorithm to uncover the memberships under a variant of SBM, where nodes can belong to multiple communities. Based on the spectral priorities of the graph, we propose an algorithm that projects the spectral representations of nodes onto random directions iteratively to distinguish pure nodes and mixed nodes. Furthermore, we conduct spectral analysis and prove the consistency of our proposed algorithm.

Acknowledgments

It has been an invaluable and unforgettable experience for me to conduct my PhD research at University of Wisconsin - Madison, where I have met so many great individuals.

First and foremost, I would like to express my deepest thanks to my thesis advisor Professor Karl Rohe for his great guidance and generous support. This thesis would not have been possible without his inspirational ideas and guidance. He is such an enthusiastic and energetic person, who has influenced me in many perspectives. Particularly, I am grateful that he introduced me into the research about spectral graph theory and community detection and helped me grow as an independent researcher.

Besides my advisor, I am also grateful to the other oral defense committee members who were more than generous with precious time and their expertise. Thank you, Professor Jerry Zhu, Professor Kam-Wah Tsui, Professor Garvesh Raskutti and Professor Daniel Sussman for serving on my committee.

I am very grateful to my collaborators: Professor Mindaugas Bloznelis, Professor Daniel Sussman, Professor Chris Wells and Ms. Yini Zhang, for the different projects

we worked on together. I greatly enjoyed those memorable experiences and learned enormously from those collaborations.

I am thankful to Professor Zhengjun Zhang for selflessly offering me enormous encouragement and help during my stay. I am thankful to Professor Peter Qian for the great guidance when I worked as his project assistant in the first summer. I am very thankful to Dr. Glenn Fung, my mentor during my internship at American Family Insurance, for introducing me into the exciting world of using machine learning to solve real problems and sharing his expertise with me. I would like to express my sincere thanks to many other professors in the department for the great courses I took from them and the invaluable discussions I had with them.

I thank the lab members Tai Qin, Norbert Binkiewicz, Juhee Cho, Thu Le, Yilin Zhang, Fan Chen, and Muze Zeng for the inspirational discussions. It is also a great pleasure to thank my classmates and friends Jared Huling, Subhrangshu Nandi, Han Chen, Shengji Jia and many others for the lasting memories we have created together. I would also like to thank Mike Cammilleri for helping me solve so many computational problems related to the department computer and Emma Krauska for editing my paper and thesis drafts.

Last but not the least, my deep and sincere gratitude goes to my family members for their continuous and unparalleled love and support. Their unwavering belief in me and endless love are the sources of strength for me. I dedicate this milestone to them.

Chapter 1

Modeling and Estimating Block Structures in Random Intersection Graphs

Abstract

Community Detection is a popular and difficult algorithmic task in network analysis. In this paper, we propose a composite model called Random Intersection Graph with Blocks (RIGB) by introducing block structures as in the Stochastic Block Model (SBM) into the RIGs. Networks sampled from the new model have both riche local structures (e.g. high number of triangles and high transitivity), which are missed by the SBM and its variants, and community structures. As such, these graphs work as better benchmark networks to understand, evaluate, and improve various community detection algorithms. Under this model, we re-study the performance of the Spectral Clustering and prove that spectral clustering still estimates communities consistently under RIGB as it does under SBM with the upper bound for the error rate converging at slower rate than it is under the SBM. In simulation, we compare local structures in terms of transitivity ratio under RIGBs and SBMs; and affirm that it is hard for spectral clustering and other various existing community algorithms to recover communities under RIGB than under SBMs.

1.1 Introduction

Networks and graphs are powerful tools to represent the complicated interactions among different people or objects. Examples include friendships on Facebook, protein interactions in biological networks, and many others. Many such networks can be decomposed into sets of tightly connected subcomponents, which are called *communities*. Identifying the communities is an essential question in the field of network research.

Among the many existing statistical models for networks with communities, the Stochastic Blockmodel (SBM) is one of the most commonly used to study community detection algorithms. In an SBM, there are n nodes from K disjoint groups or communities and the links between each of the two nodes within or across communities form independently with probabilities only dependent on the memberships of the two nodes. Under this model or its generalization, people have developed various methods to recover the block memberships, such as modularity maximization (Newman and Girvan (2004) etc), likelihood methods (Bickel and Chen (2009), Amini et al. (2013) etc), spectral clustering (Ng et al. (2001), Rohe et al. (2011), Lei and Rinaldo (2013) etc) and other methods.

However, the usefulness of this research as a guide to practice depends on the realism of the Stochastic Blockmodel. As mentioned in Newman (2006), two common features of empirical networks are long-tailed degree distribution and high transitivity. The SBM assumes that nodes in the same block are stochastically equivalent, which cannot model the variation of degrees for different nodes. To create long-tailed distribution, the degree-corrected SBM (DC-SBM) (Karrer and Newman (2011))

introduces a degree parameter for each node. Recently, Zhao et al. (2012), Qin and Rohe (2013), Jin (2012) and Lei and Rinaldo (2013) generalized the clustering results under SBMs to that under DC-SBMs, illustrating how spectral clustering algorithms must adapt when the degree distribution is skewed.

The other feature, high transitivity (also known as clustering coefficients) commonly observed in real networks, has been missed by SBMs and DC-SBMs. Take as an example the political blog network in Adamic and Glance (2005), a common benchmark network for community detection, Table 1.1 shows that there are far fewer triangles and far lower transitivity ratio in the fitted SBM than those in the original network.

Table 1.1: Number of triangles and transitivity of political blog data and fitted SBM.

	mean degree	no. of triangles	transivity
real network	27.36	101043	0.226
fitted SBM	27.33(0.21)	5235.3 (137)	0.034 (6e-4)

In the literature, Exponential Random Graph Models (ERGMs) directly model the extra dependence by incorporating counts of local structures such as number of edges, p-stars and triangles as sufficient statistics in the model; see Robins et al. (2007) and references therein. However, it is notoriously hard to sample and fit ERGMs since the normalization constant is hard to compute; and MCMC procedures commonly used in the estimation are lacking convergence guarantees (see Hunter et al. (2008); Chatterjee et al. (2013); Chandrasekhar and Jackson (2014)).

Singer-Cohen (1995) introduced Random Intersection Graphs (RIGs). A simple form of the RIG is as follows: Given a probability $p \in [0, 1]$, a vertex set $V =$

$\{1, 2, \cdots, n\}$ and an attribute set $W = \{a_1, a_2, \cdots, a_m\}$, the bipartite graph H among V and W is first constructed by pairing up nodes in V and W randomly with probability p. Then based on H, the random intersection graph on vertex set V is formed by linking two vertices in V if they share at least one neighbor in bipartite graph H. There has been a lot of interest in the various properties of RIGs and its generalizations. In particular, Deijfen and Kets (2009) studied the transitivity under RIGs with tunable degree distributions and Bloznelis (2013) study the general relationships between degree distribution and transitivity (called clustering coefficients in that paper) under RIGs. These studies show that RIGs can have high transitivity even if the graph is sparse, which is missed by SBMs.

In this paper, we propose a new model by fusing RIGs with SBMs and refer this fusion as RIGs with Blocks (**RIGB**, details in Section 1.3). This is a meaningful addition to both sides of the literature: i) introducing the new concept about blocks or global community structures into RIG literature; and ii) generating SBM-type of networks with high transitivity and thus offering a better bench model for community detection algorithms. Under this more realistic model, we re-study the spectral clustering algorithm, which is a widely-used community detection algorithm and has provable consistent performance under the Stochastic Blockmodel. We prove that it can still recover the blocks/communities under our new model consistently. However, we find the error rate of spectral clustering converges to zero at a slower rate. In the simulation study, under different settings, we affirm the properties of high transitivity of RIGB and compare the convergence rate for spectral clustering under RIGBs than under SBMs.

The paper is organized as follows: The preliminaries are in Section 3.2. We present model set-up and its properties in Section 1.3. We provide the main results including the consistency of the spectral clustering algorithm to uncover communities under RIGB in Section 1.4. We offer the simulation study to further compare the properties of RIGB with that of SBM in Section 1.5. Finally, we put all the proofs in the Appendix.

Some notations: For a matrix M, $\|M\|$ stands for the spectral norm of M and $\|M\|_F$ stands for the Frobenius norm. M_i stands for the i-th row of M, and $M_{\cdot j}$ stands for the j-th column of M. $[n]$ is the shorthand notation for the set of integers from 1 to n. I_K is identity matrix of dimension K; while J is a matrix with each cell taking value 1. For two nodes i and j in a graph, $i \sim j$ denotes that there is an edge between i and j or i and j are adjacent.

1.2 Preliminaries

Spectral Clustering

Spectral clustering is a popular clustering algorithm used in various settings, such as image segmentations (Shi and Malik (2000)) and researchers studied the theoretical properties of the algorithm (Ng et al. (2001), Von Luxburg (2007)). Recently, the algorithm has become a popular community detection algorithm and people have proved the consistency of spectral clustering under the SBM and its variants (e.g., Rohe et al. (2011); Lei and Rinaldo (2013)). There are many variants of spectral clustering available and they all involve calculating the leading eigenvectors of a

similarity matrix and clustering in the low-dimensional eigen-space. We will use the following version of spectral clustering.

Spectral Clustering

- Input: a matrix $A \in \{0,1\}^{n\times n}$ and the desired number of clusters K.
- Output: a K-set partition of node set $\{1, 2, \cdots, n\}$.

Step 1: Calculate the first K leading eigenvectors of A, denoted as $X_1, X_2, \cdots, X_K \in R^n$. Let $X = [X_1, X_2, \cdots, X_K] \in R^{n\times K}$.

Step 2: Normalize rows in X such that they all have unit norm, denoted as X^*. That is,

$$X^*_{ij} = \frac{X_{ij}}{\sqrt{\sum_{j=1}^K X_{ij}^2}}, \quad i = 1, 2, \cdots n, j = 1, 2, \cdots, K.$$

Step 3: Run k-means algorithm on rows in X^* with number of clusters K, and obtain K non-overlapping clusters for rows in X^*.

Stochastic Block Model

The Stochastic Block Model (Holland et al. (1983)) is a popular random network model to study networks with community structures. **SBM** states that the nodes $\{1, 2, \ldots, n\}$ come from K different blocks and the probability that they connect depends only on their memberships. Let $B \in [0,1]^{K\times K}$ be the probability matrix among

blocks and $z \in \mathbb{R}^n$ be the membership vector with i-th element $z_i \in \{1, 2, \cdots, K\}$ being the block to which node i belongs. For nodes $i, j (i < j)$, the probability that they connect is $B_{z_i z_j}$. Furthermore, all the links among $n(n-1)/2$ different pairs are assumed to be generated independently.

There have been various generalizations to original SBM. DC-SBM (Karrer and Newman (2011)) added a new set of degree parameters to allow degree heterogeneity within the cluster and allow the sampled networks to have the desired heavy tailed degree distributions. The new model proposed in this paper will incorporate degree corrections as DC-SBM does.

Random Intersection Graph

The random intersection graph was introduced in Singer-Cohen (1995) and Karoński et al. (1999), and has been generalizations (for details see survey paper Bloznelis et al. (2015a)). In its simplest form, the random intersection graph model is defined as follows:

1. Let $V = \{1, \cdots, n\}$ be a set of vertices and $W = \{a_1, \cdots, a_m\}$ be a set of attributes. For $p \in [0, 1]$, a bipartite graph on $V \cup W$, denoted as H, is constructed by linking vertices in V and attributes in W independently with probability p.

2. Form a link between vertices $i, j \in V$, if and only if there is an attribute $a_u \in W$ such that both i and j are adjacent to a_u in H.

In the social network setting, we can think of V as the set of people and W as the set of social clubs. People will form friendships after they meet in some common clubs or social groups. In the following, we will frequently use the terminology from the social networks with understanding that the model can be, of course, more general.

Based on different distributional assumptions to sample the bipartite graph H, there are different types of random intersection graphs featuring different properties. Examples include the Binomial Intersection Graph and the Inhomogeneous Intersection Graph (details in Bloznelis et al. (2015a), Bloznelis et al. (2015b)).

Transitivity of a Graph

Transitivity is an important property shared by many social networks and various other networks. It refers to the extent that two nodes are connected if they both connect to a common node in the network. A triad is defined as a connected subgraph consisting of three vertices and two edges, e.g. $i \sim j \sim k$. We call a triad $i \sim j \sim k$ transitive if there is an edge between i and k, i.e., $i \sim k$ (see Page 243 Wasserman and Faust (1994)).

The transitivity ratio quantifies transitivity in an undirected network based on counting the number of triads. For an undirected graph $G = (V, E)$, where $V = \{1, 2, \cdots n\}$ is the vertex set and $E = \{(i, j) : \text{there is an edge between } i \text{ and } j\}$ is the edge set. The transitivity ratio of G is defined as follows:

$$T(G) \triangleq \frac{3 \times \text{number of triangles in } G}{\text{number of triads in } G}, \tag{1.1}$$

where the 3 in the numerator comes from the fact that each triangle forms three connected triads. Intuitively, $T(G)$ characterizes the density of triangles in G. It takes values within the range of 0 and 1 with $T(G) = 1$ if G is fa union of isolated cliques (complete graph) and $T(G) = 0$ if there are no triangles.

The clustering coefficient (Watts and Strogatz (1998)) quantifies transitivity locally. For the graph G above, define the local clustering coefficient at vertex i as follows:

$$C_i \triangleq \frac{\text{number of triangles incident to } i}{\text{number of triads incident to } i} = \frac{|\{(j,k) \in E : j \in N_i, k \in N_i, \}|}{d_i(d_i-1)/2}, \tag{1.2}$$

where $N_i \triangleq \{j \in V : (i,j) \in E\}$ is the neighborhood of i, and $d_i = |N_i|$ is the degree of i and $d_i(d_i-1)/2$ is the number of all potential links among vertices in N_i. As an alternative to the transitivity ratio, the average of the local clustering coefficients of all the vertices is defined as measure for transitivity for the whole graph G:

$$C(G) \triangleq \frac{1}{n}\sum_{i=1}^{n} C_i. \tag{1.3}$$

It is worth mentioning that in RIG literature, the following conditional probability (we refer to it as transitive probability) characterizes transitivity (see, e.g., Deijfen and Kets (2009), Bloznelis (2013)):

$$\alpha_C = \mathbf{P}(i^* \sim j^* | i^* \sim k^*, j^* \sim k^*). \tag{1.4}$$

where $\sim$ refers to the adjacency relation, and $\mathbf{P}$ refers to all the sources of randomness

defining the events considered (these are the uniform sampling of vertices (i^*, j^*, k^*) and random graph generation mechanism in the present context). Bloznelis and Kurauskas (2016) studied the relationships between transitivity of a realized graph G, denoted as $T(G)$, and the conditional probability α_C of the underlying RIG models.

1.3 Random Intersection Graphs with Blocks

SBMs have global community structures, but it assumes all the edges are independent, thus in sparse graphs, they are locally tree-like and fail to represent the rich variety of local topology that we observe in empirical networks. For example, a network sampled from a sparse SBM will have far fewer triangles compared to many empirical networks, and the clustering coefficients are lower than those in empirical graphs. On the other hand, RIGs induce edge dependence through an intermediate layer of latent random variables. Several papers proved that the clustering coefficients of sparse graphs from RIGs can be bounded away from 0 (Deijfen and Kets (2009), Bloznelis (2013)). Here we will propose a new model to combine the advantages of the two types of models.

Model Setup

Let n, m be integers, and $V = \{1, 2, \ldots, n\}$ be the set of vertices ("people") and $W = \{a_1, a_2, \cdots, a_m\}$ be the set of attributes ("social clubs"). Vertices tend to have different attributes and they tend to form links if they share some common attributes.

Different from the traditional RIGs, we introduce here block structures into the

vertex and attribute set. More specifically, assume the partitions of of vertex set V and W:

$$V = V_1 \cup V_2 \cup \cdots \cup V_K, \quad W = W_1 \cup W_2 \cup \cdots \cup W_K,$$

where K is an integer, $V_1, V_2, ..., V_K$ are K non-overlapping subsets of V and $W_1, W_2, ..., W_K$ are K non-overlapping subsets of W. Vertices in V_k are expected to have higher tendency to have attributes from W_k and lower tendency to have attributes from other subsets. Let $z \in R^n, y \in R^m$ be membership vectors taking values from $\{1, 2, \cdots, K\}$ with $z_i = k(1 \leq i \leq n)$, meaning that vertex i belongs to V_k and $y_u = k(1 \leq u \leq m)$, meaning that attribute a_u belongs to W_k.

Definition 1.1 (Random Intersection with Blocks (RIGB))**.** *Let the vertex set $V = \{1, 2, \ldots, n\}$ and the attribute set $W = \{a_1, a_2, \cdots, a_m\}$ with the above partitions. Let z, y be the two membership vectors representing the partition. Also, assume there are two matrices: $\mathbf{M} \in [0,1]^{K \times K}$ and a symmetric matrix $\mathbf{B} \in [0,1]^{K \times K}$. We call the random graph A a Random Intersection Graphs with Blocks (RIGB) with parameters $(\mathbf{M}, \mathbf{B}, z, y)$ if it is sampled from the following two-step procedure.*

1. *Sample bipartite graph $H \in \{0,1\}^{n \times m}$ among V and W.*

   For vertex i of type z_i, and attribute a_u of type y_u, use $H_{iu} \in \{0,1\}$ as the indicator for whether vertex i has the attribute a_u. Assume the probability is $\mathbf{M}_{z_i, y_u}$. That is,

$$P(H_{iu} = 1) = \mathbf{M}_{z_i, y_u}, \quad i \in [n], u \in [m]. \tag{1.5}$$

Furthermore, all relationships $\{H_{ij}, i \in [n],\ u \in [m]\}$ *in* H *are formed independently.*

2. *Given* H*, sample intersection graph* $A \in \{0,1\}^{n \times n}$ *on* V*.*

   For two different vertices $i, j (i < j)$*, use* A_{ij} *as the indicator for whether* i, j *form a link. If* i, j *have shared at least one common attribute, i.e.* $\sum_{u=1}^{m} H_{iu} H_{ju} > 0$*, assume the probability that there is a link between* i *and* j *is* $\mathbf{B}_{z_i, z_j}$*. That is,*

$$P(A_{ij} = 1 | H) = \mathtt{t}(\sum_{u=1}^{m} H_{iu} H_{ju}) \cdot \mathbf{B}_{z_i z_j}, \tag{1.6}$$

   where $\mathtt{t}(\cdot)$ *is a thresholding function, defined as* $\mathtt{t}(x) = 1$*, if* $x \geq 1$*;* $\mathtt{t}(x) = x$*, if* $x < 1$*. Conditional on* H*, all the links* $\{A_{ij} = A_{ji}, i < j\}$ *in* A *are formed independently and let* $A_{ii} = 0$ *to avoid self-loops.*

Remark 1.2. *Relationship to RIG: when* $\mathbf{B}$ *takes values equal to 1 in each cell, the model becomes the RIG.*

Remark 1.3. *Relationship to SBM: when* $\mathbf{M}$ *is dense, all pair of nodes hare at least 1 attribute. i.e., all entries in* HH^T *are all non-zero and the model is like the SBM.*

Remark 1.4. *Block structure: In both Step 1 and Step 2, we see that blocks play an important role and vertices from the same block will have the same tendency to have attributes and form links with each other.*

Remark 1.5. *Randomness in Step 2: In traditional RIGs, two vertices sharing some attributes will form a link. For each attribute, the vertices having that attribute will form a clique (complete graph) and this tends to form real large cliques, which are*

less common in real networks. The randomness in Step 2 of the RIGB helps break up some of the really large cliques and introduce extra sparsity.

High Transitivity of DC-RIGB

In RIGs the conditional probability of observing $i \sim j$ given that i and j share a common neighbor k i.e. $P(i \sim j | i \sim k, j \sim k)$ characterizes the transitivity. Deijfen and Kets (2009) studied the setting where RIGs have the non-degenerate transitivity and power-law degree distribution. Bloznelis (2013) proved the general relationships between transitivity and the degree distribution. Here we adjust the result in Deijfen and Kets (2009) and prove under a simple setting that the our model can have high transitivity.

Proposition 1.6 (Transitivity in RIGB)**.** *Let J_K be a $K \times K$ matrix with 1 in each cell. Let $A \in \{0,1\}^{n \times n}$ denote the random intersection graph sampled from RIGB with parameters $(\mathbf{M}, \mathbf{B}, z, y)$, where $\mathbf{M} = p_n J_K$, and $\mathbf{B} = q_n J_K$. Assume that m, p_n, q_n change with n, $p_n = o(1)$ and furthermore $m p_n^{3/2} = o(1)$, as $n \to \infty$. Then given three distinct vertices i, j and k, we have*

$$P(i \sim j | i \sim k, j \sim k) = \frac{q_n}{1 + m p_n}(1 + o(1)). \tag{1.7}$$

The proof of this proposition is presented in the Appendix Section A.

Remark 1.7. *If $m p_n$ is bounded and q_n is constant or bounded below, then RIGB has a non-degenerate transitivity ratio. That is, the sampled graph will have transitivity bounded away from zero (as $n, m \to \infty$) if the individual attribute sets of vertices are*

(stochastic) bounded and vertices form links with high chance once they share any attribute.

Remark 1.8. *If edges are generated independently as if from DC-SBM, this conditional probability in (1.7) will be just the probability of $i \sim j$. For a semi-sparse case where $m = n, p_n = \frac{\log n}{n}$ and q_n is a constant, the mean degree with be $nmp_n^2 q_n = O(\log^2 n)$. Under the RIGB, Proposition 1.6 states that $P(j \sim k | j \sim i, k \sim i) = O(\frac{1}{\log n})$ while under the circumstance of independent edges, $P(j \sim k | j \sim i, k \sim i) = O(mp_n^2 q_n) = O(\frac{\log^2 n}{n})$. Therefore, the composite model increases the transitivity significantly.*

Degree-corrected RIGB

Similar to the generalization from SBM to DC-SBM, we can introduce the degree correction parameters and arrive at the following naturally generalized model.

Definition 1.9 (Degree-corrected RIGB, DC-RIGB)**.** *Assume we have a vertex set V and an attribute set W and the parameters $(\mathbf{M}, \mathbf{B}, z, y)$ are same as those described in RIGB. We introduce additional degree parameters associated with vertices: $\theta_i > 0, i = 1, 2, \cdots, n$, and parameters associated with attributes: $\gamma_u > 0, u = 1, 2, \cdots, m$. We call A a random graph from the Degree-corrected RIGB (DC-RIGB) with parameters $(\mathbf{M}, \mathbf{B}, z, y, \theta, \gamma)$ if it generated from the following two steps:*

1. *Sample the bipartite graph H: vertex i has attribute a_u with probability $\theta_i \gamma_u \mathbf{M}_{z_i, y_u}$.*
2. *Sample the intersection graph A: vertex i, j form a link with probability $\mathtt{t}(\sum_{u=1}^{m} H_{iu} H_{uj}) \mathbf{B}_{z_i, z_j}$.*

Remark 1.10. *If θ_i is large, then vertex i will tend to have more attributes, and if γ_u is large, then attribute a_u will tend to be shared by more vertices. It is worth noting that θ_i's can help to adjust the degree distribution so that it can be heavy-tailed. In addition, γ_u's can be used to control the distribution of clique sizes.*

Remark 1.11. *DC-RIGB is a very comprehensive random network model. It can maintain many of the desired properties observed in empirical networks: global community structures (block structures); local structures (cliques introduced by the hidden layer variables); degree heterogeneity (the degree parameters); and sparsity (shrinking probabilities). Therefore, DC-RIGB will work as a better benchmark model for community detection algorithms.*

To enforce the identifiability among the parameters, we assume that the averages of the degree corrections within each block are equal to 1 as in (e.g., Zhang et al. (2014)). That is,

$$\frac{1}{n_k}\sum_{i\in V_k}\theta_i = 1; \quad \frac{1}{m_k}\sum_{u\in W_k}\gamma_u = 1, \quad k = 1, 2, \cdots, K, \tag{1.8}$$

where n_k is the size of V_k, and m_k is the size of W_k.

Matrix representation: To facilitate the spectral analysis in the following sections, here we introduce the matrix representation of the DC-RIGB with parameters $(\mathbf{M}, \mathbf{B}, z, y, \theta, \gamma)$. Let $Z \in \{0,1\}^{n\times K}$ and $Y \in \{0,1\}^{m\times K}$ be the membership matrices. For Z, the i-th row has a single 1 in the k-th entry and 0's in all other entries if and only if $z_i = k$; and the same goes for Y. Then let $\Theta = \mathrm{diag}(\theta)$, and $\Gamma = \mathrm{diag}(\gamma)$. Let $p_n \triangleq \max_{st}\mathbf{M}_{st}, q_n \triangleq \max_{st}\mathbf{B}_{st}$ be the sparsity parameters of the graph. To better

understand the role of sparsity in the model, we can pull out the sparsity parameters p_n, q_n and assume without loss of generality that $\max_{st} \mathbf{M}_{st} = 1, \max_{st} \mathbf{B}_{st} = 1$. The probability matrices for the two steps in DC-RIGB:

$$\mathcal{H} \triangleq \mathbf{E}(H|Z, Y, M) = p_n \Theta Z \mathbf{M} Y^T \Gamma, \text{ and} \tag{1.9}$$

$$\mathcal{A}_H \triangleq \mathbf{E}(A|H, \mathbf{B}, Z) = \mathtt{t}(HH^T) \cdot q_n Z \mathbf{B} Z^T. \tag{1.10}$$

Here $\mathtt{t}(\cdot)$ is defined same as in (1.6) and applied to the matrix element-wisely. The " $\cdot$ " matrix operator is the element-wise product or Hadamard product in literature. The RIGB is the special case of the DC-RIGB, where Θ, Γ are identities matrices.

1.4 Consistency of Spectral Clustering under the DC-RIGB

Population Analysis

The intuition of why Spectral Clustering might still perform well comes from the population analysis. By substituting the sampled bipartite graph H in Equation (1.10) with its population counterpart $\mathcal{H}$, we denote the resulted expression by $\mathcal{A}$, i.e. $\mathcal{A} \triangleq \mathcal{H}\mathcal{H}^T \cdot q_n Z \mathbf{B} Z^T$. Here the $\mathtt{t}(\cdot)$ is gone because the elements in $\mathcal{H}\mathcal{H}^T$ are all smaller than 1. Furthermore,

$$\begin{aligned} \mathcal{A} &= p_n^2 \Theta Z \mathbf{M} Y^T \Gamma \Gamma Y \mathbf{M}^T Z^T \Theta \cdot q_n Z \mathbf{B} Z^T \\ &= p_n^2 q_n \Theta [Z \mathbf{M} Y^T \Gamma \, \Gamma Y \mathbf{M}^T Z^T \cdot Z \mathbf{B} Z^T] \, \Theta \end{aligned}$$

$$= \Theta Z \widetilde{\mathbf{B}} Z^T \Theta, \tag{1.11}$$

where $\widetilde{\mathbf{B}} = p_n^2 q_n \mathbf{M} Y^T \Gamma^2 Y \mathbf{M}^T \cdot \mathbf{B} \in \mathbb{R}^{K \times K}$.

$\mathcal{A}$ is the population counterpart to the observed intersection graph A and is called population graph in this paper. Clearly, $\mathcal{A}$ has the same structure as that of DC-SBM under which Spectral Clustering can estimate the memberships consistently (Rohe et al. (2011); Jin et al. (2015); Lei and Rinaldo (2013); Qin and Rohe (2013)). We prove the consistency of Spectral Clustering under this new model. Here we present the eigen-structure of $\mathcal{A}$.

Proposition 1.12 (Eigen-structure of $\mathcal{A}$)**.** *Assume that* $\widetilde{\mathbf{B}}$ *in (1.11) is positive definite with rank* K *and then the populaiton* $\mathcal{A}$ *also has rank of* K*. Let* $v_1, v_2, \cdots v_K$ *be the leading* K *eigenvectors of* $\mathcal{A}$*. Let* $\mathcal{X} = [v_1|v_2|\cdots|v_K] \in \mathbb{R}^{n \times K}$ *and* $\mathcal{X}^*$ *be the row-normalized version of* $\mathcal{X}$*. Then* $\mathcal{X}$ *can be expressed below:*

$$\mathcal{X} = \Theta Z (Z^T \Theta^2 Z)^{-1/2} U, \quad \mathcal{X}^* = ZU, \tag{1.12}$$

where the orthogonal matrix $U \in \mathbb{R}^{K \times K}$ *are the eigenvectors of* $(Z^T \Theta^2 Z)^{\frac{1}{2}} \widetilde{\mathbf{B}} (Z^T \Theta^2 Z)^{\frac{1}{2}}$.

Remark 1.13. *Let* $\mathcal{X}_i$ *be* i*-th row in* $\mathcal{X}$*, where* $z_i = k$*, then we have that* $\mathcal{X}_i = \frac{\theta_i}{\sqrt{\sum_{j \in V_k} \theta_j^2}} U_k$*, where* U_k *is the* k*-th row of* U*. As noticed in Jin et al. (2015) and Qin and Rohe (2013), this reflects that (i) the rows in* $\mathcal{X}$ *corresponding to people from the same block will share the direction, though with different lengths; (ii) rows corresponding to different rows in* $\mathcal{X}$ *are orthogonal since rows in* U *are orthogonal. The row-normalized version of* $\mathcal{X}$*,* $\mathcal{X}^*$*, will be free of the degree parameters and easy*

for community detection task. This is the logic behind the Step 2 in Spectral Clustering described in Section 1.2.

Sample Analysis: Bound on $\|A - \mathcal{A}\|$

In reality, we only observe the sampled graph A and the ability of Spectral Clustering to recover the membership from sample A is closely tied to the convergence properties of its principal eigenvectors to those of $\mathcal{A}$. In literature, people rely heavily on the independence of the edge generation process so as to achieve a tight bound for $\|A-\mathcal{A}\|$ via concentration inequalities and then control the differences between the leading eigenvectors of A and $\mathcal{A}$ with the Davis-Kahan lemma.

However, in DC-RIGB, the edges are no longer independently generated and we cannot apply the concentration inequality directly. Luckily, we found the bipartite graph H will still concentrate around its population counterpart $\mathcal{H}$, which helps us get a tight bound on $A - \mathcal{A}$ indirectly. Based on triangle inequality,

$$\|A - \mathcal{A}\| \leq \|A - \mathcal{A}_H\| + \|A_H - \mathcal{A}\|, \tag{1.13}$$

we will achieve this goal by deriving bounds on $\|A-\mathcal{A}_H\|$ and $\|\mathcal{A}_H - \mathcal{A}\|$ respectively.

Let the expected bipartite graph be $\mathcal{H} = p_n \Gamma Z \mathbf{M} Y^T \widetilde{\Gamma}$ and $N = m + n$. For the rows of $\mathcal{H}$, let $\delta_1 \triangleq \min_{1\leq i\leq n} \sum_{j=1}^{m} \mathcal{H}_{ij}$ and $\Delta_1 \triangleq \max_{1\leq i\leq n} \sum_{j=1}^{m} \mathcal{H}_{ij}$ be the minimum and maximum expected row sums. For the columns of $\mathcal{H}$, let $\delta_2 \triangleq \min_{1\leq j\leq n} \sum_{i=1}^{n} \mathcal{H}_{ij}$ and $\Delta_2 \triangleq \max_{1\leq j\leq n} \sum_{i=1}^{n} \mathcal{H}_{ij}$ be the minimum and maximum expected column sums.

Theorem 1.14. *For a given $\epsilon > 0$, assume the following:*

(1) The minimum row sum has to grow at the following rate:

$$\delta_1 \geq 3\log(8n/\epsilon).$$

(2) The dense regions in $\mathcal{H}$ have to be controlled. Assume that

$$i) \max_{|I|=2\Delta_1} \sum_{k\in I} \mathcal{H}_{jk} \leq 1; \text{ and}$$
$$ii) \Delta_1 \max_{|I|=2\Delta_1} \sum_{j=1}^{n} \sum_{k\in I} \mathcal{H}_{jk}^2 \leq \frac{2}{9}\log(8n/\epsilon).$$

(3) The maximum expected degrees satisfy that $\min\{\Delta_1, \Delta_2\} \geq 16\log(8N/\epsilon)$.

Then with probability at least $1-\epsilon/2$, we have

$$\|\mathcal{A}_H - \mathcal{A}\| \leq 6K^2 q_n \sqrt{\Delta_1\Delta_2}\sqrt{\max\{\Delta_1,\Delta_2\}\,\log(8N/\epsilon)}. \tag{1.14}$$

Additionally, if the off-diagonal terms in $\mathbf{B}$ are the same, the K^2 term on the right hand side can be removed.

Remark 1.15. *Assumptions (1) and (3) are the sparsity conditions on the population bipartite graph $\mathcal{H}$ required for the concentration bound on $\|H-\mathcal{H}\|$. This type of conditions is commonly required in literature to achieve the consistency of spectral clustering.*

Remark 1.16. *Assumption (2) is unique to our problem and it is required to control truncation effects; i.e., the differences between $\mathtt{t}(HH^T)$ and HH^T. Essentially, the*

conditions require that population biepartite graph H cannot be too dense. One sufficient condition satisfying Assumption (2) is that $p_{\max} \triangleq \max_{ij} \mathcal{H}_{ij} \leq (\frac{\log(8n/\epsilon)}{9nm^2})^{\frac{1}{4}}$. This is easy to check since $\Delta_1 \leq mp_{\max}$.

Theorem 1.17. *Assume that for a given $\epsilon > 0$, Assumptions (1) - (3) in Theorem 1.14 hold. Additionally, we assume the intersection graph should be sufficiently dense and q_n satisfies*

(4) $3\Delta_1\Delta_2 q_n > \frac{4}{9}\log(8n/\epsilon)$.

Then for sufficiently large n, we have with probability at least $1 - \epsilon/2$,

$$\|A - \mathcal{A}_H\| \leq \sqrt{12\Delta_1\Delta_2 q_n \log(8N/\epsilon)}. \tag{1.15}$$

Therefore, combining with Theorem 1.14, we have for sufficiently large n, with probability at least $1 - \epsilon$,

$$\begin{aligned} \|A - \mathcal{A}\| \quad &\leq \quad 6K^2 q_n \sqrt{\Delta_1\Delta_2}\sqrt{\max\{\Delta_1, \Delta_2\}\, \log(8N/\epsilon)} \\ &\quad + \sqrt{12\Delta_1\Delta_2 q_n \log(8N/\epsilon)}. \end{aligned} \tag{1.16}$$

Additionally, if the off-diagonal terms in $\mathbf{B}$ are the same, the K^2 term on the right hand side can be removed.

Mis-clustering Rate

With the bound on the $\|A - \mathcal{A}\|$, Davis-Kahan lemma controls the difference between the corresponding leading eigenvectors of A and $\mathcal{A}$. The consistency of spectral

clustering under the **DC-RIGB** follows from this with a common argument (e.g., Qin and Rohe (2013), Lei and Rinaldo (2013)).

Theorem 1.18 (Convergence of sample eigenvectors)**.** *Let A be the sample graph from DC-RIGB with parameters $(z, y, \mathbf{M}, \mathbf{B}, \theta, \gamma)$ and let $\mathcal{A}$ be the population graph defined as (1.11). Assume that*

(5) $\widetilde{\mathbf{B}} = \mathbf{M}Y^T\Gamma^2 Y\mathbf{M}^T \cdot \mathbf{B} \in \mathbb{R}^{K\times K}$ *is positive definite.*

Then, the rank of $\mathcal{A}$ is K. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ be the only K non-zero eigenvalues of $\mathcal{A}$. Let $X, \mathcal{X} \in R^{n\times K}$ be the matrices with the leading K eigenvectors of A and $\mathcal{A}$ as the columns. Let $X^*, \mathcal{X}^* \in R^{n\times K}$ be the row-normalized version of $X, \mathcal{X}$ as defined in the spectral clustering in Section 1.2. Let $r \triangleq \min_i \|\mathcal{X}_i\|_2$ be the minimum length among rows in $\mathcal{X}$. Then there exists a $K \times K$ orthonormal matrix $\mathcal{O}$ such that

$$\|X - \mathcal{X}\mathcal{O}\|_F \leq \frac{2\sqrt{2K}}{\lambda_K}\|A - \mathcal{A}\|, \quad \|X^* - \mathcal{X}^*\mathcal{O}\|_F \leq \frac{4\sqrt{2K}}{r\lambda_K}\|A - \mathcal{A}\|. \tag{1.17}$$

Recall $\mathcal{X} = \Theta Z(Z^T\Theta^2 Z)^{-1/2}U$, and $\mathcal{X}^* = ZU$ as in (1.12). Step 3 in Spectral Clustering in Section 1.2 applies k-means to $X^* \in \mathbb{R}^{n\times K}$ and get the membership for all the rows. For $i \in [n]$, let row vector $C_i \in R^{1\times K}$ be center of the cluster to which i-th row of X^* is assigned. Correspondingly, $\mathcal{X}^*$ has K distinct rows, and if k-means algorithm is applied to $\mathcal{X}^*$, the center of the cluster to which i-row of $\mathcal{X}^*$ is assigned is itself Z_iU. In essence, we consider node i correctly clustered if C_i is closer to $Z_iU\mathcal{O}$ than it is to any other $Z_jU\mathcal{O}$ for all $Z_j \neq Z_i$.

Definition 1.19 (set of mis-clustered nodes)**.** *Vertex i corresponding to center C_i is said to be correctly clustered if there is no $j \in [n]$, such that $Z_j U\mathcal{O}$ is closer to C_i than $Z_i U\mathcal{O}, Z_j \neq Z_i$. Based on this intuition, we define the following* ***mis-clustered set*** *of nodes:*

$$\mathcal{M} = \{i : \exists j \neq i,\ s.t.\, \|C_i - Z_j U\mathcal{O}\| < \|C_i - Z_i U\mathcal{O}\|\}.$$

Theorem 1.20 (Main Theorem)**.** *Let A be the adjacency matrix for a random graph sampled from the* ***DC-RIGB*** *with parameters $(\mathbf{M}, \mathbf{B}, z, y, \theta, \gamma)$. Let $\mathcal{A}$ be the population graph defined as in (1.11). For a given ϵ, assume that Assumptions (1) - (5) all hold. Then $\mathcal{A}$ is positive definite with rank K and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ be the only K non-zero eigenvalues. Let $r \triangleq \min_{1\leq i\leq n} \|\mathcal{X}_i\| = \min_k \frac{\theta_i}{\sqrt{\sum_{i\in V_k} \theta_i^2}}$. Let $\mathcal{M}$ be the set of mis-clustered nodes specified in Definition 3.4. Then when n is large enough, the following holds with probability at least $1 - \epsilon$,*

$$\frac{|\mathcal{M}|}{n} \leq c_0 \frac{K}{nr^2} \frac{\Delta_1 \Delta_2 q_n \log(8N/\epsilon)}{\lambda_K^2} \left(K^4 q_n \max\{\Delta_1, \Delta_2\} + \mathbf{1}\right), \tag{1.18}$$

where c_0 is a constant. Additionally, assume the off-diagonal terms in $\mathbf{B}$ are the same, the K^4 term on the right hand side can be removed.

Remark 1.21. *The mis-clustering ratio has two components: the second part is similar to the results under SBM or DC-SBM in literature (e.g., Lei and Rinaldo (2013)). However, according to Proposition 1.6, to achieve a high transitivity which is desired, q_n is expected to take constant values or shrinks very slowly. Combined with Assumption (1), which implies that $\min\{\Delta_1, \Delta_2\} p_n = \omega(1)$, we have that the first*

term will be the dominant part in determining the convergence rate. This indicates that the mis-clustering rate by Spectral Clustering will still converge to 0, but at slower rate.

Corollary 1.22 (Result in special RIGB)**.** *Let A be the adjacency matrix for a random graph sampled from the* **RIGB** *with the parameters $(\mathbf{M}, \mathbf{B}, z, y)$. Here we assume that*

i) $\mathbf{B}, \mathbf{M}$ *have the following structures:*

$$\mathbf{M} = p_n \begin{bmatrix} 1 & \eta & \cdots & \eta \\ \eta & 1 & \cdots & \eta \\ \vdots & \vdots & \ddots & \vdots \\ \eta & \eta & \cdots & 1 \end{bmatrix} \text{ and } \mathbf{B} = q_n \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \tag{1.19}$$

where $0 < \eta < 1$, and $0 < \rho < 1$ are constants.

ii) z, y *are membership vectors and assume that the blocks of vertices and of attributes are equally sized; i.e., $n_1 = n_2 = \cdots = n_K = \frac{n}{K}, m_1 = m_2 = \cdots = m_K = \frac{m}{K}$.*

Then we have $\Delta_1 = mp_n(\frac{1}{K} + \frac{K-1}{K}\eta)$, $\Delta_2 = np_n(\frac{1}{K} + \frac{K-1}{K}\eta)$, and $\lambda_K = \frac{mnp_n^2 q_n}{K}(a-b)$, where $a = \frac{1}{K} + \frac{K-1}{K}\eta^2, b = \rho(\frac{2}{K}\eta + \frac{K-2}{K}\eta^2)$. Futhermore, $nr^2 = K$ and K^4 in (1.18) can be removed. Then with probability at least $1 - \epsilon$, for sufficiently large n we have

$$\frac{|\mathcal{M}|}{n} \leq c_1 \left(\frac{K^2 \log(8N/\epsilon)}{\min\{m, n\} p_n} + \frac{K^2 \log(8N/\epsilon)}{mnp_n^2 q_n} \right), \tag{1.20}$$

where c_1 is a constant depend on η, ρ, and c_0 in (1.18).

Remark 1.23. *Semi-sparse case: $m = n$, $p_n = \frac{\log^{1+\nu} n}{n}, \nu > 0$, q_n is a constant and K is fixed. Then Assumptions (1)-(5) required by Theorem 1.20 will hold. The expected mean degree of the graph is of order $log^{2(1+\nu)}n$ and transitivity is of order $O(\frac{1}{\log^{1+\nu} n})$ from Corollary 1.22. The mis-clustering rate of SC is $O_p(\frac{1}{\log^{\nu} n})$ since the second term is dominated by the first term. On the contrast, if A is sampled from $\mathcal{A}$ as in the* **SBM**, *the mis-clustering rate of SC is $O_p(\frac{1}{\log^{(1+2\nu)} n})$, as the second term in (1.20).*

1.5 Simulation

In this section, we conducted extensive simulations to better understand the properties of RIGB and the performance of community detection algorithms under RIGB. As a complementary, we also investigate the effects of degree heterogeneity.

Setting: The simulation setting is similar to that in the Corollary 1.22. Let $m = n$, $K = 5$, $p_n = \frac{\log^{1.5} n}{n}$, and $q_n = 0.5$. $\mathbf{M}$ and $\mathbf{B}$ have the same format as in (1.19) with $\eta = 0.2, \rho = 1$. In the simulation, network size n takes values from $\{1000, 2000, 4000, 6000, \cdots, 28000, 32000\}$ and the sparsity p_n will change accordingly. So, the expected mean degree of graph is $O(\log^3 n)$.

For degree heterogeneity, we always set γ_u's to be 1. We sample θ_i's from Exponential distribution with rate equal to 1 and then add 1 to the θ_i before normalizing to satisfy the identifiable conditions in (1.8). For each n, we sample 100 networks from each model and all the analysis are conducted on the largest connect components.

Transitivity and number of triangles under RIGBs

In this simulation, we further investigate the some statistics such as transitivity, number of triangles of RIGB, DC-RIGB, SBM and DC-SBM. This exploration extends the theoretical insight in Proposition 1.6 where RIGB is simplified without blocks and degree heterogeneity to a more general setting. All these quantities are summarized in Figure 1.1.

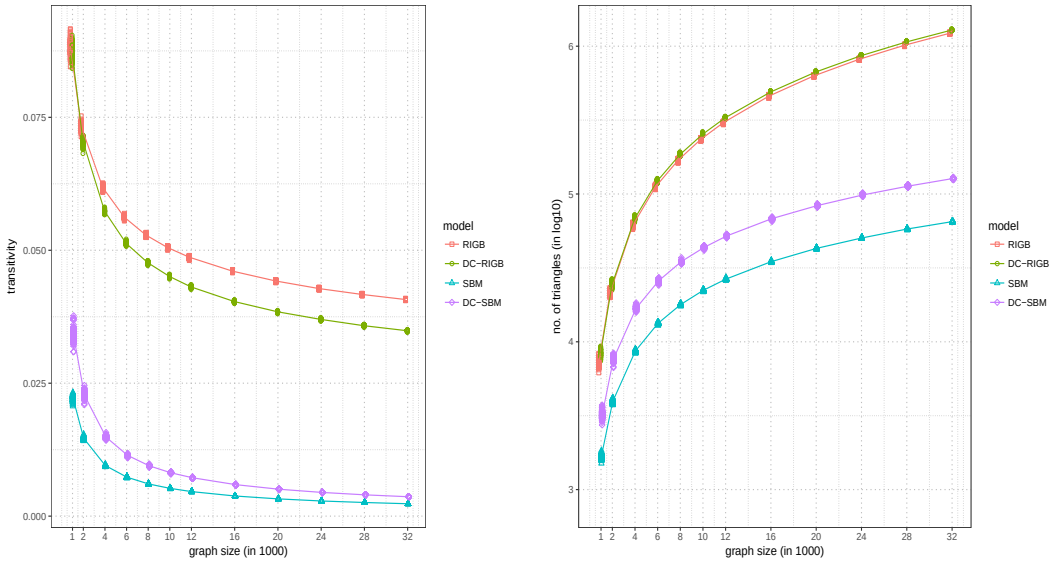


Figure 1.1: Comparison of transitivity and number of triangles for networks sampled from RIGB, DC-RIGB, SBM and DC-SBM. The left panel, shows that the transitivity ratios are higher under RIGB and DC-RIGB compared to SBM and DC-SBM. The right panel shows that RIGB And DC-RIGB have more triangles.

Mis-clustering rate for SC under RIGB

Theorem 1.20 and its corollary offer us an upper bound on the mis-clustering rate when applying SC to estimate the community under DC-RIGB and RIGB. However, we

cannot use upper bounds to compare the difficulty of two problems. This experiment affirms that the convergence rate of SC is slower under RIGB than that under SBM and this is not due to the potential limitation of our proof techniques.

For RIGB and SBM, the upper bound in Corollary 1.22 are of the order of $\frac{1}{\log^{0.5} n}$ and $\frac{1}{\log^2 n}$. The left plot in Figure 1.2 shows the actual mis-clustering rates are much lower than that in the upper bounds, which are shown as black dots and black line (we choose the constants to match the mis-clustering rates at $n = 6000$). The mis-clustering rates for SC under RIGB with and without degree corrections are higher those under SBM. The left panel in Figure 1.2 indicates that when the graph sizes are small, RIGB are easier for SC to recover memberships; this applies to DC-RIGB and DC-SBM as well. The right panel of Figure 1.2 shows that degree heterogeneity makes it harder for SC to recover community and it has bigger effect on SBM.

Performances of various community detection algorithms under RIGB.

The following experiments compare the performance of various existing community detection algorithms (including SC) under on RIGB and SBM. The methods[1] considered in the comparisons include:

– Two likelihood methods: Bickel and Chen's Profile Likelihood (BCPL) in Bickel and Chen (2009) Zhao et al. (2012), Pseudo Likelihood (APL) in Amini et al. (2013),;

[1]Ji and Jin applied several popular clustering methods including their own method to discover the community structures in statistician co-authorship and citation networks (Ji and Jin (2014)). We adapted some codes used in that paper.

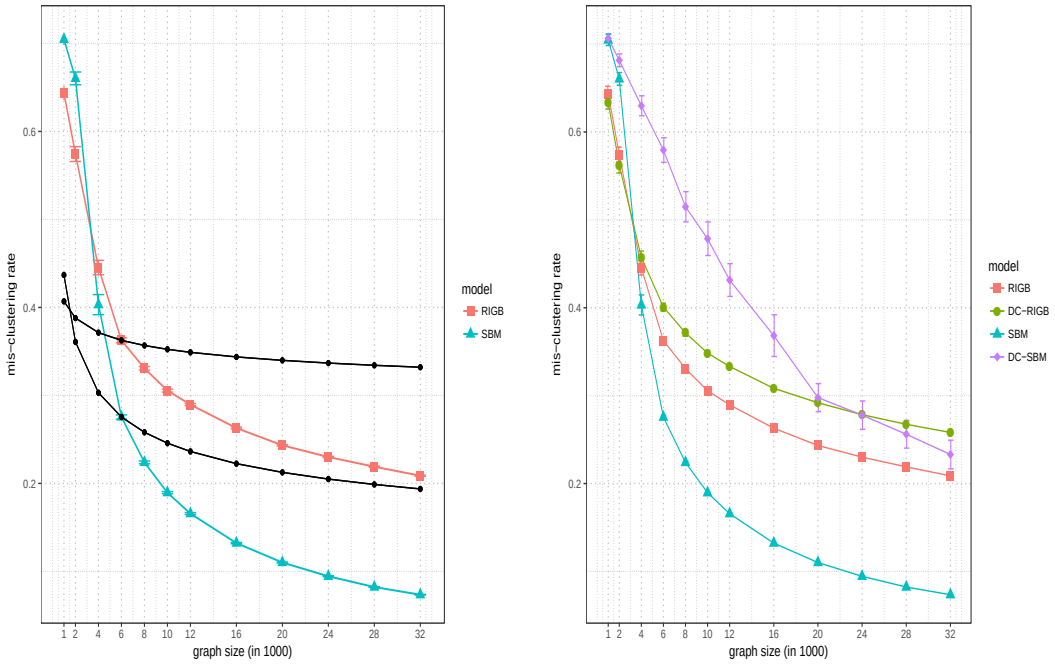


Figure 1.2: Mis-clustering rates of SC under RIGB and SBM with and without degree heterogeneity. The means and 95% confidence intervals are calculated based 100 sampled networks. On the left there is no degree heterogeneity, the mis-clustering rate is slower under RIGB; the black line are the upper bounds by Corollary 1.22. On the right, the mis-clustering rates are slower for the models with degree heterogeneity (e.g., exponential distribution tail) compared with models without degree heterogeneity.

– Four spectral methods: Newman's Spectral Clustering (NSC) in Newman (2006), Jin's SCORE in Jin (2012), Spectral Clustering on adjacency matrix (SCA) in Lei and Rinaldo (2013), Spectral Clustering on Laplacian (SCL) in Ng et al. (2001), Von Luxburg (2007); and

– One spectral methods with regularization: Regularized Spectral clustering (RSC) in Qin and Rohe (2013).

The two box plots in Figure 1.3 clearly indicate that it is harder to recover memberships from networks sampled DC-RIGB (with edge dependence) than net-

works sampled from DC-SBM (without edge dependence). RSC and BCPL works comparatively better in both cases.

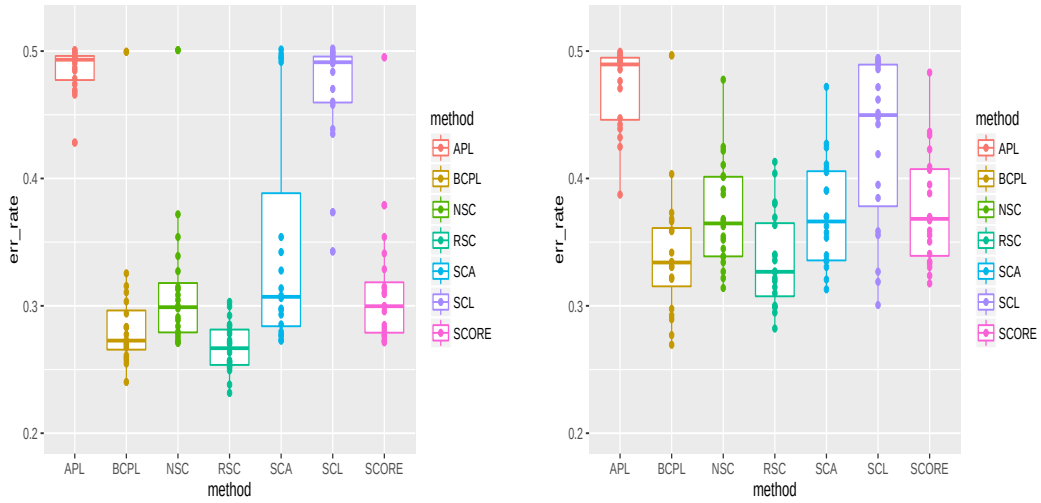


Figure 1.3: Mis-clustering rates of Different Clustering algorithms under networks sampled from DC-SBM (left) vs. networks from DC-RIGB (right). The mis-clustering rates calculated are from 20 sampled networks from each model with 2000 nodes.

1.6 Conclusion

In this study, we propose a better benchmark model called Random Intersection Graph with Blocks (RIGB) for community detection. Graphs sampled from this model have global community structures, heterogeneous degrees, and rich local structures characterized by high number of triangles and clustering coefficients. This is a valuable addition to the network modeling literature.

Also, we address a very fundamental problem about Spectral Clustering: the effects of diverse local structures or dependence among edges in networks could greatly

change the spectral properties and thus cast doubt on whether Spectral Clustering will still work consistently. In this paper, we proved that under some conditions, Spectral Clustering can still estimate the memberships consistently but the mis-clustering rate converges to 0 at a slower rate under RIGBs compared to that under SBMs.

In the simulation, we confirm the high transitivity of RIGB and affirm slower convergence rate of Spectral Clustering under the RIGB. Furthermore, we show it is a harder problem to estimate the community under RIGB than under SBM by showing that the mis-clustering rates are higher for many other existing algorithms.

Our result is the first in literature to combine literature from RIG and SBM. There can a few other directions to advance our results. 1. It would interesting to extend our conclusion about adjacency matrix to that of graph Laplacians. This is meaningful since normalized Spectral Clustering is more commonly used in practice and gives better clustering accuracy (Von Luxburg (2007), Sarkar et al. (2015)).

2. It will be a vital question to estimate the parameters in our model and reconstruct our model. Currently, we focus solely on the memberships of vertices and our next question is to recover the hidden bipartite graph from A to some degree.

Chapter 2

Eigenspace MLE to Estimating the Communities under Stochastic Block Model

Abstract

Many community detection algorithms, such as modularity maximization and Spectral Clustering, use the eigenvectors of certain matrix representation of a graph. Recent development in Random Dot Product Graph (RDPG) showed that the scaled eigenvectors of RDPG converge to a mixture of of Gaussian distributions, which implies that scaled eigenvectors from Stochastic Block Models (SBM) converge to a finite Gaussian distribution. In this paper, we model the scaled eigenvectors of the adjacency matrix using a Gaussian Mixture Model (GMM) with covariance structure given by a Central Limit Theorem under the RDPG. Furthermore we propose an approximate Expectation-Maximization algorithms to estimate parameters and the hidden memberships under the SBMs and under its variant, Degree-corrected SBM (DC-SBM). In simulation, our approach to incorporates the limiting covariance structure in the fitting process shows advantages over fitting Gaussian mixture models blindly. For the algorithm developed under DC-SBM, it works better than the state-of-art method, and potentially be come alternative way to handle degree heterogeneity in community detection.

2.1 Introduction

Networks and graphs represent the complex relationships among people or objects. For example, Facebook shows the friendships among people and co-authorship networks display the collaboration relationship among researchers. There has been growing interest in analyzing the network data to under the underly mechenism behind these networks.

Identifying the community structure or clusters of tightly connected nodes, which is often called community detection, has gained enormous attention from various research fields. People have proposed many different algorithms to tackle this problem. Many of existing algorithms such Spectral Clustering (Von Luxburg (2007); Ng et al. (2002)) and modularity maximization (Newman (2006)) among others are gaining popularity in utilizing the eigenvectors of certain graph matrix to partitioning the graph due to their computational tractability. This paper, based on recent development in literature, will provide another intelligent way to utilize the eigenvectors of a graph matrix to conduct community detection.

Recently, Athreya et al. (2016) proved a central limit theorem (CLT) for the weighted leading eigenvectors (spectral embedding) of the adjacency matrix of random dot product graph (RDPG). As the corollary of the result, the spectral embedding of adjacency matrix sampled under the stochastic blockmodel (SBM) converges to a finite Gaussian mixture distribution. A more recent paper (Tang and Priebe (2016)) extended the CLT for sparse RDPG and also proved the CLT results hold for the leading eigenvectors from the normalized Laplacian.

In terms of the implication of the CLT on statistical inference, the authors

demonstrated through simulation that under the SBM, there are potential advantages of applying Gaussian mixture models to recover the memberships over k-means algorithm. As know that k-means algorithm is equivalent to GMM with an identity covariance matrix, which is not the case for the scaled eigenvectors of SBM. This might explain why Fitting GMM which allows for elliptical covariance structure can improve inference accuracy.

In fact, the CLT results in these two papers also detail the expression of covariance matrix of the limiting distribution. This can save us from estimating the covariance matrices, which can be expensive, and offer another potential improvement in inferences. In this study, we model the spectral embedding of the adjacency matrix (see Definition) from SBM using a Gaussian mixture model (GMM) with the covariance structures specified by the CLT. We propose an approximate EM algorithm called eigenspace MLE (eMLE) to estimate the parameters and the hidden community memberships. Simulation shows that our algorithm can work better than fitting GMM blindly without using the limiting covariance structure.

Furthermore, we propose a different algorithm to handle networks sampled from the Degree-corrected SBM (DC-SBM, Karrer and Newman (2011)) where nodes from same community have different expected degrees. Currently, to alleviate the effects of degree heterogeneity on community memberships, people resort to row normalization which projects the rows in the leading eigenvectors onto unit sphere (e.g. Ng et al. (2002), Jin et al. (2015), Qin and Rohe (2013)). However, row normalization will have the less favorable effects of inflating the influence of rows with small lengths while down weighting the points with big lengths, which corresponding to high-degree

nodes and should contain more information. Our algorithm provides a different and more powerful procedure to estimate the community memberships.

The paper is organized as follows. Section 2.2 includes the preliminaries about concept of Stochastic Block Model (SBM), Random Dot Product Graph (RDPG) and others. Section 2.3 proposes an Expectation-Maximization (EM) algorithm to estimate the memberships from spectral embedding under SBM. Section 2.4 derives a new algorithms called variational EM to estimate memberships under the DC-SBM. Section 2.5 provides another EM algorithm to handle degree heterogeneity under DC-SBM. Section 2.6 gives a different objective function other than k-means to recover community from DC-SBM. Section 2.7 gives some simulation results to demonstrate the advantages of our proposed algorithms. Finally, Section 2.8 discuss the successes and limitations of the new algorithms. The appendix B contains all the technical derivations and proofs.

Some notations: The vectors in the paper are column vectors unless stated otherwise. $[n]$ where n is an integer, used as the shorthand for the set $\{1, 2, ..., n\}$. For a vector $\theta \in \mathbb{R}^n$, $\text{diag}(\theta)$ denotes the $n \times n$ diagonal matrix with elements from θ located along the diagonal. For a matrix M, M_i denote the i-th row of M and $M_{\cdot j}$ denote the j-th column of M. For two real numbers a and b, $a \wedge b$ stands for the minimum of a and b. $Z \in \{0, 1\}^{n \times K}$ is called membership matrix, where each row has one and only one non-zero entry and we may assume there is columns all 0's. We call $z \in \mathbb{R}^n$, where $z_i \in \{1, 2, ..., K\}$ denote the position of nonzero entry in i-th row in Z, the membership vector.

2.2 Preliminaries

Random Dot Product Graph states that each vertex is associated with a hidden position characterized by a d-dimension vector and conditioned on the hidden positions, the presence or absence of edges is independent. The presence probability between two nodes is determined by the inner product of their latent positions. Mathematically, we have the following definition.

Definition 2.1 (Random Dot Product Graph)**.** *Given a distribution F on a set $\mathcal{X} \subset \mathbb{R}^d$ satisfying $x^T y \in [0,1]$ for all $x, y \in \mathcal{X}$ and a sparsity factor $\alpha_n \leq 1$, we say (X, A), where we only observe A and X is hidden, is sampled from Random Dot Product Graph with distribution F with sparsity factor α_n, denoted as RDPG(F) with sparsity factor α_n, if*

(i) rows in $X \in \mathbb{R}^{n \times d}$, X_i's, are i.i.d. samples from F; and

(ii) for $i < j$, $A_{ij}|X \sim$ Bernoulli$(\alpha_n X_i^T X_j)$, and conditioning on X, all the A_{ij}'s are independent.

In matrix form, we have $P \triangleq \mathbf{E}(A|X) = \alpha_n X X^T$, which has a low rank structure. Since A is a noisy version of P, a natural way to estimate the pattern positions is based the spectral decomposition of A. Here is the spectral embedding defined in Athreya et al. (2016) .

Definition 2.2 (Spectral embedding of the adjacency matrix A)**.** *Given an adjacency matrix A, let $\widehat{\lambda}_1, \widehat{\lambda}_2, \cdots, \widehat{\lambda}_n$ be the eigenvalues of A satisfying $|\widehat{\lambda}_1| \geq |\widehat{\lambda}_2| \cdots \geq |\widehat{\lambda}_n|$ and $\widehat{v}_1, \widehat{v}_2, \cdots, \widehat{v}_n \in \mathbb{R}^n$ be the eigenvectors. Thus the full eigen-decomposition of*

A is $A = \sum_{i=1}^{n} \widehat{\lambda}_i \widehat{v}_i \widehat{v}_i^T$. For an integer $d > 0$, let $\widehat{V} \triangleq [\widehat{v}_1|\widehat{v}_2|\cdots|\widehat{v}_d]$ and $\widehat{S} \triangleq \text{diag}(|\widehat{\lambda}_1|, |\widehat{\lambda}_2|, \cdots, |\widehat{\lambda}_d|)$. We call

$$\widehat{X} \triangleq \widehat{V}\widehat{S}^{\frac{1}{2}}$$

the d-dimension spectral embedding of A.

Athreya et al. (2016) proves a Central Limit Theorem for the spectral embedding of adjacency matrix under RDPG. Later, Tang and Priebe (2016) extends the result by proving a CLT under the sparse RDPG and the CLT for the spectral embeddings of another type of graph matrix, Graph Laplacian matrix.

Lemma 2.3 (CLT from Athreya et al. (2016) and Tang and Priebe (2016))**.** *Let $(X, A) \sim RDPG(F)$ with sparsity α_n, where F is a distribution for points in $\mathbb{R}^d$, and let $\widehat{X}$ be spectral embedding A as in previous slide. Under some regularity conditions, we have there exists a sequence of orthogonal matrices W_n such that for each component i and $z \in \mathbb{R}^d$, we have*

$$Pr\{\sqrt{n}(W_n\widehat{X}_i - \sqrt{\alpha_n}X_i) \leq t\} \to \int \Phi(t, \Sigma(x))dF(x). \tag{2.1}$$

Here $\Phi(t, \Sigma)$ is the cumulative distribution function for multivariate normal, with mean zero and covariance matrix Σ, evaluated at t and

$$\mathbf{\Sigma}(x) = \begin{cases} \Delta^{-1}E[X_1X_1^T(x^TX_1)]\Delta^{-1}, & \text{if } \alpha_n = o(1) \text{ and } n\alpha_n = \omega(\log^4 n); \\ \Delta^{-1}E[X_1X_1^T((x^TX_1) - (x^TX_1)^2)]\Delta^{-1}, & \text{if } \alpha_n = 1, \end{cases} \tag{2.2}$$

where $\Delta = E(X_1 X_1^T)$ *with* $X_1 \sim F$ *and* $F = \sum_k \pi_k \delta_{\boldsymbol{\mu}_k}$. *Additionally, for any fixed index set* $(i_1, \cdots, i_k)$, *the variables* $\widehat{X}_{i_1}, \cdots, \widehat{X}_{i_k}$ *are asymptotically independent.*

The covariance matrices have different forms for sparse ($\alpha_n = o(1)$), and dense ($\alpha_n = 1$) RDPGs . However, if we assume that X_i are sampled from scaled F distribution, samples multiplied by $\sqrt{\alpha_n}$, then based on intermediate steps in the proof, we can consider the sparse version of covariance matrix is obtained by omitting the lower order term in the dense version. Moreover, our simulations show that keeping the lower order usually provides better accuracy for finite graph. Therefore, To universalize the notation for both sparse and dense graphs, we use F_n to denote the F distribution scaled by $\sqrt{\alpha_n}$, and use the dense form of covariance structure for both dense and sparse RDPGs i.e.

$$\Sigma(x) = E[X_1 X_1^T((x^T X_1) - \sqrt{\alpha_n}(x^T X_1)^2)]\Delta^{-1}, \text{ where } X_1 \sim F_n. \tag{2.3}$$

It is equivalent to

$$\Sigma(x) = E[X_1 X_1^T((x^T X_1) - \sqrt{\alpha_n}(x^T X_1)^2)]\Delta^{-1}, \text{ where } X_1 \sim F. \tag{2.4}$$

In community detection literature, Stochastic Block Model is a popular random network model, commonly used as building blocks to model networks with community structures or used to generate benchmark networks to evaluate community detection algorithms.

Definition 2.4 (Stochastic Block Model (SBM))**.** *Given* $\pi \in [0,1]^K$ *with* $\sum_{k=1}^K \pi_k = 1$

and $\mathbf{B} \in [0,1]^{K\times K}$*, we call* (Z, A) *is a network sampled from SBM with parameters* $(\pi, \mathbf{B})$ *(Holland et al. (1983)), if*

1. $Z_i \in \{0,1\}^K \overset{i.i.d.}{\sim}$ *multinomial*(π)*, for* $i \in [n]$*; and*

2. *for nodes* i, j*, they form a link with probability* $Z_i^T \mathbf{B} Z_j$*. That is,*

$$P(A_{ij} = 1 | Z_i, Z_j) = Z_i^T \mathbf{B} Z_j. \tag{2.5}$$

Additionally, given $Z = [Z_1 | Z_2 | \cdots | Z_n]^T$*,* A_{ij}*'s are independent.*

Additionally, $\mathbf{E}(A|Z) = Z\mathbf{B}Z^T$. When $\mathbf{B}$ is positive semidefinite with rank d, then there exists $\boldsymbol{\mu} \in \mathbb{R}^{K\times d}$ such that $\mathbf{B} = \boldsymbol{\mu}\boldsymbol{\mu}^T$. It is easy to verify that SBM with $(\pi, \mathbf{B})$ is RDPG(F), where

$$F = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\mu}_k}, \tag{2.6}$$

and δ is the Dirac measure. That is F is a combination of K points masses.

A popular extension of SBM is Degree-corrected SBM (DC-SBM) (Karrer and Newman (2011)), which introduces a degree parameter $\theta_i > 0$ for each node i. Under this model, for a pair of nodes i and j, the probability there is a link between them is $\theta_i\theta_j Z_i^T \mathbf{B} Z_j$. In matrix form, we have $\mathbf{E}(A|\Theta, Z) = \Theta Z B Z^T \Theta$, where $\Theta = \text{diag}(\theta)$, i.e., the diagonal matrix with θ located along its diagnoal. We denote (Θ, Z, A) is sampled from DC-SBM with parameters$(\pi, f, \mathbf{B})$.

Assume that $\theta_i \overset{i.i.d.}{\sim} f$, where f is a distribution on $(0, \infty)$. It is easy to verify that DC-SBM is RDPG(F), where F is a distribution on $\mathbb{R}^d$ with the support of

$\{a\,\boldsymbol{\mu}_k : k = 1, 2, \cdots, K, a > 0\}$ defined as follows:

$$F(x) = \begin{cases} f(\theta)\pi_k, & \text{if } x = \theta\boldsymbol{\mu}_k; \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

Remark 2.5. *To enforce identifiability about the scaling, we assume that the distribution for the degree parameters θ_i has mean equal to 1. That is,*

$$\int_0^\infty t\, f(t)dt = 1. \tag{2.8}$$

Remark 2.6. *F's in (2.6) and (2.7) are not unique since the $\boldsymbol{\mu}$ in the the decomposition that $\mathbf{B} = \boldsymbol{\mu}\boldsymbol{\mu}^T$ is not unique. Different $\boldsymbol{\mu}$'s differ up to a $d \times d$ rotation and we ensure the uniqueness by requiring that $\Delta \triangleq \boldsymbol{\mu}^T \mathrm{diag}(\pi)\boldsymbol{\mu}$, which the second moment of F, be diagonal. Through this paper, $\boldsymbol{\mu}$ is assumed to this unique version.*

To emphasize the parameters $\boldsymbol{\mu}$, when we talk about SBM, we call (Z, A) or (z, A) is sampled from the SBM with parameters $(\pi, \boldsymbol{\mu})$ instead of $(\pi, \mathbf{B})$ or from the DD-SBM with parameters $(\pi, f, \boldsymbol{\mu})$ instead of $(\pi, f, \mathbf{B})$.

2.3 eMLE algorithm under SBM

Corollary 2.7 (CLT under SBM, Athreya et al. (2016) and Tang and Priebe (2016))**.** *Assume that (Z, A) is sampled from SBM with parameters $(\alpha_n, \pi, \boldsymbol{\mu})$. Let $X = Z\boldsymbol{\mu} \in \mathbb{R}^{n\times d}$ and $\widehat{X}$ be the d-dimensional embedding of A as in Definition 2.2, then there*

exists some $d \times d$ orthonormal matrix W_n, such that for any index i,

$$\sqrt{n}(W_n\widehat{X}_i - \sqrt{\alpha_n}X_i)|_{X_i=\boldsymbol{\mu}_k} \xrightarrow{d} N(0, \Sigma(\boldsymbol{\mu}_k)). \tag{2.9}$$

Here,

$$\boldsymbol{\Sigma}(x) = \Delta^{-1}E[X_1X_1^T((x^TX_1) - \sqrt{\alpha_n}(x^TX_1)^2)]\Delta^{-1}, \tag{2.10}$$

where $\Delta = E(X_1X_1^T)$ with $X_1 \sim F$ and $F = \sum_k \pi_k\delta_{\boldsymbol{\mu}_k}$. Additionally, for any fixed index set $(i_1, \cdots, i_k)$, the variables $\widehat{X}_{i_1}, \cdots, \widehat{X}_{i_k}$ are asymptotically independent.

Remark 2.8. *For sparse graphs where $\alpha_n = o(1)$, we see that α_n in Equation (2.9) reduces the length of the centers $\sqrt{\alpha_n}X_i$, but it has minor effects on the limiting covariance matrix. This indicates that as the graph become sparser, it will become more difficult to distinguish points from different normal distributions, which coincides the similar results in SBM literature.*

For $F = \sum_k \pi_k\delta_{\boldsymbol{\mu}_k}$, after assuming that $\alpha_n = o(1)$ and ignoring the lower order term, we have

$$\Delta = \sum_j \pi_j\boldsymbol{\mu}_j\boldsymbol{\mu}_j^T \text{ and } E[X_1X_1^T(\boldsymbol{\mu}_k^TX_1)] = \sum_{j=1}^{K} \pi_j(\boldsymbol{\mu}_k^T\boldsymbol{\mu}_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^T. \tag{2.11}$$

And thus the covariance matrix $\Sigma(\boldsymbol{\mu}_k)$ is:

$$\Sigma(\boldsymbol{\mu}_k) = [\boldsymbol{\mu}^T\text{diag}(\pi)\boldsymbol{\mu}]^{-1}\,(\boldsymbol{\mu}^T\text{diag}(\pi)\text{diag}(\boldsymbol{\mu}\boldsymbol{\mu}_k)\boldsymbol{\mu})\,[\boldsymbol{\mu}^T\text{diag}(\pi)\boldsymbol{\mu}]^{-1}. \tag{2.12}$$

Analysis on the population covariance matrix

Assume that the off-diagonal terms in $\mathbf{B}$ are the same, which is equivalent to saying that inner products of rows in $\boldsymbol{\mu}$ are the same, e.g.,

$$\mathbf{B} = \begin{bmatrix} \eta_1 & \eta_0 & \cdots & \eta_0 \\ \eta_0 & \eta_2 & \cdots & \eta_0 \\ \vdots & \vdots & \ddots & \vdots \\ \eta_0 & \eta_0 & \cdots & \eta_K \end{bmatrix}, \quad \text{or} \quad \boldsymbol{\mu}_i^T \boldsymbol{\mu}_j = \begin{cases} \eta_k, & \text{if } i = j = k; \\ \eta_0, & \text{if } i \neq j. \end{cases} \tag{2.13}$$

Here $\eta_0 \geq 0$ and $\eta_k \geq \eta_0, 1 \leq k \leq K$, where at most one of η_k's equal to η_0. Then $\mathbf{B}$ has rank K, and

$$\begin{aligned} E[X_1 X_1^T (\boldsymbol{\mu}_k^T X_1)] &= \eta_0 \sum_{j=1}^{K} \pi_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T + (\eta_k - \eta_0) \pi_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T \\ &= \eta_0 \Delta + (\eta_k - \eta_0) \pi_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T. \end{aligned} \tag{2.14}$$

Using Sherman-Morrison formula, we have

$$\begin{aligned} \Sigma(\boldsymbol{\mu}_k)^{-1} &= \frac{1}{\eta_0} \left(\Delta - \frac{(\frac{\eta_k}{\eta_0} - 1)\pi_k}{1 + (\frac{\eta_k}{\eta_0} - 1)\pi_k \boldsymbol{\mu}_k^T \Delta^{-1} \boldsymbol{\mu}_k} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T \right) \\ &= \frac{1}{\eta_0} \left(\Delta - \frac{(\frac{\eta_k}{\eta_0} - 1)}{1 + (\frac{\eta_k}{\eta_0} - 1)} \pi_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T \right) \\ &= \pi_k \boldsymbol{\mu}_k^* \boldsymbol{\mu}_k^{*T} + \sum_{j \neq k} \frac{\eta_j}{\eta_0} \pi_j \boldsymbol{\mu}_j^* \boldsymbol{\mu}_j^{*T}, \end{aligned} \tag{2.15}$$

where $\boldsymbol{\mu}_j^* = \frac{\boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_j\|} = \frac{\boldsymbol{\mu}_j}{\sqrt{\eta_j}}, j = 1, 2, \cdots, K$. Note that the second equality comes from that

$$\pi_k \boldsymbol{\mu}_k^T \Delta^{-1} \boldsymbol{\mu}_k = [\mathrm{diag}(\sqrt{\pi})\boldsymbol{\mu}(\boldsymbol{\mu}^T \mathrm{diag}(\pi)\boldsymbol{\mu})^{-1}\boldsymbol{\mu}^T \mathrm{diag}(\sqrt{\pi})]_{kk} = [I]_{kk} = 1.$$

The expression in (2.15) indicates that the precision matrix $\Sigma(\boldsymbol{\mu}_k)^{-1}$ spanned by $\boldsymbol{\mu}_j$'s with the coefficient π_k for $\boldsymbol{\mu}_k^* \boldsymbol{\mu}_k^{*T}$ and coefficients $\frac{\eta_j}{\eta_0}\pi_j$ for $\boldsymbol{\mu}_j^* \boldsymbol{\mu}_j^{*T}$ with $j \neq k$.

Remark 2.9. *When to expect the covariance to be helpful in clustering over k-means. Intuitively, the covariance matrix will provide better boundaries when the points from different Gaussian distributions have overlapping. There are two cases, the points can be well separated:*

- *Small η_0. In the case where η_0 is close to 0 and π_j's are equal, the coefficient for $\boldsymbol{\mu}_k^* \boldsymbol{\mu}^{*T}$ in this precision matrix is dwarfed by others; thus the precision matrix is close to a matrix spanning only by $\boldsymbol{\mu}_j^*, j \neq k$. On the other hand, the covariance matrix $\Sigma(\boldsymbol{\mu}_k)$ will be dominant by the direction of $\boldsymbol{\mu}_j^*$ and thus embedded points from the k-th group will spread more along the direct of $\boldsymbol{\mu}_k^*$. So points from different clusters spread along nearly orthogonal directions (the inner product η_0 is small). See leftmost panel in Figure 2.1.*

- *Large n. The error is shrinking at the rate of $\frac{1}{\sqrt{n}}$, which is a faster rate than α_n, the rate for centers under the sparse SBM. When n is large, the points are well-separated. See rightmost panel in Figure 2.1. However, generally for many sparse graph with finite graph sizes, we expect GMM to improve over k-means.*

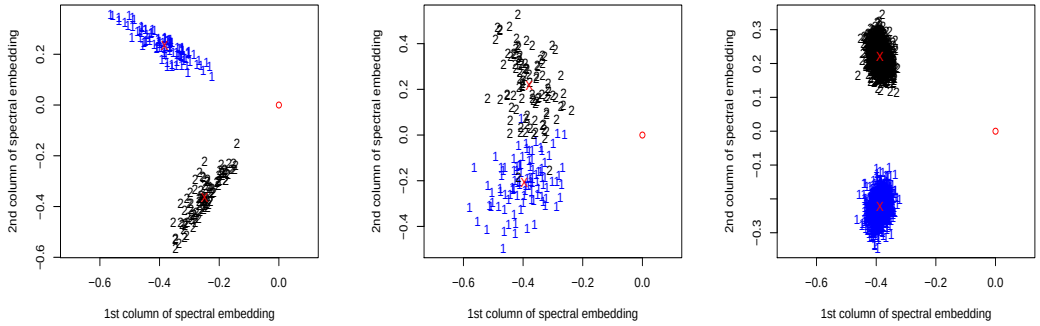


Figure 2.1: Illustrative plots see effects of η_0 and graph size n on how points are spread. Three plots are sampled from SBM with two blocks with $\alpha_n = 1$. Left: $\eta_1 = \eta_2 = 0.2, \eta_0 = 0.01, n = 200$; Middle: $\eta_1 = \eta_2 = 0.2, \eta_0 = 0.1, n = 200$; Right: $\eta_1 = \eta_2 = 0.2, \eta_0 = 0.1, n = 2000$. The red circle is the origin and the crosses denote the centers of clusters.

When η_0 is not very small compared to η_k, $k \in [K]$, and n is very extremely large, we see fitting GMM will provide a more accurate separating boundaries. Here is an illustration of advantages of fitting GMM in Figure 2.2.

Eigenspace MLE for SBM

Assume that (Z, A) is sampled from SBM with parameters $(\pi, \boldsymbol{\mu})$. Let $X = Z\boldsymbol{\mu} \in \mathbb{R}^{n\times d}$ and $\widehat{X}$ be the d-dimensional embedding of A as in Definition 2.2. Then we consider the rows in the spectral embeddings $\widehat{X}$ as i.i.d. samples from following generative model:

(i) $z_i \sim$ multinomial(π), for $i \in [n]$ with $P(z_i = k) = \pi_k, 1 \le k \le K$;

(ii) $\widehat{X}_i|_{z_i=k} \sim N(\boldsymbol{\mu}_k, \Sigma_k)$, where $\Sigma_k \triangleq \frac{1}{n}\Sigma(\boldsymbol{\mu}_k)$ with $\Sigma(\boldsymbol{\mu}_k)$ defined as in (2.12). Assume that conditioned on z_i's, $\widehat{X}_i$'s are independent.

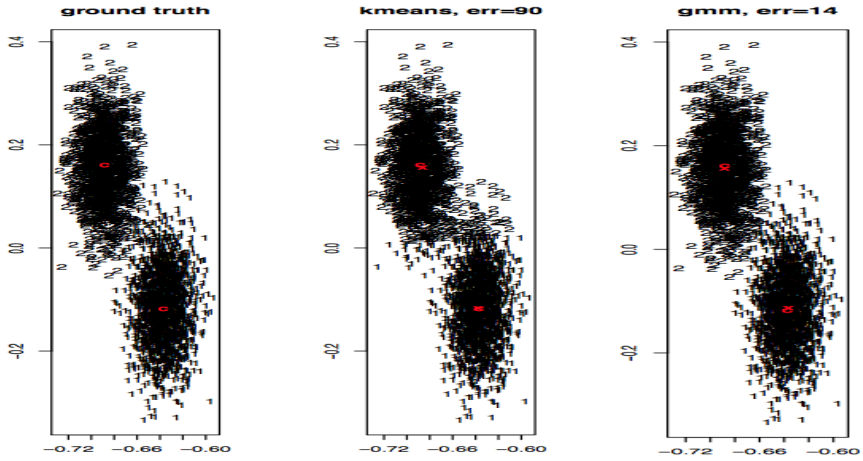


Figure 2.2: Covariance matrix can help find accurate separating boundaries. These are the scatter plots for the spectral embedding of an adjacency matrix with $n = 3000$ nodes from SBM with parameters $(\alpha_n = 1, \pi = [0.6, 0.4], \eta_0 = 0.42, \eta_1 = 0.42$ and $\eta_2 = 0.5)$ as in Athreya et al. (2016). Panel 1 displays the points with ground-truth labels. Panel 2 displays points with labels learned from k-means, which assumes spherical covariance structure and thus makes the more mistakes (90/3000) and GMM (using R package `mclust`) in Panel 3 fit elliptical covariance to the data and make fewer errors (14/3000).

Remark 2.10. *The distribution comes after* ***reasonable*** *approximations or simplification on* $\widehat{X}$*:*

1. *Asymptotic distribution* $\rightarrow$ *finite* n*;*

2. *Finitely asymptotic independence* $\rightarrow$ *independence among all rows;*

Let $\widehat{X}$, Z, and its corresponding z be defined as in Lemma 2.7. Based on the CLT in Lemma 2.7, we can reasonably assume the rows in $\widehat{X}$ are approximately i.i.d. samples from the following mixture Gaussian distribution.

Assumption 1. *Assume the rows in* $\widehat{X}$ *are sampled with following procedure:*

(i) $z_i \sim$ *multinomial*(π), *for* $i \in [n]$ *and* z_i*'s are independent.*

(ii) $\widehat{X}_i|_{z_i=k} \sim N(\boldsymbol{\mu}_k, \Sigma_{n,k})$, *where* $\Sigma_{n,k} \triangleq \frac{1}{n}\Sigma(\boldsymbol{\mu}_k)$ *with* $\Sigma(\boldsymbol{\mu}_k)$ *defined as in (2.12). Assume that conditioned on* z_i*'s,* $\widehat{X}_i$*'s are independent.*

Remark 2.11. *From Lemma 2.7 to the assumption, we did following simplification. 1, we absorb sparsity* α_n *into* X *and* $\boldsymbol{\mu}$ *by denote the* $\sqrt{\alpha_n}X$ *as the new* X *and the* $\sqrt{\alpha_n}\boldsymbol{\mu}$ *as the new* $\boldsymbol{\mu}$*. This won't have effect on expression of the covariance matrix since* $\Sigma(\boldsymbol{\mu}_k) = \mathbf{\Sigma}(\sqrt{\alpha_n}\boldsymbol{\mu}_k)$*. 2, For no rotation matrix* W_n *in the normality, we absorb* W_n *into the* $\boldsymbol{\mu}_k$*. In particular, let* $\widetilde{\boldsymbol{\mu}_k} = W_n^T\boldsymbol{\mu}$*, then* $W_n^T\Sigma(\boldsymbol{\mu}_k)W_n = \Sigma(\widetilde{\boldsymbol{\mu}}_k)$ *from Eqn. (2.12). Note that* $\widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T = \mathbf{B}$*, and* $\widetilde{\boldsymbol{\mu}}^T\mathrm{diag}(\pi)\widetilde{\boldsymbol{\mu}} = W_n^T\Delta W_n$ *may not be diagonal any more. That is, sample* $\widehat{X}$ *is estimating* $Z\widetilde{\boldsymbol{\mu}}$*, where* $\widetilde{\boldsymbol{\mu}}$ *is a rotated version of* $\boldsymbol{\mu}$ *and maintains the same inner product. Without confusion, we still denote* $\widetilde{\boldsymbol{\mu}}$ *as* $\boldsymbol{\mu}$*, without assuming* $\boldsymbol{\mu}^T\boldsymbol{\mu}$ *being diagonal.*

Under Assumption 1, the log likelihood function of $(\boldsymbol{\mu}, \pi)$ given data $\widehat{X}$ is:

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \pi|\widehat{X}) &= \sum_{i=1}^{n} \log \sum_{z_i} \pi_{z_i} N(\widehat{X}_i; \boldsymbol{\mu}_{z_i}, \mathbf{\Sigma}_{n,z_i}) \\
&= \sum_{i=1}^{n} \log \sum_{z_i} q_i(z_i) \frac{\pi_{z_i} N(\widehat{X}_i; \boldsymbol{\mu}_{z_i}, \mathbf{\Sigma}_{n,z_i})}{q_i(z_i)} \\
&\geq \sum_{i=1}^{n} \sum_{k} w_{ik} \log \frac{\pi_k N(\widehat{X}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_{n,k})}{w_{ik}} \triangleq Q((w_{ik}), \boldsymbol{\mu}, \pi) \qquad (2.16)
\end{aligned}$$

where $q_i(\cdot)$ is a mass function on $\{1, 2, \cdots, K\}$, $w_{ik} = q_i(k)$, and the inequality is from the Jensen inequality. The EM algorithm alternates between maximizing Q with respect to (w_{ik}) and $(\pi, \boldsymbol{\mu})$, respectively, holding the other fixed.

Algorithm 1: approximate EM algorithm for SBM.

Input: spectral embedding $\widehat{X} \in \mathbb{R}^{n\times K}$ and number desired clusters K.

Output: estimates for $(w_{ik}), \pi, \boldsymbol{\mu}$.

(i) Initialization $\pi, \boldsymbol{\mu}$: If $K \leq 20$ or $n \leq 2000$, we initialize them by the results from k-means on $\widehat{X}$. Otherwise, initialize them with hierarchical clustering (k-means cannot find good initial points).

(ii) E-step: Given $(\pi_k, \boldsymbol{\mu}_k, \Sigma(\boldsymbol{\mu}_k))_{k=1}^K$, to update $w_{ik} \triangleq P\left(z_i = k|\widehat{X}, \pi, \boldsymbol{\mu}\right)$.

Calculating probabilities:

$$w_{ik} = \pi_k|\Sigma_{n,k}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\widehat{X}_i - \boldsymbol{\mu}_k)^T\Sigma_{n,k}^{-1}(\widehat{X}_i - \boldsymbol{\mu}_k)\right); \quad w_{ik} \leftarrow \frac{w_{ik}}{\sum_j w_{ij}}. \tag{2.17}$$

(iii) M-step: Given (w_{ik}), to update the parameters $(\pi_k, \boldsymbol{\mu}_k)$. Note that We cannot get the exact maximizer, and we arrive at following update by ignoring the dependence $\Sigma(\boldsymbol{\mu}_k)$ on the π and $\boldsymbol{\mu}$:

$$\pi_k \leftarrow \frac{\sum_{i=1}^n w_{ik}}{n}; \quad \boldsymbol{\mu}_k \leftarrow \frac{\sum_{i=1}^n w_{ik}\widehat{X}_i}{\sum_{i=1}^n w_{ik}}. \tag{2.18}$$

Then we calculate the variance matrix using the formula provided by CLT:

$$\mathbf{B} \leftarrow \boldsymbol{\mu}\boldsymbol{\mu}^T; \tag{2.19}$$

$$\Delta = \sum_{k=1}^K \pi_k\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T, \quad \Sigma_{n,k} \leftarrow \frac{1}{n}\Delta^{-1}\left(\sum_{j=1}^K \pi_j\mathbf{B}_{kj}\boldsymbol{\mu}_j\boldsymbol{\mu}_j^T\right)\Delta^{-1}. \tag{2.20}$$

(iv) Repeat E-step and M-step: Repeat until reaching the following stopping

criteria: the change of $\boldsymbol{\mu}$ and π is relatively small; or the $Q((w_{ik}), \pi, \boldsymbol{\mu})$ as in (2.16) is decreasing.

The detailed derivation for this EM algorithm is provided in Section B in the Appendix.

There is another way to update $\Sigma_{n,k}$ by calculating quantity in (2.10) using sample estimate from $\widehat{X}_i$'s directly:

$$\Delta \leftarrow \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i^T, \quad \Sigma_{n,k} \leftarrow \frac{1}{n}\Delta^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i^T (\widehat{X}_i^T \boldsymbol{\mu}_k)\right)\Delta^{-1} \tag{2.21}$$

We prefer the update in (2.18) since it requires less computation. But (2.21) can be a good idea when the estimates for $\boldsymbol{\mu}, \pi$ are not accurate.

Remark 2.12. *We want to emphasize that the maximization step requires us to calculate the MLE from an curved multivariate normal, where $\Sigma_{n,k}$'s have a complicated dependence on $\boldsymbol{\mu}_k$'s. We could not get an tractable MLE estimator. Instead, we provide a consistent estimator as in (2.18) obtained by ignoring the dependence. The limiting covariance structure is used in calculating covariance matrix and further the class probabilities in the E-step. The severe drawback of this approximation is that Q-score as in (2.16) is no longer monotonically increasing over iterations.*

2.4 Variational eMLE under DC-SBM with random degree

Lemma 2.13 (CLT for DC-SBM)**.** *Assume that (θ, z, A) is sampled from DC-SBM with parameters $(\alpha_n, f, \pi, \boldsymbol{\mu})$, where $\alpha_n = o(1)$ and $n\alpha_n = \omega(\log^4 n)$. Assume that f is a* ***discrete*** *distribution. Let $\Theta = \mathrm{diag}(\theta)$ and Z be the membership matrix defined based on z. Let $X = \Theta Z \boldsymbol{\mu}$ and $\widehat{X}$ be the d-dimensional spectral embedding of A, then there is a $K \times K$ orthonormal matrix W_n such that*

$$\sqrt{n}(W_n \widehat{X}_i - \sqrt{\alpha_n} X_i)|_{X_i = \theta_i \boldsymbol{\mu}_{z_i}} \xrightarrow{d} N(0, \boldsymbol{\Sigma}(\theta_i \boldsymbol{\mu}_{z_i})), \tag{2.22}$$

where $\boldsymbol{\Sigma}(x) := \Delta^{-1} E[X_1 X_1^T (x^T X_1)] \Delta^{-1}$ and $\Delta = \mathbf{E}(X_1 X_1^T)$.

Note the lemma holds for any discrete distribution f, so we assume plausibly that (2.22) holds for general f defined on $(0, \infty)$. Plugging in $x = \theta_i \boldsymbol{\mu}_k$ and after some calculation, we have

$$\Delta = \mathbf{E}\theta_1^2 \sum_j \pi_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T, \quad E[X_1 X_1^T (x^T X_1)] = \theta_i \mathbf{E}\theta_1^3 \sum_{j=1}^K \pi_j (\boldsymbol{\mu}_k^T \boldsymbol{\mu}_j) \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T, \tag{2.23}$$

where $\theta_1 \sim f$. Furthermore,

$$\Sigma(\theta_i \boldsymbol{\mu}_k) = \theta_i \frac{\mathbf{E}\theta_1^3}{(\mathbf{E}\theta_1^2)^2} (\boldsymbol{\mu}^T \mathrm{diag}(\pi) \boldsymbol{\mu})^{-1} \left[\boldsymbol{\mu}^T (\mathrm{diag}(\pi) \mathrm{diag}(\boldsymbol{\mu} \boldsymbol{\mu}_k)) \boldsymbol{\mu}\right] (\boldsymbol{\mu}^T \mathrm{diag}(\pi) \boldsymbol{\mu})^{-1}. \tag{2.24}$$

The covariance matrix is same as in Eqn. (2.12) and has the same properties as those in previous section, except the scalar $\theta_i \frac{\mathbf{E}\theta_1^3}{(\mathbf{E}\theta_1^2)^2}$. For f, We have only one requirement

that is mean equal to 1. Assume that f is Gamma distribution with shape and rate parameter equal to λ, i.e., Gamma(λ, λ). Then for $\theta_1 \sim Gamma(\lambda, \lambda)$, we have $\mathbf{E}\theta_1 = 1, \mathbf{E}\theta_1^2 = \frac{\lambda+1}{\lambda}, \mathbf{E}\theta_1^3 = \frac{(\lambda+1)(\lambda+2)}{\lambda^2}$. Therefore,

$$\Sigma(\theta_i \boldsymbol{\mu}_k) = \frac{\lambda+2}{\lambda+1} \left(\boldsymbol{\mu}^T \mathrm{diag}(\pi)\boldsymbol{\mu}\right)^{-1} \left[\boldsymbol{\mu}^T (\mathrm{diag}(\pi)\mathrm{diag}(\boldsymbol{\mu}\boldsymbol{\mu}_k))\boldsymbol{\mu}\right] \left(\boldsymbol{\mu}^T \mathrm{diag}(\pi)\boldsymbol{\mu}\right)^{-1}. \tag{2.25}$$

That is, $\Sigma(\theta_i \boldsymbol{\mu}_k)$ is a function of $\pi, \boldsymbol{\mu}$ and λ.

With the same argument as in previous section, we have the following distribution for rows in $\widehat{X}$. Let $\theta, \widehat{X}$, Z, and its corresponding z be defined as in Lemma 2.13. Based on Based on the CLT in Lemma 2.13, we can reasonably assume the rows in $\widehat{X}$ are approximately i.i.d. samples from the following mixture Gaussian distribution.

Assumption 2. *We assume rows in $\widehat{X}$ are sampled from following procedure:*

(i) $\theta_i \sim f, z_i \sim$ multinomial(π), for $i \in [n]$ and all θ_i, z_i's are independent.

(ii) $\widehat{X}_i|_{z_i=k} \sim N(\theta_i \boldsymbol{\mu}_k, \theta_i \Sigma_{n,k})$, where $\Sigma_{n,k} \triangleq \frac{1}{n}\Sigma(\boldsymbol{\mu}_k)$ with $\Sigma(\boldsymbol{\mu}_k)$ defined as in (2.12). Assume that conditioned on z_i's, $\widehat{X}_i$'s are independent.

The log likelihood function is:

$$\begin{aligned} \ell(\lambda, \pi, \boldsymbol{\mu}|\widehat{X}) &= \sum_{i=1}^{n} \log \int_{\theta_i \in [0,\infty)} \sum_{z_i=1}^{K} f(\theta_i)\pi_{z_i} N(\widehat{X}_i; \theta_i \boldsymbol{\mu}_{z_i}, \theta_i \boldsymbol{\Sigma}_{n,z_i}) d\theta_i \qquad (2.26) \\ &\geq \sum_{i=1}^{n} \mathbf{E}_{q(\theta_i, z_i)} \log \frac{\lambda^\lambda / \Gamma(\lambda) \theta_i^{\lambda-1} e^{-\lambda\theta_i} \pi_{z_i} N(\widehat{X}_i; \theta_i \boldsymbol{\mu}_{z_i}, \theta_i \boldsymbol{\Sigma}_{n,z_i})}{q(\theta_i, z_i)} \\ &\triangleq Q(q, \lambda, \pi, \boldsymbol{\mu}), \qquad (2.27) \end{aligned}$$

where $q(\cdot,\cdot)$ is a density function on $[0,\infty)\times\{1,2,\cdots,K\}$ and inequality in (2.27) holds due to the Jensen inequality. The EM algorithm alternates between maximizing Q with respect to q and $(\lambda,\pi,\boldsymbol{\mu})$, respectively, holding the other fixed. It is easy to show that the maximum in the E step obtained by setting q as the posterior, i.e, $q(\theta_i,z_i)=P(\theta_i,z_i|\widehat{X},\lambda,\boldsymbol{\mu},\pi)$. However, the posterior distribution is computationally intractable. We use the following variational approximation to EM algorithm.

For conjugacy and computation purpose, we use the following class of factorable distributions to approximate the posterior:

$$q(\theta_i,z_i|\widehat{X}_i,\lambda,\boldsymbol{\mu},\pi)=q_1(\theta_1|\alpha_i,\beta_i)\ \cdot\ q_2(z_i|w_i),$$

where q_1 is Gamma distribution and q_2 is multinomial distribution. Let

$$\begin{aligned}
& Q(\alpha_i,\beta_i,w_i,\lambda,\boldsymbol{\mu},\pi)\\
\triangleq\ & \sum_{i=1}^{n}\mathbf{E}_{q_i(\theta_i)q_2(z_i)}\log\frac{\lambda^\lambda\theta_i^{\lambda-1}e^{-\lambda\theta_i}\pi_{z_i}N(\widehat{X}_i;\theta_i\mu_{z_i}\Sigma_{n,z_i}}{q(\theta_i,z_i)}\\
=\ & \sum_{i=1}^{n}6n\mathbf{E}_{q_1(\theta_i)q_2(z_i)}(\lambda\log(\lambda)-\log\Gamma(\lambda)+(\lambda-1)\log\theta_i-\lambda\theta_i)\\
& +\mathbf{E}_{q_1(\theta_i)q_2(z_i)}\left(\log\pi_{z_i}-\frac{K}{2}\log\theta_i-\frac{1}{2}\log|\Sigma_{n,z_i}|\right)\\
& -\mathbf{E}_{q_1(\theta_i)q_2(z_i}\left(\frac{1}{2\theta_i}(\widehat{X}_i-\theta_i\mu_{z_i})^T\Sigma_{n,z_i}^{-1}(\widehat{X}_i-\theta_i\mu_{z_i}\right)\\
& -\mathbf{E}_{q_i(\theta_i)\log q_1(\theta_i)-\sum_{i=1}^n q_2(z_i}\log q_2(z_i)\\
=\ & -n(\lambda\log(\lambda)-\log\Gamma(\lambda))+(\lambda-1)(\psi(\alpha_i)-\log(\beta_i))-\lambda\sum_{i=1}^{n}\frac{\alpha_i}{\beta_i}\\
& +\sum_{i=1}^{n}\sum_{k=1}^{K}w_{ik}\log\pi_k-\frac{K}{2}\sum_{i=1}^{n}(\psi(\alpha_i)-\log(\beta_i))
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log|\mathbf{\Sigma}_{n,k}| - \frac{1}{2}\sum_{i=1}^{n}\frac{\beta_i}{\alpha_i-1}\widehat{X}_i^T\sum_{k=1}^{K} w_{ik}\Sigma_{n,k}^{-1}\widehat{X}_i \\
&+\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\,\widehat{X}_i^T\Sigma_{n,k}^{-1}\mu_k - \frac{1}{2}\sum_{i=1}^{n}\frac{\alpha_i}{\beta_i}\sum_{k=1}^{K} w_{ik}\,\mu_k^T\mathbf{\Sigma}_{n,k}^{-1}\mu_k \\
&+\sum_{i=1}^{n}[\alpha_i - \log(\beta_i) + \log(\Gamma(\alpha_i)) + (1-\alpha_i)\psi(\alpha_i)] \\
&-\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log w_{ik}. \qquad (2.28)
\end{aligned}
$$

Remark 2.14. *In the formula above, we used some properties of Gamma distribution. If* $\theta_i \sim$ *Gamma*(α_i, β_i)*, then* $\mathbf{E}\theta_i = \frac{\alpha_i}{\beta_i}$, $\mathbf{E}\frac{1}{\theta_i} = \frac{\beta_i}{\alpha_i - 1}$, $\mathbf{E}\log(\theta_i) = \psi(\alpha_i) - \log\beta_i, \mathbf{E}(\log\theta_i)^2 = (\log\beta_i + \psi(\alpha_i))^2 + \psi'(\alpha_i)$*, and* $\mathbf{E}[-\log q_1(\theta_i)] = \alpha_i - \log(\beta_i) + \log(\Gamma(\alpha_i)) + (1-\alpha_i)\psi(\alpha_i)$*, where* $\psi(\cdot) \triangleq \Gamma'(\cdot)$ *is called digamma function and* $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$.

Here, we get rid of the integration in the original expression by using q to approximate the posterior of hidden variables $(\theta_i, z_i)_{i=1}^n$. Using coordinate accent algorithm to update $(\alpha_i, \beta_i, w_i)_{i=1}^n$ and $(\lambda, \boldsymbol{\mu}, \pi)$, we have the following algorithm which we will call Variational EM algorithm (details are in Section B in the Appendix).

Algorithm 2: Variational EM algorithm.

Input: The desired number of cluster K and the K-spectral embedding of adjacency matrix A: $\widehat{X} \in \mathbb{R}^{n\times K}$.

Output: Memberships for the rows in $\widehat{X}$.

(i) Initialization: Let $\widehat{X}^*$ be defined as follows: i-th row in $\widehat{X}^*$ is the i-th row in $\widehat{X}$ divided by its norm. $\lambda = 1$. $\boldsymbol{\mu}$ are initialized by following procedure: applying k-means to $\widehat{X}^*$ to get the membership estimate and initialize $\boldsymbol{\mu}$ using the cluster centers based on the estimated membership.

(ii) E-step: given $(\lambda, \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{n,k})_{k=1}^K$, update (β_i, w_i). Assume that $\alpha_i > 1$ is a given constant (default equal to 2):

Update the parameter for β_i,

$$\beta_i \leftarrow \frac{B_i + \sqrt{B_i^2 + 4A_iC_i}}{2A_i}, \tag{2.29}$$

where $A_i = \frac{1}{2(\alpha_i - 1)} \widehat{X}_i \sum_{k=1}^K w_{ik} \Sigma_{n,k}^{-1} \widehat{X}_i$, $B_i = (\frac{K}{2} - \lambda)$ and $C_i = \alpha_i \left(\lambda + \frac{1}{2} \sum_{k=1}^K w_{ik} \boldsymbol{\mu}_k^T \Sigma_{n,k}^{-1} \boldsymbol{\mu}_k \right)$.

Calculate the membership probability:

$$\begin{aligned} w_{ik} &= \pi_k |\Sigma_{n,k}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \frac{\beta_i}{\alpha_i - 1} \widehat{X}_i \Sigma_{n,k}^{-1} \widehat{X}_i + \widehat{X}_i \Sigma_{n,k}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \frac{\alpha_i}{\beta_i} \boldsymbol{\mu}_k \Sigma_{n,k}^{-1} \boldsymbol{\mu}_k \right); \\ w_{ik} &\leftarrow \frac{w_{ik}}{\sum_{k=1}^K w_{ik}}. \end{aligned} \tag{2.30}$$

(iii) (approx.) M-step: given $(\alpha_i, \beta_i, w_i)_{i=1}^n$, update $(\lambda, \pi, \boldsymbol{\mu})$. For $k = 1, 2, \cdots, K$, Because the normalization equaions for λ are very complicated, we decide to use its moment estimate.

$$\lambda \leftarrow \left(\frac{1}{n} \sum_{i=1}^n (\frac{\alpha_i}{\beta_i})^2 - (\frac{1}{n} \sum_{i=1}^n \frac{\alpha_i}{\beta_i})^2 \right)^{-1}. \tag{2.31}$$

We use approximation again by ignoring the dependence of $\Sigma_{n,k}$ on $\pi, \boldsymbol{\mu}$ when taking derivative w.r.t. π_k, $\boldsymbol{\mu}_k$.

$$\pi_k \leftarrow \frac{1}{n}\sum_{i=1}^{n} w_{ik}, \text{ and } \boldsymbol{\mu}_k \leftarrow \frac{\sum_{i=1}^{n} w_{ik}\widehat{X}_i}{\sum_{i=1}^{n} \frac{\alpha_i}{\beta_i} w_{ik}}. \tag{2.32}$$

Calculate the covariance matrix use the formula provided in Eqn. (2.24)

$$\Delta = \sum_j \pi_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T; \quad \Sigma_{n,k} \leftarrow \frac{\lambda+2}{\lambda+1}\Delta^{-1}\left(\sum_j \pi_j (\boldsymbol{\mu}_j^T \boldsymbol{\mu}_k)\boldsymbol{\mu}_j \boldsymbol{\mu}_j^T\right)\Delta^{-1}. \tag{2.33}$$

(iv) Repeat E-step and M-step: Stop if Q starts decreasing or the change of $(\lambda, \pi, \boldsymbol{\mu})$ is small.

Alternatively, we can update $\boldsymbol{\Sigma}_{n,k}$ in (2.33) by substituting the population quantities in $\Sigma_{n,k} = \Delta^{-1}\mathbf{E}[X_1 X_1^T (\boldsymbol{\mu}_k^T X_1)]\Delta^{-1}$ with their sample versions:

$$\Delta \approx \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i^T, \text{and } \mathbf{E}[X_1 X_1^T (\boldsymbol{\mu}_k^T X_1)] \approx \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i^T (\boldsymbol{\mu}_k^T \widehat{X}_i). \tag{2.34}$$

Remark 2.15. *During the M-step we used some consistent estimates as approximation to the true optimizer. First , $\frac{1}{\lambda}$ is the variance of the Gamma(λ, λ) and its MLE normal equation Eqn (B.11) is very complicated. So we adopt (2.31) in the algorithm. Second, same as in SBM, when we update $(\lambda, \pi, \boldsymbol{\mu})$, we assume the derivative of $\Sigma_{n,k}$ with respect to $(\lambda, \pi, \boldsymbol{\mu})$ are zero. Compare with fitting with GMM to the rows in $\widehat{X}$ without knowing the covariance matrix, this algorithm makes a distinction when updating the covariance matrix $\Sigma_{n,k}$. This will eventually affect the weights when*

estimating $(\pi_k, \boldsymbol{\mu}_k)$ *in the M-step as in (2.32) and (2.32).*

2.5 eMLE for DC-SBM with fixed degree parameters

Definition 2.16 (DC-SBM with $(\alpha_n, \theta, \pi, \boldsymbol{\mu})$)**.** *Given* $0 < \alpha_n \leq 1, \mathbf{B} \in [0,1]^{K\times K}$ *and* $\pi \in \mathbb{R}^K$ *same as in Definition 4.1. Assume* $\alpha_n = o(1)$ *and* $n\alpha_n = \omega(\log^4 n)$*. Additionally, for each* i*, we have one extra parameter* $\theta_i > 0$*. We call a random graph* (z, A)*, where we only observe* A *and* z *is hidden, is sampled from DC-SBM with* $(\alpha_n, \theta, \pi, \mu)$ *if it is sampled from the following procedure:*

(i) Membership $z_i \overset{i.i.d.}{\sim}$ *Multinomial*(π) *and* z_i*'s are independent; and*

(ii) For node i, j*, Conditional on* z_i, z_j*,* $A_{ji} = A_{ij}$ *is a Bernoulli variable with success probability* $\alpha_n\theta_i\theta_j(\boldsymbol{\mu}\boldsymbol{\mu}^T)_{z_iz_j}$*; and* $A_{ii} = 0$*. All the* $\{A_{ij}, i < j\}$ *are independent.*

In matrix form, we have $X = \Theta\mathbf{Z}\boldsymbol{\mu}$; and $\mathbf{E}(A|X) = \alpha_n XX^T$. The difference from DC-SBM with random degree parameters is that this model is no long exchangeable, and further we don't have a proved CLT theorem for this spectral embedding of the adjacency under this model anymore.

Assume that the rows in $\boldsymbol{\mu}$ have lengths equal to 1 and the overall mean for θ_i's is equal to 1 (identifiability condition as in Lei et al. (2015) and Zhang et al. (2014)). All the parameters are free of n except α_n.

Assumption 3 (Normal distribution in the eigenspace)**.** *Let* (z, A) *be the sampled graph from DC-SBM with parameters* $(\alpha_n, \theta, \pi, \mathbf{B})$ *as above. Let* $\widehat{X}$ *be the*

K-dimensional embedding of A. From analogy from DC-SBM with random degree, we assume that,

(i) normality: $\widehat{X}_i \sim N(\theta_i\mu_{z_i}, \theta_i\mathbf{\Sigma}_{n,z_i})$, *where* $\mathbf{\Sigma}_{n,k} = \frac{1}{n}\Delta^{-1}E[X_1X_1^T(\mu_k^TX_1)]\Delta^{-1}$ *and* $X_1 = \alpha\mu_{z_i}$, $z_i \sim multinormial(\pi)$, *and* α *is a random sample from* $[\theta_1, \cdots, \theta_n]$*; and*

(ii) independence: assume that $\widehat{X}_i$*'s are independent.*

Remark 2.17. *Under the assumption that* θ_i*'s are prefixed as parameters and only* z_i*'s are random following the multinomial distribution, the exchangeability of* A *conditional on* $(\alpha_n, \theta, \pi, \mathbf{B})$ *will not hold. Exchangeability assumption is essential the current proof of the CLT under RDPG, and thus this model is not well supported yet.*

Under the model with the Assumption 3, the log likelihood is

$$\begin{aligned}
\ell(\theta, \pi, \boldsymbol{\mu}|\widehat{X}) &= \sum_{i=1}^{n} \log \sum_{z_i} N(\widehat{X}_i; \theta_i\mu_{z_i}, \theta_i\mathbf{\Sigma}_{n,z_i})\pi_{z_i} \\
&= \sum_{i} \log \sum_{z_i} q_i(z_i) \frac{\pi_{z_i} N(\widehat{X}_i; \theta_i\mu_{z_i}, \theta_i\mathbf{\Sigma}_{n,z_i})}{q_i(z_i)} \\
&\geq \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log \frac{\pi_k N(\widehat{X}_i; \theta_i\mu_k, \theta_i\mathbf{\Sigma}_{n,k})}{w_{ik}} \triangleq Q, \qquad (2.35)
\end{aligned}$$

where q_i is a distribution over $\{1, 2, \cdots, K\}$ and $w_{ik} \triangleq q_i(z_i = k)$ are the intermediate variables satisfying $\sum_{k=1}^{K} w_{ik} = 1, \forall i$.

As we know that the EM algorithm to get the MLE is equivalent to the coordinate accent algorithm to maximize the quantity Q with repsect to $\mathbf{W}$ and $(\theta, \pi, \boldsymbol{\mu})$ alternatively.

Algrithm 3: EM under DC-SCM with fixed degree.

Input: $\widehat{X} \in \mathbb{R}^{n\times K}$ spectral embedding of adjacency matrix A, and the number of blocks K.

Ouput: Estimates for π, $\boldsymbol{\mu}$, θ, $\boldsymbol{\Sigma}_{n,k}, \mathbf{W}$

Step 1: Given parameters $\theta, \boldsymbol{\pi}, \boldsymbol{\mu}$, update w_{ik} to maximize Q. The solution is:

$$w_{ik} \leftarrow P(z_i = k|\widehat{X}, \pi, \theta, \boldsymbol{\mu}). \tag{2.36}$$

Calculate the covariance matrices:

$$\boldsymbol{\Sigma}_{n,k} \leftarrow \frac{\sum_i w_{ik}\frac{1}{\theta_i}(\widehat{X}_i - \theta_i\mu_k)(\widehat{X}_i - \theta_i\mu_k)^T}{\sum_i w_{ik}}. \tag{2.37}$$

Then, calculate the posterior probability (assignment probability):

$$w_{ik} \leftarrow \pi_k|\boldsymbol{\Sigma}_{n,k}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\theta_i}(\widehat{X}_i - \theta_i\mu_k)^T\boldsymbol{\Sigma}_{n,k}^{-1}(\widehat{X}_i - \theta_i\mu_k)\right); \tag{2.38}$$

$$w_{ik} \leftarrow \frac{w_{ik}}{\sum_{k'\in[K]} w_{ik'}}. \tag{2.39}$$

Step 2: Given w_{ik}'s, Maximize Q with respect to π, μ_k's and θ_i's. Solving for the stationary point gives the following updates:

$$\begin{aligned} \pi_k &\leftarrow \frac{1}{n}\sum_i w_{ik}; \\ \mu_k &\leftarrow \frac{\sum_i w_{ik}\widehat{X}_i}{\sum_i w_{ik}\theta_i},\ \mu_k \leftarrow \frac{\mu_k}{\|\mu_k\|}. \end{aligned} \tag{2.40}$$

$$\theta_i \leftarrow \frac{-K + \sqrt{K^2 + 4B_iA_i)}}{2A_i}, \tag{2.41}$$

where $A_i = \sum_{k=1}^{K} w_{ik}\mu_k^T \mathbf{\Sigma}_{n,k}^{-1}\mu_k$ and $B_i = \sum_k w_{ik}\widehat{X}_i^T \mathbf{\Sigma}_{n,k}^{-1}\widehat{X}_i$.

Remark 2.18. *The update used in (2.37) doesn't utilize the detailed structure of* $\Sigma_{n,k}$*, which is function of* $\boldsymbol{\mu}$*. It only utilize the information that rows from same community, the covariance is proportional to the degree parameters. Base on the covariance structure in the limiting distribution, we have the other ways to update* $\Sigma_{n,k}$ *in (2.37).*

- *Method 1: The population quantities will be substituted by their sample versions:*

$$\Sigma_{n,k} \leftarrow \frac{1}{n}\Delta^{-1}\mathbf{E}[X_1X_1^T(\mu_k^T X_1)]\Delta^{-1}, \tag{2.42}$$

  where $\Delta \approx \frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i^T$*, and* $\mathbf{E}[X_1X_1^T(\mu_k^T X_1)] \approx \frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i^T(\mu_k^T\widehat{X}_i)$.

- *Method 2: update with the representation in terms of* (π, μ, θ)*, and then plug in their estimate:*

$$\Sigma_{n,k} \leftarrow \frac{1}{n}\Delta^{-1}\left(\sum_{k'}\phi_{k'}\pi_{k'}\mu_{k'}\mu_{k'}^T(\mu_k^T\mu_{k'})\right)\Delta^{-1} \tag{2.43}$$

  Here, $\Delta^{-1} = \sum_{k=1}^{K}\psi_k\pi_k\mu_k\mu_k^T$*, where* $\psi_k = \sum_{z_i=k}\theta_i^2$ *and* $\phi_{k'} = \sum_{z_i=k'}\theta_i^3$.

2.6 Another remedy for degree heterogeneity: k-directions algorithm

As indicated in the CLT result, $\widehat{X}_i \sim N(\theta_i\sqrt{\alpha_n}\boldsymbol{\mu}, \theta_i\Sigma(\boldsymbol{\mu})$. The variances change proportionately to the changes made to the means and the standard deviation is proportional to the square root of the degree parameter. From the signal to noise ratio or precision perspective, the rows with bigger norms will be more informative.

To give high importance to the rows with big norm in $\widehat{X}$. We propose the following objective function:

$$(ii) \max_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_K} \sum_{i=1}^{n} \max_{1\le k\le K} \langle \widehat{X}_i, \boldsymbol{\mu}_k\rangle \text{ s.t. } \|\boldsymbol{\mu}_k\|_2 = 1, \forall k. \tag{2.44}$$

Algorithm

There is quite some similarity among k-directions algorithm and traditional k-means, as well as k-directions algorithm and the normalized k-means[1]. Objective functions similar to (2.44) has been proposed and studied theoretically in the context of market segmentation in Kleinberg et al. (1998) and text mining Dhillon and Modha (2001).

Similar to k-means, the following EM-algorithm can be used to optimize the above objective fucntion.

Algorithm 4: k-directions algorithm for (2.44).

Input: $\widehat{X} \in \mathbb{R}^{n\times K}$, the desired number of clusters K .

[1]The term normalized k-means means the algorithms applying k-means after row normalizations.

i) E-step: if partition $[n] = \mathcal{C}_1 \cup \mathcal{C}_2 \cup ... \cup \mathcal{C}_K$ is given, update centers $\boldsymbol{\mu}$ as follows.

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \widehat{X}_i, \; \boldsymbol{\mu}_k \leftarrow \frac{\boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_k\|} \tag{2.45}$$

ii) M-step: if the centers $\boldsymbol{\mu}_k, k = 1, 2, ..., K$ are given, with norm = 1, to update the partition. Let $z_i := \arg\max_k \langle \widehat{X}_i, \boldsymbol{\mu}_k \rangle$, Assign $\widehat{X}_i$, the i-th row, to $\mathcal{C}_{z_i}$.

iii) Repeat i) and ii) until convergence.

MLE Justification

We provide an MLE justification for the normalization in K-directions and thus some support for the whole k-directions algorithm (The notation in this Proportion is separated from the notations used in the paper).

Proposition 2.19. *Given unknown parameters $\mu \in \mathbb{R}^K$ with $\|\mu\| = 1$, $\theta_i \in \mathbb{R}, i = 1, 2, ..., n$, and $\Sigma \in \mathbb{R}^{K \times K}$, assume $\widehat{X}_i \overset{ind}{\sim} N(\theta_i \mu, \theta_i \Sigma)$. Then the MLE for μ is a function of $\sum_{i=1}^n \widehat{X}_i$ and the direction is adjusted by the estimator for covariance Σ. When $\Sigma = \sigma^2 I_K$, then $\frac{\sum_i \widehat{X}_i}{\|\sum_i \widehat{X}_i\|}$ is MLE estimator for μ.*

comparison with other algorithms

- Comparison with k-means:

  E-step in this algorithm is different from k-means since it normalize the norm of centers to have unit length.

M-step in k-directions algorithm is essentially the same as that in k-means. This is due to the facts that $\|\widehat{X}_i - \boldsymbol{\mu}_k\|^2 = \|\widehat{X}_i\|^2 + \|\boldsymbol{\mu}_k\|^2 - 2\langle \widehat{X}_i, \boldsymbol{\mu}_k \rangle$ and $\|\boldsymbol{\mu}_k\| = 1$ imply that $\arg\min_k \|\widehat{X}_i - \boldsymbol{\mu}_k\|^2 = \arg\max_k \langle \widehat{X}_i, \boldsymbol{\mu}_k \rangle$

The normalization of the centers in E-step will reduce the effects that points with bigger norms will drag the centers far away from the origin and form clusters there.

- Comparison with normalized k-means:

  Normalized k-means will inflate the influence of points closer to the origin by the row normalization, which will give bad performance when there are a lot of noisy low-norm points.

- Comparison with spherical k-means:

  The algorithm was first proposed in Dhillon and Modha (2001) and analyzed theoretically in Banerjee et al. (2005). The objective function for spherical k-means is as follows:

$$\arg \max_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K} \sum_{i=1}^{n} \max_{1 \le j \le K} \cos(X_i, \boldsymbol{\mu}_j) \quad s.t. \, \|\boldsymbol{\mu}_j\| = 1, \forall j. \tag{2.46}$$

  Spherical k-means usually applies to cluster unit-norm vectors and projects the centers on the sphere as well. It will be very similar to normalized k-means with row normalization with the difference that this algorithm requires the center to have unit norms. This algorithm also doesn't take into account the length of rows in $\widehat{X}$.

2.7 Simulation

eMLE under SBM

We simulate networks from the four-parameter SBM, which assumes that all the probabilities along the diagonal in $\mathbf{B}$ are equal denoted as η_1, all the probabilities in the off-diagonal entries are equal denoted as η_0 and that the clusters have the same size, i.e. $\pi_k = \frac{1}{K}, 1 \leq k \leq K$, where K is the number of cluster. Including the graph size n, we denote this model as SBM(η_1, η_0, K, n). This is clearly a special case of the covariance analyzed in Section 2.3 and thus the analysis there can help explain the performances here.

To generate sparse graph, we introduce a sparsity parameter as $c_0 \frac{(\log n)^4}{n}$, where the constant c_0 is chosen to make the graph has expected degree equal to 10 when graph size $n = 1000$. We call the ratio $\frac{\eta_1}{\eta_0}$ the signal noise ratio (SNR), which will affect the difficulty of estimating the community memberships, so are the number of clusters K and the graph size n.

The algorithms considered in the comparison are 1) k-means, 2) GMM from highly optimized R package `mclust` without specifying the covariance structure (**g**mm) when K is large.

The following figures show the performance of different algorithms as n increases under different settings determined by the combinations of SNR and K. Certain settings are omitted, since the trend is already clear. Figure 2.4 shows the advantage of algorithm in term of clustering accuracy and Figure **??** shows the time taken by each algorithm under different settings. K-means is conducted with $K * 10$ random

initializations, where K is the number of clusters.

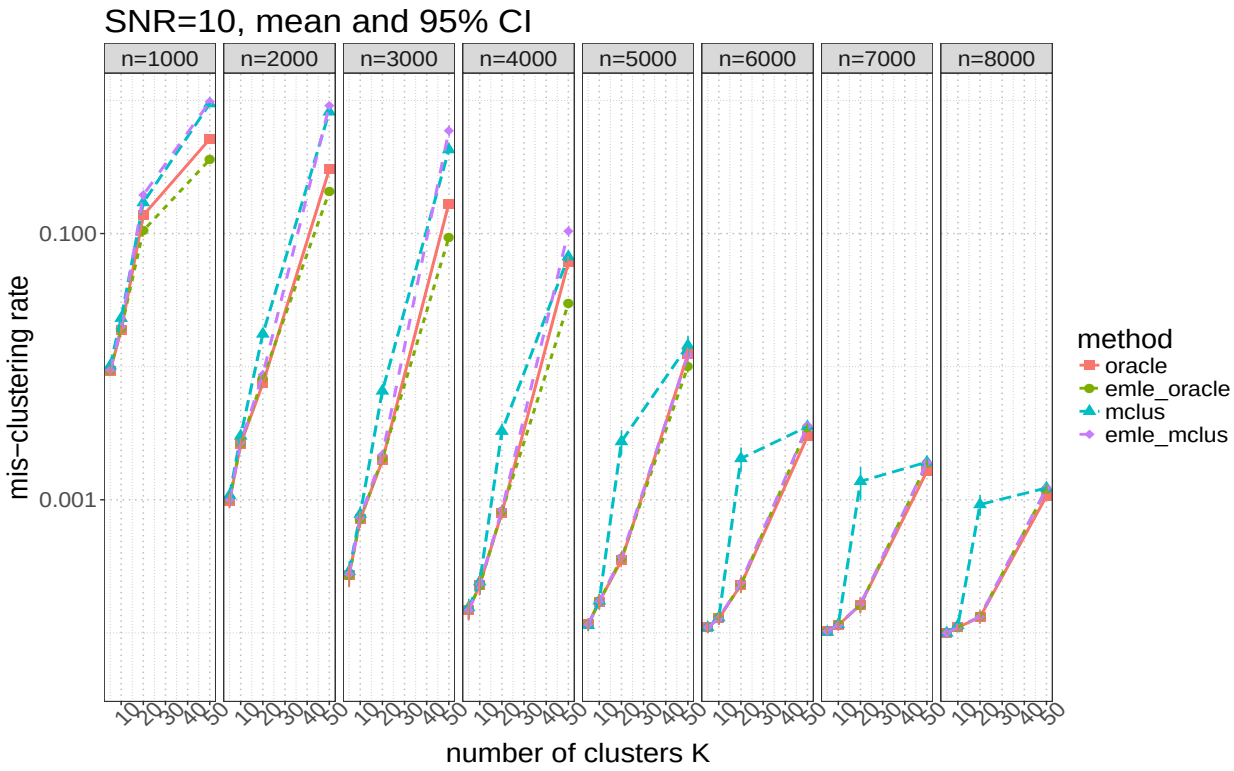


Figure 2.3: Misclustering rates of eMLE initialized by mclust and ground truth. We want to highlight the following. 1) eMLE outperforms the GMM in some of the cases where K is big and when there initializations have good accuracy. 2) oracle classifier based on limiting covariance matrices performs the best in all cases and this indicates the valuable information in the curvature.

eMLE under DC-SBM

Here are the description of the experiment setting(more experiments under more settings included later on):

graph size: $n = 2000$, sparsity:$\alpha_n = \frac{21}{n/2*0.35}$, cluster size proportion: $\pi = (0.6, 0.4)$

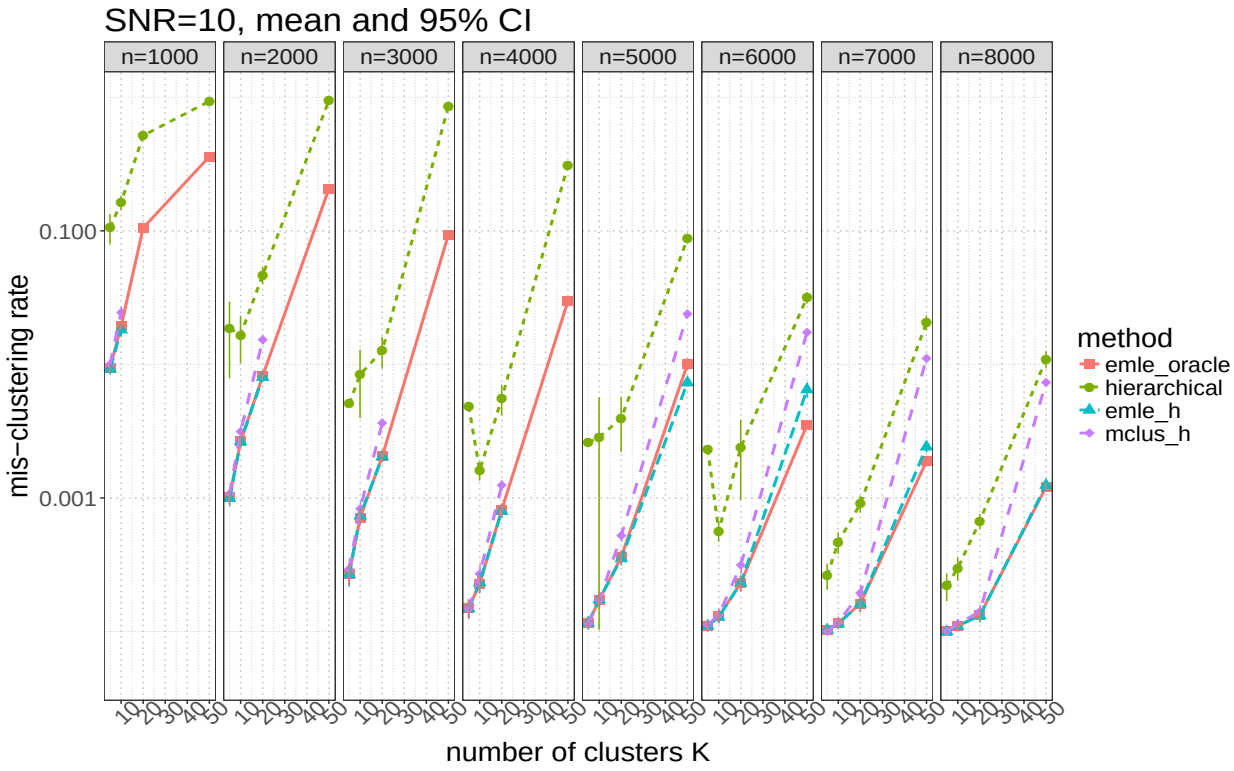


Figure 2.4: Misclustering rates of eMLE initialized by hierarchical clustering. We want to highlight the following. 1) eMLE outperforms the GMM in some of the cases where K is big and when there initializations have good accuracy. 2) oracle classifier based on limiting covariance matrices performs the best in all cases and this indicates the valuable information in the curvature.

and the block probability matrix:

$$B = \alpha_n \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}. \tag{2.47}$$

Additionally, the degree parameters θ_i's are sampled from exponential distribution with mean equal to 1.

Convergence behavior in a single experiment

From this single experiment, we can understand how the variance structure helped correct the mistakes introduced by initial points by k-means, and overcome the potential pitfalls for GMM without the scaling and considering of variance structures. The convergence speed is also quite fast.

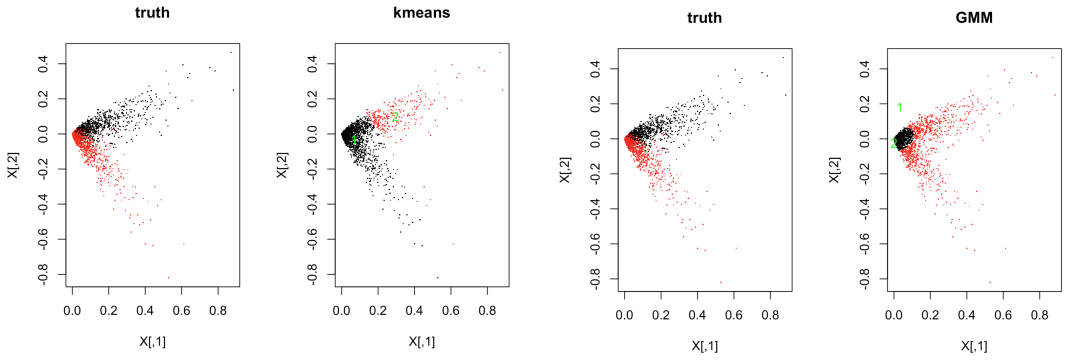


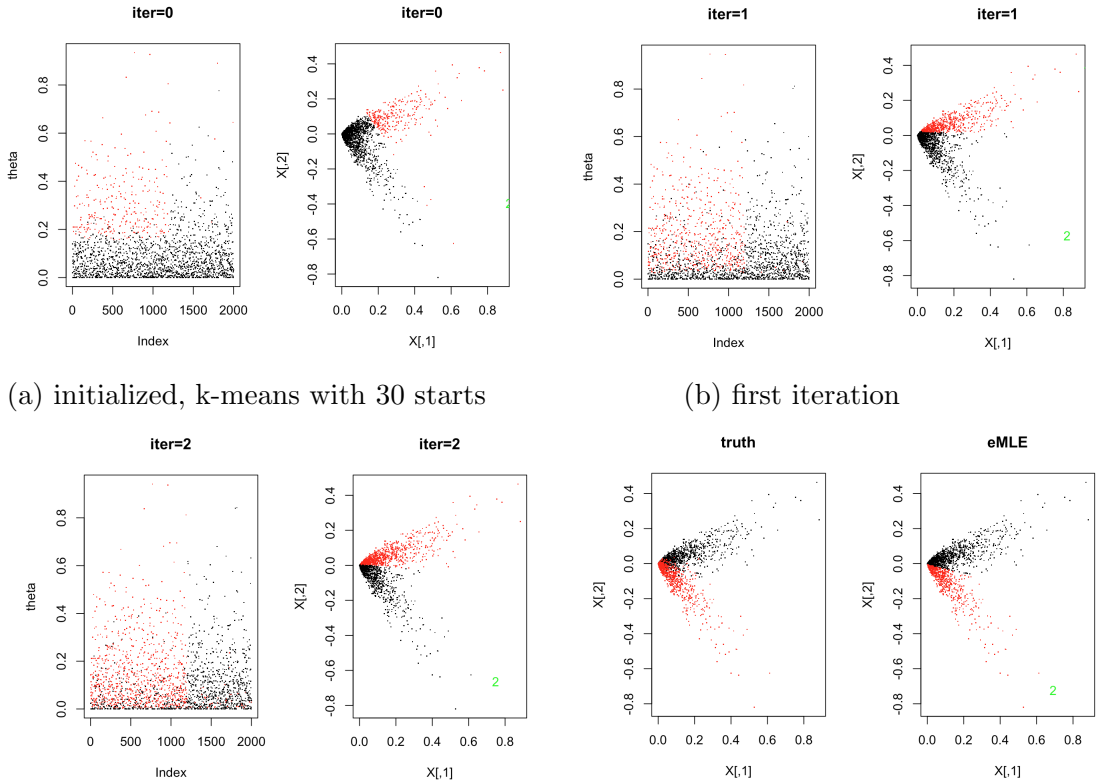
Figure 2.5: K-means (left) and GMM (right) fail to uncover the true clusters.

Mis-clustering Rate over Repeated Experiments

For the comparison with many other algorithms including the k-directions, normalized k-means, Spectral Clustering based on random walk, symmetric graph laplacian, you can see the "simulation.html".

For the same setting above, we repeat the experiments 100 times and the error rates for each methods are summarized in Figure 2.7 and Table 2.1 below.

There are total 21 pairwise comparisons, the critical value at $\alpha = 05$ for t-distribution after Bonferroni correction is 3.118. The paired t-test for `kmr` and `elme`



(a) initialized, k-means with 30 starts

(b) first iteration

(c) 2nd iteration

(d) converged results after 11 iterations

Figure 2.6: Convergence of eMLE starting with centers from k-means with 30 random initializations

	`kmeans`	`km+row`	`kangle`	`gmm`	`emle`	`lrw`	`lsys`
mean	0.4587	0.1205	0.1191	0.4323	0.1166	0.1453	0.0952
std dev	0.0162	0.0079	0.0080	0.0201	0.0082	0.0165	0.0069
std error	0.0016	0.0008	0.0008	0.0020	0.0008	0.0017	0.0007

Table 2.1: The means, standard deviations, standard errors for the error rates for 7 methods over 100 replicates

is 7.878, which indicates that `elme` improves `km+row` significantly by decreasing error rate by 3%. From the results above, `elme` is only worse than regularized Spectral Clustering based on symmetric Laplacian.

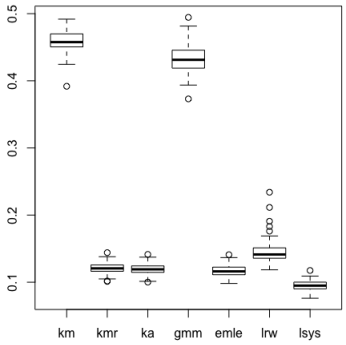


Figure 2.7: *Boxplot of mis-clustering rates for 7 different methods.* The first 5 methods utilize the eigenvectors from adjacency matrix, while the last two use eigenvectors from regularized Laplacians. Experiments are repeated 100 times.

2.8 Discussion and Future work

In this paper, we proposed a new EM-type algorithm called eMLE to recover community structures from the spectral embedding of the adjacency matrix of graphs. The algorithm is derived based on the maximum likelihood idea after applying the CLT for rows in the spectral embeddings provided in paper Athreya et al. (2016) and Tang and Priebe (2016). We have shown in simulation that the newly developed algorithms under SBM have advantages under SBM over k-means algorithm and gmm from R package `mclust` without knowing the limiting covariance structure.

Addtionally, our few algorithms motivated by the CLT under DC-SBM works very well in simulation compare to the competitive methods like applying k-means after normalizing the rows in eigenvectors. Our new way to handle the degree correction will also be a big contribution.

However, there are still some places that we can make improvements.

1: Issues with approximation in M-step of the eMLE.

Currently, the update for μ_k is not the exact solution for the in the M-step, It is just an consistent estimator is the MLE in the case where $\Sigma_{n,k}$ is free of μ_k. This has some drawbacks since it may miss the real curvature. In the numerous experiments, I find the the Q in (2.16), which is the lower bound of the MLE, is no longer always increasing. On the contrary, in many cases, about 95%, it will decrease at some point. In the algorithm, I have to set the following stopping criteria: Q starts to decrease, or μ, π stabilizes. These are two common cases:

(i) algorithm stops very early (0 or 1 step) due to bad initialization – causing Q decrease.

(ii) algorithm stops early due to good initialization – μ, π stabilizes quickly.

This makes the current version eMLE algorithm under SBM heavily dependent the initialization an work like a soft assignment calculated based on the initial memberships and covariance matrix structure. For example, when the number of clusters and and the graph size are big, k-means has a hard time finding the good centers and eMLE with initialization from k-means perform poorly.

Potential improvements: (a) Iterative method to get the exact solution in M-step; (b) assume the off-diagonal of $\mathbf{B}$ are constant, use the nice formula in (2.15).

2: CLT assumption for DC-SBM with fixed degree parameters.

Under the assumption that θ_i's are prefixed as parameters and only z_i's are random variable following the multinomial distribution, the exchangeability of A conditional on $(\alpha_n, \theta, \pi, \mathbf{B})$ won't hold. The exchangeability is used in the proof to help convert

the result $\|\widehat{X} - X\|_F^2$ to that for individual row. So It will be great if we can prove an CLT justify the mixture normal assumptions under this partial random model.

The eMLE algorithm under this setting is easier compared the one under DC-SBM under random degrees. Some simple simulations (in the simulation part) based on this version of eMLE works well, too.

Chapter 3

Spectral Clustering in Stochastic Block Model with Transitivity-based Dependent Edges

Abstract

The Stochastic Block Model (SBM) is commonly used to model networks with community structures. However, the edges in the SBM and its existing generalizations are all assumed to be generated independently. We find that the number of small structures like triangles or the cluster coefficients for samples from SBM are significantly smaller than those in real network. In this paper, we propose a new generalization of Stochastic Block Model which generates a second round of edges based on shared nodes in original graph sampled from SBM, and thus allows dependence among the edges (called as Transitive SBM, or T-SBM). We exhibit that graphs generated from T-SBM enjoy higher clustering coefficients (both local and global) than SBM when the graph is sparse. Also we give an asymptotic spectral bound for this type of random graph from its expectation; furthermore we prove that under this model with dependent edges, spectral clustering still enjoys weakly consistent clustering results.

3.1 Introduction

Networks appear very commonly in modern lives, for example social networks (friendships between Facebook users), biological networks (gene interactions), information networks (email exchanges), and many others. From a research point of view, networks are a powerful tool to represent the relationships among different objects. Therefore, network analysis is of great interest to researchers in many fields. A review of modeling and inference on network data can be found in Kolaczyk (2009) and Newman (2010).

Among the many existing statistical models for networks with communities, the Stochastic Block Model (SBM) is one of most commonly used. In an SBM, there are n nodes, clustered in K disjoint groups or *communities*. The SBM assumes stochastic equivalence among nodes within the same community, and that each pair of nodes will be connected independently with probabilities only dependent on memberships of the two nodes (see Section 3.2 for details). Based on this model or its generalization, people have proposed various community detection algorithms including modularity maximization (Newman and Girvan (2004)), likelihood methods (Bickel and Chen (2009), Amini et al. (2013)), spectral clustering (Ng et al. (2001), Rohe et al. (2011), Lei and Rinaldo (2013)) and other methods. We will mainly focus on the spectral method for clustering.

Classical results for parametric spaces of fixed dimension study how the MLE performs under a general misspecified model (e.g. White (1982)). Analogously, we would like to understand how network clustering algorithms perform when the standard SBM does not hold. As mentioned in literature (e.g. Newman (2006)),

two common features of empirical networks are long tailed degree distribution and high transitivity. These features are not preserved by the SBM. The SBM assumes that nodes in the same block are stochastically equivalent, which cannot model the variation of degrees for different nodes. To create long detailed distribution, the degree-corrected SBM (DC-SBM) is proposed (Karrer and Newman (2011)), which can generate any degree distribution as specified before. Recently, Zhao et al. (2012), Qin and Rohe (2013), Jin (2012) and Lei and Rinaldo (2013) have generalized the SBM clustering results to the DC-SBM, illustrating how clustering algorithms must adapt when the degree distribution is skewed.

One feature that both the SBM and DC-SBM have missed is the high transitivity(also known as clustering coefficients) in real networks. Table 3.2 and Figure 3.1a illustrates how the the political blog network has an excessive number of triangles and bigger average local clustering coefficients compared to estimates under the SBM. These phenomena suggest that there should be more dependence among the edges. In the literature, Exponential Random Graph Models (including Markov random graphs) are used to directly model this extra dependences in network data (Robins et al. (2007), Hunter et al. (2008) etc). However, ERGM is computation intractable and MCMC methods used to estimate parameters have no converge guarantees.

Different from those in literature, we propose a new generalized SBM that maintains the good properties of SBM and allows for additional dependence among the edges. The contributions of this paper are as follows. Firstly, we propose a new generative model (T-SBM), which generates a random network through two steps: sample a network with independent edges from an SBM ; then conditional on the gen-

erated network, generate some more *conditionally independent* edges with probability proportional to their transitivity, or common neighbors the two nodes share. The network generated from this model significantly improves dependence among edges and increases clustering coefficient. This is consistent with the observation in real networks. Secondly, we obtain a non-asymptotic bound for the difference between the sample adjacency matrix and its population version, and prove the consistency of spectral clustering under this non-traditional SBM.

The rest of article is organized as follows. In Section 3.2, we give formal introduction to the Stochastic Block Model and spectral clustering. We propose the T-SBM and conduct the population analysis in Section 3.3. We present the main theorems about the consistency of spectral clustering under this new model in Section 3.4. Simulation and real data analysis are presented in Section 3.5. Discussion remarks are given in Section 3.6. All the proofs are given in the Appendix.

Notation: For a graph $G = (V, E)$, where V is the node or vertex set, E is the edge set. Say the graph has N nodes. $V = \{1, 2, ..., N\}$ and pair $(i, j) \in E$ if there is a link between node i and j. The graph is represented by adjacency matrix A with $A_{ij} = 1$ if there is a link between node i and j and $A_{ij} = 0$ otherwise. Let $d_i = \sum_j A_{ij}$ be the degree of node i,for $i = 1, 2, ..., N$. Let $d_{mean}(A) := \frac{1}{N}\sum_i d_i$, i.e. mean degree of A. $d_{max}(A) := \max_i d_i$, $d_{min}(A) = \min_i d_i$ are maximum /minimum degree of A respectively. $D_A = \text{diag}(d_1, d_2, .., d_N)$ is a diagonal matrix made up of the degrees of A. $\|\cdot\|$ is spectral norm of a matrix or ℓ_2 norm of a vector. $\|\cdot\|_F$ standards for the Frobenius norm of a matrix. In this paper, terms like random network /random graph/ random matrix are interchangeable.

3.2 The SBM and Preliminaries

Stochastic Block Model

For an graph $G = (V, E)$ with N nodes, clustered in K blocks, let $z : \{1, 2, ..., N\} \to \{1, 2, ..., K\}$ be the block membership mapping. So, $z_i = k$ means that node i is in block k. Let B be a $K \times K$ symmetric matrix with $B_{ab} \in [0, 1]$ for all $a, b = 1, 2, \cdots, K$. This is called block probability matrix where B_{ab} is the probability that there is a edge between a pair, given one node of the pair is from block a and the other is from block b. We say a random matrix A is sampled from SBM with z and B, if $A_{ij} = A_{ji} \sim$ Bernoulli (B_{z_i,z_j}), $i < j$. Given mapping z, assume that different edges are generated independently.

Matrix expression for SBM: Let $Z \in \mathbb{R}^{N \times K}$ be the block membership matrix with i-th row having 1 in z_i-th cell and 0's in the rest of the cells. Then $\mathbf{E}A = ZBZ^T - \text{diag}(ZBZ^T)$.

Spectral Clustering

The standard versions of the spectral clustering algorithm are based on the graph Laplacian matrice $D_A^{-1}A$ or $D_A^{-1/2}AD_A^{-1/2}$, (e.g. Ng et al. (2002), Von Luxburg (2007), Rohe et al. (2011)). In this paper, we are introducing the algorithm based on adjacency matrix (e.g. Lei and Rinaldo (2013), Lyzinski et al. (2013)):

- Input: adjacency matrix A, number of clusters K.

– Output: clusters $V_1, V_2, \cdots, V_K$.

Step 1: Calculate the first K leading eigenvectors $X_1, X_2, \cdots, X_K \in R^N$. Put the K vectors together, $X = [X_1, X_2, \cdots, X_K] \in R^{N \times K}$.

Step 2: Consider each row of X as a point in R^K and run k-means on X with K clusters. Obtain K non-overlapping clusters $V_1, V_2, \cdots, V_K$ with union equal to V.

Step 3: Output the $V_1, V_2, \cdots, V_K$.

Eigen Structure for standard SBM

Let $\mathcal{A} := ZBZ^T$ and its eigen decomposition is $\mathcal{A} = \mathcal{X}\Lambda\mathcal{X}^T$. Then $\mathcal{X} = Z(Z^TZ)^{-1/2}U$, where U and Λ are eigenvectors and eigen values of $(Z^TZ)^{1/2}B(Z^TZ)^{1/2}$. This means that (i) For i and j with $z_i = z_j$, the i-th and j-th row of $\mathcal{X}$, $\mathcal{X}_{i*}$ and $\mathcal{X}_{j*}$, are equal; and (ii) For i and j with $z_i \neq z_j$, the i-th and j-th row of $\mathcal{X}$, $\mathcal{X}_{i*}$ and $\mathcal{X}_{j*}$ the are orthogonal. If we denote the size of k-th cluster is n_k, then $\|\mathcal{X}_{i*} - \mathcal{X}_{j*}\| = \sqrt{\frac{1}{n_{z_i}} + \frac{1}{n_{z_j}}}$ (as in Rohe et al. (2011)).

Therefore, the eigen-decomposition of $\mathcal{A}$ yields $\mathcal{X}$; applying k-means with specified number of clusters equal to K on $\mathcal{X}$ gives perfect clustering results. From matrix perturbation, sample A is expected to be close to its population $\mathcal{A}$ and we will expect to get similar result from A.

Clustering Coefficients

Clustering coefficients are defined to be the ratio of of triangles in the networks. There are two versions of clustering coefficients: the global and the local.

The global clustering coefficient (also called transitivity in many literatures) gives an indication of the clustering in the whole network (global). Luce and Perry (1949) first defined the quantity based on counting the number of triplets. A triplet consists of three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties. A triangle consists of three closed triplets, one centered on each of the nodes. The global clustering coefficient of graph G is defined as:

$$C_{global} = \frac{\text{number of closed triangles in G } \times 3}{\text{number of triplets in G}}.$$

The local clustering coefficient of a vertex (node) in a graph quantifies how close its neighbors are to being a clique (complete graph). Denote $\mathcal{N}_i$ the neighborhood of node i excluding i itself, then the local clustering coefficient at node i is defined :

$$C_i = \frac{\text{number of edges among nodes in } \mathcal{N}_i}{|\mathcal{N}_i|(|\mathcal{N}_i| - 1)/2}.$$

As an alternative to global clustering coefficient, the overall level of clustering in a network is measured by Watts and Strogatz (1998) as the average of the local clustering coefficients $C_{local} = \frac{1}{n}\sum_{i=1}^{n} C_i$.

3.3 T-SBM model and its population analysis

Transitivity-based SBM

A random graph from the T-SBM is generated in two steps:

(1) Generate A_1 from the regular SBM, i.e. $P([A_1]_{ij} = 1) \sim Bernoulli(B_{z_i z_j}), i < j$; $[A_1]_{ji} = [A_1]_{ij}, i < j$; and $[A_1]_{ii} = 0$.

(2) Adding some more edges based on A_1. Let $\mathbf{P}([A_2]_{ij} = 1|A_1) = [P_{A_1}]_{ij}, i < j$, where $P_{A_1} = D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2}$; $[A_2]_{ji} = [A_2]_{ij}, i < j$; and $[A_2]_{ii} = 0$.

The final adjacency matrix $A = t(A_1 + A_2)$, where t is a cell-wise operator on $\mathbb{R}^1$ with $x \mapsto 1$, if $x > 1; x \mapsto x$ otherwise. So $A = A_1 + A_2 - A_1 \circ A_2$, where $\circ$ stands for entry-wise product (Hadamard product).

Remark 3.1. *: The choice of P_{A_1} is indeed a good and reasonable one. It has the following good properties: 1) each number in P_{A_1} doesn't exceeds one; 2) the mean degree of A_2 is comparable to that of A_1, which means we add non-negligible number of new edges.*

Population Analysis

Let $\mathcal{A} = \bar{A}_1 + D_{\bar{A}_1}^{-1/2} \bar{A}_1^2 D_{\bar{A}_1}^{-1/2}$, where $\bar{A}_1 = ZBZ^T$. The (i, j) th element of $\mathcal{A}$ is

$$\mathcal{A}_{ij} = B_{z_i, z_j} + d_{z_i}^{-1/2} (BZ^T ZB)_{z_i z_j} d_{z_j} (\bar{A}_1)^{-1/2}$$

It is easy to see tell that $\mathcal{A}_{ij}$ only depends on the memberships of nodes i, j i.e. z_i, z_j. Actually, $\mathcal{A}$ can be expressed in matrix as follows:

$$\mathcal{A} = Z\widetilde{B}Z^T, \tag{3.1}$$

where $\widetilde{B} = B + D_B^{-1/2}B(Z^TZ)BD_B^{-1/2}$, and $D_B = diag(BZ^TZ\mathbf{1}_K)$. Similar to the result in traditional SBM setting, the expectation of adjacency matrix in this new setting has clear block structure as well. Therefore, we should expect that spectral cluttering algorithm will work under this model and our main result provides an affirmative guarantee for this.

3.4 Main Results

Theorem 3.2 (**Spectral Bound**). *Let A and $\mathcal{A}$ be defined in Section 3.3 and Section 3.3. Let $\Delta(\delta)$ be the maximum(minimum) expected node degree and $b := \sqrt{\frac{\log(2n/\epsilon)}{\delta}} < 1/\sqrt{3}$. Assume that $\frac{d_{mean}(\bar{A}_1)^2(1-b)^2}{d_{max}(\bar{A}_1)(1+b)} > \frac{4}{9}\log(\frac{2n}{\epsilon})$. With probability at least $1-3\epsilon$, we have the following bound:*

$$\|A - \mathcal{A}\| \quad \leq \quad O(b\Delta), \ \textit{where } \Delta = O(n\alpha_n) \tag{3.2}$$

The proof of this theorem and the detailed expression of the upper bound are in the appendix of this paper.

Remark 3.3. *Due to efforts to bound A_2, the bound here is actually weaker than*

the bound in standard SBM as in Lemma C.1: 1) have some conditions on δ; 2) $\|A - \mathcal{A}\| \leq \Delta\sqrt{\frac{4\log(2n/\epsilon)}{\Delta}}$ in Lemma C.1 gives a narrower bound. If the Δ and δ of the same order, the difference will be negligible.

Based on the above spectral bound on $\|A - \mathcal{A}\|$, we can use the version Dave-Kahan's theorem Lemma 5.1 in Lei and Rinaldo (2013) (attached as Lemma C.2 in Appendix) to bound the differences of their corresponding eigenvectors. According the result, the following holds: *There is a $K \times K$ rotation matrix $\mathcal{O}$ such that*

$$\|X - \mathcal{X}\mathcal{O}\| \leq \frac{2\sqrt{2K}\|A - \mathcal{A}\|}{\lambda_K}. \tag{3.3}$$

According to the algorithm in Section 3.2, we will consider each row of X as a point in K dimensional space and run k-means on set of points. Let $C_1, C_2, \cdots, C_n$ (K distinct ones) denote the corresponding centers for each row in X; and $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_n$ be the centers for each row in the $\mathcal{X}$, when we run k-means on population version $\mathcal{X}$. Because the Davis-Kahan's result above, we expect that C_i is closer to $\mathcal{C}_i\mathcal{O}$ than other population centers.

Definition 3.4 (Mis-clustering)**.** *Call node i in the graph is correctly clustered if there is no j, such that $\mathcal{C}_j\mathcal{O}$ is closer to C_i than the center $\mathcal{C}_i\mathcal{O}$ is. Based on this we define the following mis-clustered set of nodes:*

$$\mathcal{M} = \{i : \exists j \neq i, s.t. \|C_i - \mathcal{C}_i\mathcal{O}\| > \|C_i - \mathcal{C}_j\mathcal{O}\|\}.$$

Theorem 3.5 (**Mis-clustering Error Rate**). *Assume* $\frac{d_{mean}(\bar{A}_1)^2(1-b)^2}{d_{max}(\bar{A}_1)(1+b)} > \frac{4}{9}\log(\frac{2n}{\epsilon})$ *and* $b := \sqrt{\frac{\log(2n/\epsilon)}{\delta}} < \frac{1}{\sqrt{3}}$ *as in Theorem 3.2. Let* $\lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$ *be the* K *positive eigenvalues of* $\mathcal{A}$. *Let* $\mathcal{M}$ *be defined as in Definition 3.4. Let* n_i *be the size of cluster* i, *and* $n_{\max} = \max_{1\le i\le K} n_i$ *be maximum cluster size. Then with probability at least* $1-3\epsilon$, *we have*

$$\frac{|\mathcal{M}|}{n} \le c_0 \frac{K n_{\max}\,\Delta^2}{n\lambda_K^2} b^2 = c_0 \frac{K n_{\max}\,\Delta^2}{n\lambda_K^2}\,\frac{\log(2n/\epsilon)}{\delta}. \tag{3.4}$$

Remark 3.6. *The mis-cluster rate bound is higher than under SBM by a factor* $\frac{\Delta}{\delta}$.

Remark 3.7. *(Four-parameter SBM)The four-parameter Stochastic block model SBM*(K, s, p, r) *is as follows:* K *is the number of blocks,* s *is the number of nodes within each block,* p *is the probability of an edge occurring between two nodes from the same block, and* r *is the probability of for two nodes from different blocks.*

Then $d_{max}(\bar{A}_1) = d_{min}(\bar{A}_1) = s(p + r(K-1))$. *We have* $\mathcal{A} = Z\widetilde{B}Z^T$ *where*

$$B = \begin{pmatrix} p + \frac{p^2+(K-1)r^2}{p+(K-1)r} & r + \frac{(K-2)r^2+2pr}{p+(K-1)r} \\ r + \frac{(K-2)r^2+2pr}{p+(K-1)r} & p + \frac{p^2+(K-1)r^2}{p+(K-1)r} \end{pmatrix}.$$

So $\lambda_K(\mathcal{A}) = s(p_1 - r_1) = s(p-r)(1 + \frac{p-r}{p+(K-1)r}) = s(p-r)\frac{2p+(K-2)r}{p+(K-1)r}$. *Therefore assume that* p, r *are constants and allow* K *to change, we have*

$$\frac{\mathcal{M}}{n} = O(\frac{\log(2n/\epsilon)}{s}\,\frac{(p + r(K-1))^3}{(p-r)^2(2p+(K-2)r)^2}) = O(\frac{K^2\log(2n/\epsilon)}{n})$$

with probability $\ge 1-3\epsilon$. *This result coincides with the result in Qin and Rohe (2013).*

3.5 Simulation & Data Analysis

Simulation

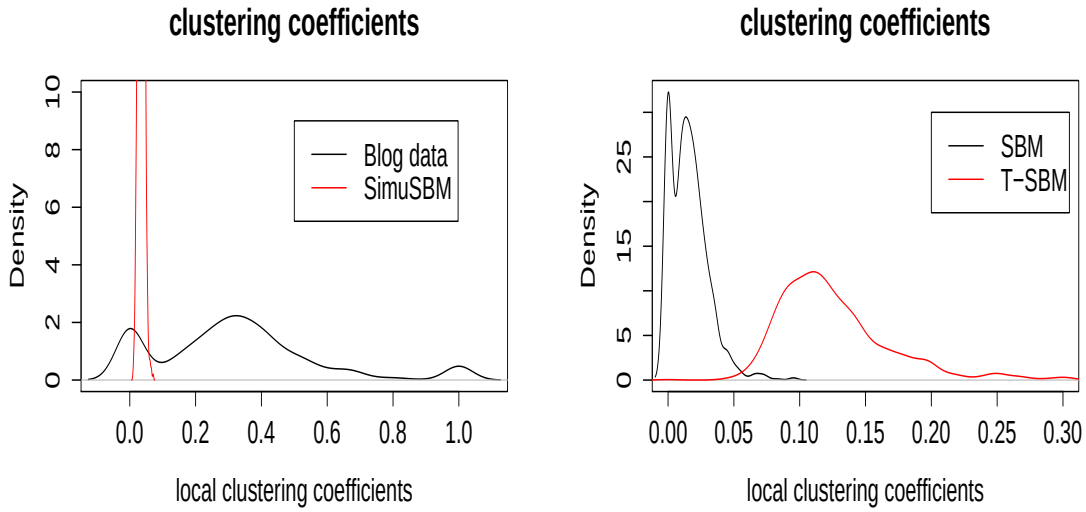
First, we did do simulation study with four parameter-model SBM$(K, s, p, r) = (3, 500, 0.01, 0.004)$, and SBM$(K, s, p, r) = (3, 1000, 0.005, 0.002)$. The SBM model is generated from $\widetilde{B}$ as defined in remark of Theorem 3.5, TSBM is generated from two-step procedure. The comparison between SBM and T-SBM are summarized in the Table 3.1 and in Figure 3.1b: We can see that T-SBM will increase local/global

Table 3.1: Clustering Coefficients in T-SBM v.s. SBM.

Source	mean degree	triangles	mean(local)	global
SBM(500)	17.97(0.19)	2975.4(148.9)	0.0123(0.0003)	0.0123(0.0005)
T-SBM(500)	18.61(0.35)	20428.9(589.2)	0.0837(0.0010)	0.0922(0.0011)
SBM(1000)	17.98(0.16)	3012.9(103.1)	0.0062(0.0002)	0.0062(0.0002)
T-SBM(1000)	18.56(0.15)	38389.8(513.2)	0.0802(0.0006)	0.0874(0.0006)

clustering coefficients and number of triangles significantly even when the two models have the similar edge density (degree). This is especially true when the graph is sparse.

Let's take a look at the performance of spectral clustering algorithm on T-SBM. For $A = A_1 + A_2 - A_1 \circ A_2$, in simulation, we can see that it can still do a reasonable job of clustering when we run spectral clustering on A, which is corresponding to the observed graph. But the mis-clustering rate is usually higher compared to applying spectral clustering to the hidden A_1 directly, even though A_1 is sparser. This may indicate that when newly added edges blur the community structure and makes it more difficult to cluster the nodes. This is actually consistent with the theorem we



(a) Clusering coefficients from Blog data (b) Clusering coefficients from T-SBM

Figure 3.1: (a) shows that clustering coefficients from 2004 political blog data Adamic and Glance (2005) is significantly higher than that in simulated SBM network. (b) SBM and T-SBM have the similar mean node degree, but T-SBM significantly improves the local clustering coefficients

proved, the mis-clustering rate converge to zero more slowly.

Political Blog Data

We use SBM and DC-SBM as well as T-SBM to the fit the political blog data Adamic and Glance (2005). The results are as in Table 3.2. As we can see in figure 3.1b, T-SBM improves the local clustering coefficients and number of triangles in the network by doubling or tripling them. This normally should be true, since the edges we added in the second step in the generating process more likely to create triangles and increase the local clustering coefficients. This may be also in part due to the fact that

adding edges will significantly improve the local clustering coefficient. For example, for a node with degree = 2 and there are no edge between these two neighbors first, adding one edge between the two neighbors can increase local clustering coefficients from 0 to 1.

From the table, it seems that DC-SBM works even better in terms of mimicing the clustering coefficients and number of triangles in the original network. Maybe the varying degree can contribute a lot more to the number of triangles and cluttering coefficients. We should try to incorporate both DC-SBM with edge dependence idea.

Table 3.2: Comparisons of transitivity under different models (SBM, DC-SBM, T-SBM). We fit different models to political blog network and use bootstrapping to do the resampling to see the sensitivity of each model.

Models	mean degree	triangles	mean(local)	global
Blog data	27.36	101043	0.3203	0.2259
SBM	27.33(0.21)	5235.3(137.0)	0.0344(0.0006)	0.0344(0.0006)
DC-SBM	26.82(0.18)	88073(1796)	0.2175(0.0059)	0.2221(0.0025)
T-SBM	27.81(0.33)	11593.9(288.1)	0.0826(0.0006)	0.0826(0.0006)

YouTube Data

YouTube is a video-sharing web site that includes a social network. In the YouTube social network, users form friendship each other and users can create groups which other users can join. We consider such user-defined groups as ground-truth communities. This data is provided by Mislove et al. (2007), Yang and Leskovec (2012).

Originally there are 8385 communities, and we only keep those communities with more than 100 members, leaving us with 133 of such communities. We remove members that don't belong to any of these communities and work on the largest

Table 3.3: Comparisons of transitivity under different models (SBM, DC-SBM, T-SBM). We fit different models to YouTube network and use bootstrapping to do the resampling to see the sensitivity of each model.

Models	mean degree	triangles	mean(local)	global
YouTube	12.47	254349	0.1761	0.0587
SBM	12.46(0.032)	2976(70)	0.0040(0.0001)	0.0045(0.0001)
DC-SBM	12.24(0.029)	232085(2388)	0.0697(0.0008)	0.0617(0.0004)
T-SBM	12.88(0.086)	60898(659)	0.1202(0.0013)	0.1030(0.0006)

connected component, which in the end has 132 communities, totally 20208 users. For those users who are in multiple communities, we assign them to one of the communities they belong to randomly. Based on the true memberships, we fit SBM, DC-SBM, an approximate TSBM to the data set, and do a bootstrapping to resample networks based the given parameters to see the properties of these three types of models. The results are shown in Table 3.3.

3.6 Discussion

We have proposed a modified model based on the traditional Stochastic Block Model by adding second round of edges with probability proportional to the number of their shared nodes. In this way, we create the dependence among edges in the graphs and further high transitivity of the real network and more triangles in the sampled networks. This indicates that networks sampled under T-SBM can better mimic the properties of real networks and can server better benchmark model for community detection algorithms. This paper proves that spectral clustering is not very sensitive to edge dependence, as long as in the expectation the graph has balanced distinct

block structure.

There also remain some places to work on in the future.

1: This is a generative model, we think the graph is generated from this process and want to do some inference about the observed graph. We are using an optimization package to estimate the block probability, which can be quite expensive when the number of clusters K is large. We need to find efficient estimation method.

2: In simulation, DC-SBM enjoys even a much bigger improvement of cluster coefficients compared with both T-SBM and SBM. It seems reasonable to extend our work and build a Transitive DC-SBM.

3: This model is likely to be extended to adding edges through multiple runs (currently 2), it will be a very interesting problem to investigate this dynamic process.

Chapter 4

Applying Random Projection to Detecting Mixed Memberships in Stochastic Blockmodel

Abstract

Community detection is a fundamental problem in network analysis. In the past decade, Researchers from different fields have made a lot of progress by proposing various statistical networks models and community detection algorithms. It remains a challenging problem to recover community structure from networks especially those with overlapping community structure. In this paper, we propose a simple algorithm, which keeps projecting rows in the leading eigenvectors onto randomly selected directions and select those achieving extreme values. This algorithm utilizes the geometric relationships among the spectral representations of mixed nodes and the pure nodes and use the projection to select those potential pure nodes. Our algorithm will detect the pure nodes first and then figure out the memberships for the mixed nodes. Under the sparsity assumption and some other constraint, our algorithm will be able to estimate the memberships consistently with high probability.

4.1 Introduction

The Stochastic Block Model is a popular random network model for community detection. It has many variants including the Degree-Corrected SBM, and many of them assume that each node belongs to a single block. However, in real-world networks, each node may belong to multiple communities and the links will reveal these multiple memberships. For example, in social network, a person may connect to co-workers, classmates from same school or friends in the same neighborhood etc. In the past few years, various models and algorithms about detecting overlapping memberships in networks have been proposed. They can be categorized into the following groups: Bayesian models and inferences (Airoldi et al. (2008),Gopalan and Blei (2013)),Tensor methods (Anandkumar et al. (2013)), spectral methods(Zhang et al. (2014), Mao et al. (2017), Rubin-Delanchy et al. (2017)). Compared with the other algorithms, spectral methods, which utilizes the spectral properties of the eigenvectors of the graph matrices, are computationally tractable and easy to implement.

In this paper, we have proposed a different and simple algorithm utilizing the eigenvectors of the graphs to estimate the community memberships under the model proposed by Zhang et al. (2014). Compared with algorithm in Zhang et al. (2014), where its conditions for consistency very vague and difficult to check, our algorithm is easy to understand and conditions are easier to check. Also the model used in this paper is more general than that in Mao et al. (2017) and Rubin-Delanchy et al. (2017).

The intuition of our methods comes from the fact that for points located within a

polytope such as a triangle in a plane, the vertex of triangles are more likely to take biggest or smallest values when those points are projected onto a random direction. This indicates that we can use iterative projection to select those pure nodes since in the spectral space, the mixed nodes and pure nodes have similar relationships. More accurately, as we will see later, mixed nodes are located within the affine space spanned by pure nodes. We prove the distance of the rows in the eigenvectors of the population graph and sampled graph and further use this as the tool to show that our algorithm will be consistent.

The organization of the paper are as follows. In Section 4.2, we present the model under which the networks with mixed memberships are sampled and the algorithm we propose to estimate the memberships. In Section 4.3, we analyze the consistency of a variant version of this algorithm and proved its consistency using the bound we developed using sample spitting. In Section 4.4, we conclusion this chapter. All the proofs are provided in the Appendix 4.

4.2 Model & Algorithm

Definition 4.1 (Degree-Corrected SBM)**.** *Denote the Network by its adjacency matrix as G, a symmetric binary matrix with $\{G_{ij}, i < j\}$ independent Bernoulli random variables. Let $\bar{G} = \boldsymbol{E}\, G$. Assume the $\bar{G}$ has the following strcture:*

$$\bar{G} = \alpha_n \Theta Z P Z^T \Theta.$$

Here $Z \in \mathbb{R}^{N\times K}$ is the membership matrix, where ***each row has a single 1 and the rest 0****; $P \in [0,1]^{K\times K}$ is the block connectivity matrix; and $\Theta \in \mathbb{R}^{N\times N}$ is a diagonal matrix, with $\Theta_{ii} = \theta_i > 0$ as the degree parameter at the node i.*

In this paper, we are tackling multiple membership problems of nodes in a graph, where rows of Z can have multiple non-zero entries.

Definition 4.2 (Overlapping Continuous Community Assignment Model (OCCAM) in Zhang et al. (2014).)**.** *Graph G is sampled in the same way as described in Definition 4.1 with $\bar{G}$ having the following structure:*

$$\bar{G} = \alpha_n \Theta Z P Z^T \Theta. \tag{4.1}$$

The only difference from Definition 4.1 is that rows in Z are relaxed and satisfies that $\|Z_i\|_2 = 1$ for $1 \le i \le N$ and $Z_{ij} \ge 0$. This model is called Overlapping Continuous Community Assignment Model (OCCAM) in Zhang et al. (2014).

To enforce the identifiability of the model, we propose the following conditions (same as in Zhang et al. (2014)) on the parameters:

(1) P is positive definite with $P_{ii} = 1, 1 \le i \le K$; and

(2) $Z_{ik} \ge 0$ $\|Z_i\|_2 = 1$, $i = 1, 2, 3, \ldots, N$. Additionally, assume that there is at least one "pure" node in every community, i.e. for each $k = 1, 2, \ldots, K$, there exists at least one i such that $Z_{ik} = 1$.

(3) The mean of the degree parameters is 1. i.e., $\frac{1}{n}\sum_i \theta_i = 1$.

Furthermore, we can represent Assumption (2) as follow.

$$Z = \begin{bmatrix} Z_{\mathrm{P}} \\ Z_{\mathrm{M}} \end{bmatrix}, \tag{4.2}$$

where Z_{P} is the membership matrix for the pure nodes, i.e. each row has exact one non-zero element in the entry; while Z_{M} is the membership matrix for the mixed nodes, i.e. each row can have multiple non-zero entries.

The population graph $\bar{G}$ in (4.1) has the following unique spectral structure.

Lemma 4.3 (Population eigenvectors). *Let $\bar{G} = \bar{X}\bar{\Lambda}\bar{X}^T$ with $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_K > 0$. Then there is orthogonal matrix $U \in \mathbb{R}^{K\times K}$ such that*

$$\bar{X} = \Theta Z(Z^T\Theta^2 Z)^{-1/2}U \text{ and } \bar{X}^* = ZU \tag{4.3}$$

Proof. Its eigenvectors: $\mathcal{X} = \Theta Z(Z^T\Theta^2 Z)^{-1/2}U$, where U comes from the eigen-decomposition of $(Z^T\Theta^2 Z)^{1/2}B(Z^T\Theta^2 Z)^{1/2}$. The i-th row of $\mathcal{X}$ is $\mathcal{X}_i = \theta_i Z_i V$ where $V = (Z^T\Theta^2 Z)^{-1/2}U$.

The normalized version is $\mathcal{X}_i^* = \|Z_i V\|^{-1} Z_i V$.

This indicates that if node i is a pure node, $\mathcal{X}_i = V_{Z_i}$ and that node i is a mixed node, $\mathcal{X}_i$ in the cone spanned by the pure nodes and $\mathcal{X}_i^*$ is on sphere with pure nodes as its vertices. □

Remark 4.4. *The geometry among the spectral representations among the pure nodes and mixed nodes can be shown in the following plot Figure 4.1.*

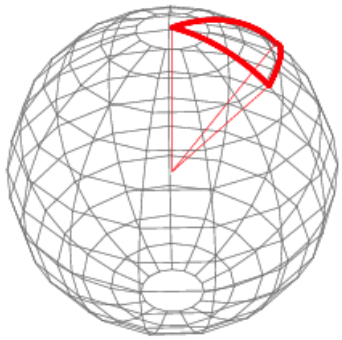


Figure 4.1: Visualization of $\mathcal{X}^*$. The endpoint are corresponding to the pure nodes, while the points in the middle part on the sphere are corresponding to the mixed nodes

Due to this geometric structure, we propose the following algorith, which find the extreme points first and then estimate memberships for the mixed nodes.

Eigenvectors+Projection (EigenProjection)

Input: (Given adjacency matrix G, number of clusters K, parameters $\delta_n, \tau_n > 0$ to be determined.)

S1 Calculate eigenvectors X of G , $X \in \mathbb{R}^{n \times K}$.

S2 Normalize each row in X.

$$X^*_{ij} \leftarrow \frac{X_{ij}}{\sqrt{\sum_j X^2_{ij}}}.$$

For those $\|X_i\|$ is too small, we can add some constant to the norm of the vector during the normalization as in Zhang et al. (2014).

S3 Use random project to deselect mixed nodes while maintain those pure nodes (for given $\delta_n, \tau_n > 0$, their exact values will are given in the next section). for $i = 1, 2, ..$

- generate $w \in \mathbb{R}^{K\times 1}$, $w_i \overset{iid}{\sim} N(0,1)$. Project each row of Y onto w, we have Yw. Look at its top $1-\delta_n$ and bottom δ_n quantiles.
- combined all the indices, select the indices which appears $\geq N * \tau_n$, denoted the set as S, and their corresponding row vectors in X, denoted as X_S

S4 run k- kmeans on X_S, get centers $X_c \in \mathbb{R}^{K\times K}$, membership matrix $\widehat{Z}_S$

S5 calculate $\widehat{Z}_{S'} = (X \backslash X_S)X_c^{-1}$

S6 return $\widehat{Z} = [\widehat{Z}_S, \widehat{Z}_{S'}]$.

Lemma 4.5 (Uniform mixture)**.** *Assume that the non-pure nodes, their corresponding rows in $\mathcal{X}^*$ are uniformly or continuously distributed on the sphere, say Dirichlet distribution. For $\mathcal{X}^*$ defined as above, adopt the following projection rule: For every projection, we pick the the indices producing the* ***biggest and smallest*** *values under each projection. When the number of random projections, $N = O(K)$, is big enough, we can include all the endpoints among rows of $\mathcal{X}^*$. Additionally, they will be the one*

appear very frequently.

$$\frac{\textit{times of selecting node } i}{N} \to \begin{cases} 1/(1+n_{min}) \to 0, & \textit{if } i \textit{ is a mixed node;} \\ 1/K > 0, & \textit{if } i \textit{ is a pure node.} \end{cases} \tag{4.4}$$

Based on the Lemma 4.5, it is easy to see that $\delta_n = \frac{1}{N}, \tau_n = \frac{1}{2K}$, the algorithms can estimate the correct memberships if applied to the population graph $\bar{G}$. To make sure the algorithm will work properly on the sampled graph G, we hope that the rows in sample X and rows in population $\mathcal{X}$ are close. To achieve a good row-wise bound, we use the sample splitting procedure as in Lyzinski et al. (2013) and Lei and Zhu (2014) and further we prove that a variant of this algorithm is consistent. We present these analyses in the following section.

4.3 Consistency of the EigenProjection Algorithm

Denote the vetex set of G as V. V_A, V_B is a parition of V. Let A, B, C be the adjacency matrix for the sub-network induced by vertex sets, i.e. $A = G|_{V_A \times V_A}$ and $B = G|_{V_A \times V_B}, C = G|_{V_B \times V_B}$ are defined in a similar way. After reordering the vertex, we have the following partition the observed graph and population graph:

$$G = \begin{pmatrix} A & B^T \\ B & C \end{pmatrix} \text{ and } \bar{G} = \begin{pmatrix} \bar{A} & \bar{B}^T \\ \bar{B} & \bar{C} \end{pmatrix} = \begin{pmatrix} \Theta_A Z_A \\ \Theta_B Z_B \end{pmatrix} \mathbf{P} \begin{pmatrix} Z_A \Theta_A \\ Z_B \Theta_B \end{pmatrix}^T \tag{4.5}$$

Based on A , the estimate for the leading K eigenvectors of A can obtained, such that $A = X\Lambda X^T$. We want to estimate the corresponding row representation for node in B using formula: $Y = BX\Lambda^{-1}$. Let $U = [X^T, Y^T]^T$ be the our estimate for top K eigenvectors of G.

Lemma 4.6 (**Population eigenvectors**). *Let* $\bar{A} = \bar{X}\bar{\Lambda}\bar{X}^T$ *with* $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_K > 0$. *Let* $\bar{Y} = \bar{B}\bar{X}\bar{\Lambda}^{-1}$. *then the followings results hold:* .

1. $\bar{X} = \Theta_A Z_A (Z_A^T \Theta_A^2 Z_A)^{-1/2} U$
2. $\bar{Y} = \Theta_B Z_B (Z_A^T \Theta_A^2 Z_A)^{-1/2} U$

Lemma 4.7 (**Concetrationon of adjacency matrix in Chung and Radcliffe (2011)**). *Let* A, $\bar{A}$ *defined as above. Let* Δ *denote the maximum expected degree of* A. *Let* $\epsilon > 0$ *and suppose that for* n_A *sufficiently large,* $\Delta > \frac{4}{9}\log(2n/\epsilon)$, *Then with probability at least* $1 - \epsilon$, *for* n *sufficently large, we have:*

$$\|A - \bar{A}\|_2 \leq \sqrt{4\Delta \log \frac{2n_A}{\epsilon}} \tag{4.6}$$

Lemma 4.8 (**Davis-Kahan Theorem in Lei and Rinaldo (2013)**). *Assume the conditions in Lemma 4.7, let* $\bar{A}$ *has rank* K, *and* λ_K *is the smallest non-zero singular values of* $\bar{A}$. *Let* $X, \bar{X} \in \mathbb{R}^{n_A \times K}$, *be the top* K *eigenvectors of* A, $\bar{A}$, *then there exists an orthogonal matrix* $\mathcal{O} \in \mathbb{R}^{K \times K}$. *With probability at least* $1 - \epsilon$

$$\|X - \bar{X}\mathcal{O}\|_F \leq \frac{2\sqrt{2K}\|A - \mathcal{A}\|}{\bar{\lambda}_K} \tag{4.7}$$

Note that $\mathcal{O}$ can be chosen as a block-diagonal matrix with each block corresponding the different eigenvalues of $\bar{A}$ therefore $\bar{\Lambda}\mathcal{O} = \mathcal{O}\bar{\Lambda}$. Especially $\mathcal{O}$ will be diagonal with values ± 1 when the eigenvalues of $\bar{A}$ are distinct.

Theorem 4.9 (**Sample splitting and Row bound on** Y). *Let* $Y = BX\Lambda^{-1}$ *and* $\bar{Y} = \bar{B}\bar{X}\bar{\Lambda}^{-1}$. *Assume the following conditions:*

i) Assume conditions on $\bar{A}$ *in Lemma 4.7, 4.8;*

ii) Let $\gamma := \min\{\lambda_i(\bar{A}) - \lambda_{i+1}(\bar{A}) : i = 1, 2, \ldots, K\}$, *note* $\lambda_{K+1} = 0$. *Assume that there is* $c_0 > 0$, *such that* $\gamma \geq c_0\, n\alpha_n$; *and*

iii) Let $\bar{A}$ *have rank* $= K$, *and* $\bar{A} = \bar{X}\bar{\Lambda}\bar{X}^T$ *be the full eigen-decomposition. Let* $\mu(\bar{X}) := \frac{n_A}{K}\max_i \|\bar{X}_i\|_2^2$, *called coherence in Jianqing Fan and Zhong (2016). Assume that* $\mu(\bar{X})$ *is bounded.*

Then, there exists diagonal matrix $\mathcal{O}$ *which is a diagonal matrix with values* ± 1, *with probability* $\geq 1 - \epsilon$, *the followings results hold:*

$$\mathbb{P}\left(\|y_i - \bar{y}_i\mathcal{O}\| \geq C_0 n^{-1/2}\sqrt{\frac{K\log(2n/\epsilon)}{n\alpha_n^2}}, \text{for all } i \in B\right) \leq \epsilon \tag{4.8}$$

Remark 4.10. *The norm of the* i*-th row in* $\mathcal{Y}$

$$\|\bar{y}_i\| = \|[\theta_B]_i(Z_B)_i(Z_A^T\Theta_A^2 Z_A)^{-1/2}U\| = O(n^{-1/2}) \tag{4.9}$$

So we have $\|y_i - \bar{y}_i\| \leq r_n\|\bar{y}_i\|$, *where* $r_n = O(\sqrt{\frac{K\log 2n/\epsilon}{n\alpha_n^2}})$ *This is a concentration result when* $r_n = o(1)$.

Remark 4.11. *From the results above, we can see rows in $\bar{X}\mathcal{O}$ and $\bar{Y}$ are from an affine space spanned by the same vertexes, $\bar{V} \equiv (Z_A^T Z_A)^{-1/2} U\mathcal{O}$. Our results is new in terms that this provides a row wise bound for the rows in Y and rows in $\bar{Y}$, which will facilitate the derivation of our main theorem and may be of independent interest. Our focus will to recover $\mathcal{V}$ from Y based on realized graph.*

Remark 4.12. *Row bounds for X. If we change substitute B by A, we have a row-wise bound for $\|X_i - \bar{X}_i\|_2$ as well, which is of order $\frac{\sqrt{K \log(2n/\epsilon)}}{n\alpha_n}$. This result is better than that obtained directly using Davis-Kahan and coincides with the result in Lyzinski et al. (2013). Unfortunately, we cannot put X and Y calculated in this way together and learn the pure nodes at once. Because they may have two different rotation matrices.*

Algorithm

This algorithm incorporating the sample splitting and learns the overlapping membership from the network.

Splitting + Projection + Clustering (SPC)

Input: (Given graph $G = (V, E)$, number of clusters K, parameters Λ)

S0 Divide the nodes set V into two equal parts randomly, denoted as V_A, V_B.

S1 Calculate eigenvectors X of $A = G|_{V_A \times V_A}$.

S2 Calculate the corresponding representation for V_B, by $Y = D_B^{-1/2} B X \Lambda^{-1}$.

S3 Let $F = [X, Y]$, then normalize each row in F. For those $\|F_i\|$ too small, we either throw it away, or use regularization by adding a small constant before normalization.

S4 Use random project to de-select mixed nodes; (Need to determine $0 < \delta_n, \tau_n < 1$, their meanings are given in later sections), for $i = 1, 2, .., N = O(K)$:

- Generate $w \in \mathbb{R}^{K\times 1}$, $w_i \overset{iid}{\sim} N(0, 1)$. Project each row of X onto w, we have $y = Xw$. Look at its top $1 - \delta_n$ and bottom δ_n quantiles.
- Combined all the indices, select the indices which appear more than $\frac{N\tau_n}{K}$ times, denoted the set as S, and their corresponding row vectors in X, denoted as X_S

S5 Run k- kmeans on Y_S, and denote the centers as $Y_c \in \mathbb{R}^{K\times K}$, membership matrix $\widehat{Z}_S$.

S6 Solve for $\widehat{Z}$ such that $\|F - ZY_c\|_F$ is minimized.

S7 Similarly, we re-run S1- S6, and have $\widehat{Z}_{S'} = (F \backslash Y_S)Y_c^{-1}$. Let $\widehat{Z} = [\widehat{Z}_S, \widehat{Z}_{S'}]$ and output.

For the algorithm above, we have the following consistency result.

Theorem 4.13 (Main result : K-means combined with Random Projection)**.** *Let $Y \in \mathbb{R}^{n\times K}$ be defined as above and Y^* is the row normalized version of Y. Use*

the following projection procedure to select a set of potential pure points: For $k = 1, 2, \ldots, N$*, generate a random direction* $d_k \in \mathbb{R}^K$*, Let*

$$S_k = \{i \in B : d_k^T y_i^* > (1 - \epsilon_n) \max_i d_k^T y_i^* \text{ or } d_k^T y_i^* < (1 + \epsilon_n) \min_i d_k^T y_i^*\}, \tag{4.10}$$

where $\epsilon_n = 2 * r_n$*, and* $r_n = O(\sqrt{\frac{K^2 \log(2n/\epsilon)}{n\alpha_n^2}})$*. Let* $S = \cup_{k=1}^N S_k$*, allowing for duplicates. Count the frequency, keep those points with frequency greater than* $N/3K$ *and Denote the final set of index* S*. Then we will have the following conclusions:*

1. S *contains* $O(n)$ *of pure points while at most* $O(nr_n)$ *mixed points*

2. *Run the k-means algorithm on* Y_S *to discover the centers, denoted as* $v_1, v_2, .., v_K$*. Estimate for the membership nodes in* A, B*,* $\widehat{Z}_A = XV^{-1}$*,* $\widehat{Z}_B = YV^{-1}$*. Let* $\widehat{Z} = [\widehat{Z}_A^T, \widehat{Z}_B^T]^T$*. Then*

$$\frac{1}{\sqrt{n}} \|\widehat{Z}^* - Z\|_F = O(\sqrt{\frac{\log(2n/\epsilon)}{n\alpha_n^2}}). \tag{4.11}$$

Remark 4.14. *The sparsity condition required for consistency, which is* $n\alpha_n^2 = \omega(\log n)$*, stronger than that in Zhang et al. (2014). But we don't put the condition that the proportion of pure nodes is not decreasing to 0. For our model method to work, we only requires there existence of pure nodes for each cluster.*

4.4 Conclusion

In this paper, we propose a novel and simple algorithm to estimate the mixed memberships under the model proposed in Zhang et al. (2014). This algorithms keeps projecting the rows of the eigenvectors of adjacency matrix onto random sampled directions and select those rows produced biggest/smallest values in the projection. This simple strategy serves to down sample the mixed nodes and upsample the pure nodes. K-means can estimate spectral representations of pure nodes from the selected rows easily and then estimate the memberships for mixed nodes. We proved that this algorithm can estimate the memberships consistency given mean expected degree grows at the rate of square root of n.

We believe that our algorithm will be consistent under sparser graph. We can achieve this by adapting the new bound developed in Mao et al. (2017). Also, we may refine our results by obtaining a bound on number of projects. This will give a better guide in practice.

Appendix A

Appendix for Chapter 1

Proof of Proposition 1.6 We revised the proof of Theorem 1.2 in Deijfen and Kets (2009) for traditional random intersection to address our model.

Lemma A.1. *Let $V = \{1, 2, \cdots, n\}$ be the set of vertices and $W = \{a_1, a_2, \cdots, a_m\}$ be the set of attributes. For $p_n \in [0, 1]$, vertices in V have every attribute independently with probability p_n. Let E_{ij} be the event that two vertices i, j share at least one common attribute. For three distinct vertices $i, j, k \in V$, denote by E_{ijk} the event that there is at least one attribute that is shared by vertices i, j, k. Write $E_{ij,ik,jk}$ for the event that there are at least 3 distinct attributes shared by i and j, i and k, and j and k respectively. Similarly, the event that there are two distinct attributes shared by vertices i and k, and j and k respectively is denoted as $E_{ik,jk}$. Assume that $mp_n^{3/2} = o(1)$. Then for any three distinct vertices $i, j, k \in V$, we have*

(a) $P(E_{ijk}) = 1 - (1 - p_n^3)^m = mp_n^3 + O(m^2 p_n^6)$

(b) $P(E_{ij,ik,jk}) = m^3 p_n^6 + O(m^4 p_n^8)$

(c) $P(E_{ik,jk}) = m^2 p_n^4 + O(m^3 p_n^6)$

(d) $P(E_{ijk}E_{ik,jk}) = m^2 p_n^5 + O(m^3 p_n^7)$

Proof. Under the assumption that $mp_n^{3/2} = o(1)$, we have $mp_n^a \leq mp_n^{3/2} = o(1), \forall a \geq 3/2$.

As for (a), $P(E_{ijk}) = 1 - (1-p_n^3)^m = mp_n^3 + O(m^2 p_n^6)$.

To prove (b), note that the probability that there is exactly one attribute that is shared by both i and j is $mp_n^2(1-p_n^2)^{m-1} = mp_n^2 + O(m^2 p_n^4)$. Given i and j share one attribute, the probability that i and k share exactly one of the other $m-1$ attribute is $(m-1)p_n^2(1-p_n^2)^{m-2} = mp_n^2 + O(m^2 p_n^4)$. Finally, conditional probability that there is a third group to which j and k belong given that i, j and i, k share one attribute is that $1-(1-p_n^2)^{m-2} = mp_n^2 + O(m^2 p_n^4)$. Combining these estimates, and noting that scenarios in which i, j or i, k share more than one attribute have negligible probability in comparison. We get that

$$P(E_{ij,ik,jk}) = m^3 p_n^6 + O(m^4 p_n^8)$$

Part(c) can be derived analogously to Part (b).

As for (d), note that the event $E_{ijk}E_{ik,jk}$ occurs when there is at least one attribute that is shared by all three vertices i, j, k and a second group shared by either i, k or j, k. Denote by r the probability that vertex k and at least one of the vertices i and j belong to a fixed group. Then $r = p_n(2p_n - p_n^2)$. Conditional on that there is exactly one attribute shared by i, j, k (the probability of this is $mp_n^3(1-p_n^3)^{m-1} = mp_n^3 + O(m^2 p_n^6)$), the probability that there is at least one other attribute that is shared either by i, k

or j, k is $1-(1-r)^{m-1} = mr + O(m^2r^2)$. It follows that

$$(E_{ijk}E_{ik,jk}) = (mp_n^3 + O(m^2p_n^6))(mr + O(m^2r^2)) = m^2p^5 + O(m^3p_n^7)$$

This finishes the proof of the Lemma A.1. ■

The conditional probability can be represented below.

$$\begin{aligned} P(A_{ij}=1|A_{ik}=1, A_{jk}=1) &= \frac{P(A_{ij}=1, A_{ik}=1, A_{jk}=1)}{P(A_{ik}=1, A_{jk}=1)} = \frac{P(E_{ij}, E_{ik}, E_{jk})q_n^3}{P(E_{ik}, E_{jk})q_n^2} \\ &= \frac{P(E_{ijk} \cup E_{ij,ik,jk})}{P(E_{ijk} \cup E_{ik,jk})} q_n \end{aligned} \tag{A.1}$$

lower bound:

$$\begin{aligned} \frac{P(E_{ijk} \cup E_{ij,ik,jk})}{P(E_{ijk} \cup E_{ik,jk})} &\geq \frac{P(E_{ijk})}{P(E_{ijk}) + P(E_{ik,jk})} \\ &= \frac{mp_n^3 + O(m^2p_n^6)}{mp_n^3 + O(m^2p_n^6) + m^2p_n^4 + O(m^3p_n^6)} \\ &= \frac{1 + O(mp_n^3)}{1 + mp_n(1 + O(mp_n^2))} = \frac{1}{1+mp_n}(1+o(1)) \end{aligned} \tag{A.2}$$

upper bound:

$$\begin{aligned} \frac{P(E_{ijk} \cup E_{ij,ik,jk})}{P(E_{ijk} \cup E_{ik,jk})} &\leq \frac{P(E_{ijk}) + P(E_{ij,ik,jk})}{P(E_{ijk}) + P(E_{ik,jk}) - P(E_{ijk}E_{ik,jk})} \\ &= \frac{mp_n^3 + O(m^2p_n^6) + m^3p_n^6 + O(m^4p_n^8)}{mp_n^3 + O(m^2p_n^6) + m^2p_n^4 + O(m^3p_n^6) - (m^2p_n^5 + O(m^3p_n^7))} \\ &= \frac{1 + m^2p_n^3 + O(mp_n^3)}{1 + mp_n + mp_n^2 + O(m^2p_n^3)} = \frac{1}{1+mp_n}(1+o(1)) \end{aligned} \tag{A.3}$$

The last equality is due to the assumption that $mp_n^{3/2} = o(1)$.

Combining the two bounds above gives us that

$$P(A_{ij}=1|A_{ik}=1,A_{jk}=1)=\frac{q_n}{1+mp_n}(1+o(1)) \tag{A.4}$$

We finish the proof.■

Appendix for Section 1.4

Lemma A.2 (Concentration on Bipartite Graph H)**.** *Let $N=m+n$ and Δ be the maximum degree of $\mathcal{H}$, that is $\Delta=\max\{\max_i\sum_{j=1}^m\mathcal{H}_{ij},\max_j\sum_{i=1}^n\mathcal{H}_{ij}\}$. For a given $\epsilon>0$, assume that $\Delta>\frac{4}{9}\log(2N/\epsilon)$, then*

$$\|H-\mathcal{H}\| \leq \sqrt{4\Delta\log(2N/\epsilon)} \tag{A.5}$$

Proof of Lemma A.2: *Proof*: Define the symmetrized version of H and $\mathcal{H}$ as follow:

$$\widetilde{H}=\begin{pmatrix}0 & H^T\\ H & 0\end{pmatrix},\ \widetilde{\mathcal{H}}=\begin{pmatrix}0 & \mathcal{H}^T\\ \mathcal{H} & 0\end{pmatrix}$$

Note the maximum expected degree of $\widetilde{\mathcal{H}}$ is Δ, Lemma A.13 gives the following result.

$$\|\widetilde{H}-\widetilde{\mathcal{H}}\|\leq\sqrt{4\Delta\log(2N/\epsilon)}. \tag{A.6}$$

For $\|x\|=1,\|y\|=1$,

$$x^T(H-\mathcal{H})y = \frac{1}{2}[x^T,y^T](\widetilde{H}-\widetilde{\mathcal{H}})[x^T,y^T]^T$$

$$\leq \frac{1}{2}\|(x^T, y^T)\|^2 \|\widetilde{H} - \widetilde{\mathcal{H}}\| \leq \sqrt{4\Delta \log(2N/\epsilon)}. \tag{A.7}$$

Therefore, $\|H - \mathcal{H}\| \leq \sqrt{4\Delta \log(2N/\epsilon)}$, which completes the proof. ■

Lemma A.3 (Bound on element-wise product)**.** *Given a matrix $A \in \mathbb{R}^{n\times n}$, $u, v \in R^{n\times 1}$, then*

$$\|A \cdot uv^T\|_2 \leq u_{\max} v_{\max} \|A\|_2, \tag{A.8}$$

where $u_{\max} = \max_i |u_i|, v_{\max} = \max_i |v_i|$.

Proof of Lemma A.3:

$$\begin{aligned}
\|A \cdot uv^T\|_2 &= \max_{x,y:\|x\|=1,\|y\|=1} x^T(A \cdot uv^T)y = \max_{x,y:\|x\|=1,\|y\|=1} A_{ij}u_i v_j x_i y_j \\
&= \max_{x,y:\|x\|=1,\|y\|=1} (diag(u)x)^T A (diag(v)y) \\
&\leq \max_{x,y:\|x\|=1,\|y\|=1} \|A\| \, \|diag(u)\| \|x\| \, \|diag(v)\| \|y\| \\
&= u_{\max} v_{\max} \|A\|_2.
\end{aligned}$$

Here $diag(v)$ denotes the diagonal matrix with $diag(v)_{ii} = v_i$.

Lemma A.4 (Effect of Truncation)**.** *Assume that $H \in \{0,1\}^{n\times m}$ is sampled from probability matrix $\mathcal{H} = p_n \Theta Z \mathbf{M} Y^T \Gamma$ as in (1.9). Assume for a given $\epsilon > 0$, sufficiently large n, that (i) the minimum row degree $\delta_1 > 12\log(16n/\epsilon)$ and (ii) $\max_{|I|=2\Delta_1} \sum_{k\in I} \mathcal{H}_{jk} \leq 1$ and (iii) $\Delta_1 \max_{|I|=2\Delta_1} \sum_{j=1}^n \sum_{k\in I} \mathcal{H}_{jk}^2 \leq \frac{2}{9}\log(8n/\epsilon)$. Then*

the following holds with probability at least $1-\epsilon/4$,

$$\|HH^T - \mathtt{t}(HH^T)\| \leq 4\Delta_1 \log(8n/\epsilon). \tag{A.9}$$

Proof of Lemma A.4: Since $HH^T - \mathtt{t}(HH^T)$ is a symmetric nonnegative matrix, we have

$$\|HH^T - \mathtt{t}(HH^T)\| \leq \|HH^T - \mathtt{t}(HH^T)\|_1 = \max_{1\leq i\leq n} \sum_{j=1}^{n} [H_i H_j^T - \mathtt{t}(H_i H_j^T)].$$

For a fixed i, let $d_i = \sum_{j=1}^{m} H_{ij}$ denote row degree of bipartite graph H, and $t_i = \sum_j \mathcal{H}_{ij}$ denote the row degree of the population bipartite graph $\mathcal{H}$. we will show that d_i will concentrate around t_i. We can apply one-sided concentration inequality on sum of independent Bernoulli variables, and have

$$P(d_i - t_i > bt_i) \leq e^{-\frac{b^2 t_i^2}{2(t_i + bt_i/3)}}, \text{ for any } b > 0. \tag{A.10}$$

Let $b = \sqrt{\frac{3\log(8n/\epsilon)}{t_i}} < 1$, we have $e^{-\frac{b^2 t_i^2}{3t_i}} \leq \epsilon/(8n)$. That is,

$$P(d_i < 2t_i) \geq 1 - \frac{\epsilon}{8n}. \tag{A.11}$$

Let row vectors $H_i, \mathcal{H}_i \in R^{1\times m}$ be the i-th row of H and $\mathcal{H}$ respectively. Let $X_j = H_i H_j^T, j \neq i$, and $Y_j = X_j - \mathtt{t}(X_j)$. Let $I = \{k \in [m] : H_{ik} = 1\}$ and $|I| = u$. Conditional on the set I, $\sum_{j\neq i}(X_j - \mathtt{t}(X_j))$ is a summand of $n-1$ independent random variables. The Bernstein concentration inequality for sum of bounded independent

variables has the following expression:

$$P\left(\sum_{j\neq i} Y_j - \mathbf{E}\sum_{j\neq i} Y_j > \lambda \middle| I\right) \leq e^{-\frac{\lambda^2}{2(v^2+M\lambda/3)}}, \text{ for any } \lambda > 0. \tag{A.12}$$

where $|Y_j| \leq M = u$ and $v^2 = \sum_{j\neq i} Var(Y_j)$. Furthermore

$$\begin{aligned}
v^2 &= \sum_{j\neq i} Var(Y_j) \leq \sum_{j\neq i} E(Y_j^2) = \sum_{j\neq i} E(X_j - \mathtt{t}(X_j))^2 \\
&= \sum_{j\neq i} \{E[(X_j)^2] - 2E[X_j\,\mathtt{t}(X_j)] + E[\mathtt{t}(X_j)^2]\} \\
&= \sum_{j\neq i} [\sum_{k\in I} \mathcal{H}_{jk} + \sum_{k\neq k'} \mathcal{H}_{jk}\mathcal{H}_{jk'} - 2\sum_{k\in I} \mathcal{H}_{jk} + 1 - \prod_{k\in I}(1-\mathcal{H}_{jk})] \\
&\leq \sum_{j\neq i} [\sum_{k\neq k'} \mathcal{H}_{jk}\mathcal{H}_{jk'} - \eta_j + (\eta_j - \sum_{k<k'} \mathcal{H}_{jk}\mathcal{H}_{jk'} + \sum_{k<k'<k''} \mathcal{H}_{jk}\mathcal{H}_{jk'}\mathcal{H}_{jk''} + \cdots).] \\
&\leq \sum_{j\neq i} \eta_j^2/2 + \sum_{j\neq i} \eta_j^3/3! + \sum_{j\neq i} \eta_j^4/4! + \cdots + \sum_{j\neq i} \eta_j^u/u! \leq \sum_j \eta_j^2,
\end{aligned}$$

where $\eta_j = \sum_{k\in I} \mathcal{H}_{jk}$. This comes from the fact $\eta_j \leq 1$ when $u < 2t_i$ and assumption (ii) holds. For $u > 0$, let

$$\psi_u \triangleq \max_{|I|=u} \sum_{j=1}^{n} \sum_{k\in I} \mathcal{H}_{jk}^2 \tag{A.13}$$

It is easy to see that ψ_u is an increasing function of u.

$$\sum_j \eta_j^2 \leq \sum_{j=i} u \sum_{k\in I} \mathcal{H}_{jk}^2 = u\psi_u \tag{A.14}$$

Let $\lambda = 4/3u\log(8n/\epsilon)$, we have $v^2 \leq u\psi_u \leq u\lambda/3$ when $u \leq 2t_i$ and assumption (iii)

holds. Then

$$P\left(\sum_{j\neq i} Y_j - \mathbf{E}\sum_{j\neq i} Y_j > \frac{4}{3}u\log\frac{8n}{\epsilon}\middle| I\right) \leq e^{-\frac{\lambda^2}{2(v^2+u\lambda/3)}} \leq e^{-\frac{\lambda^2}{4u\lambda/3}}$$
$$\leq e^{-\frac{(4/3u\log(8n/\epsilon))^2}{4/3u\,4/3u\log(8n/\epsilon)}} \leq \frac{\epsilon}{8n} \quad \text{(A.15)}$$

$$\begin{aligned}\mathbf{E}\left[\sum_{j\neq i} Y_j \middle| I\right] &= \sum_{j\neq i}[\mathbf{E}(X_j) - \mathbf{E}\mathtt{t}(X_j)] = \sum_{j\neq i}(\eta_j - [1-\prod_{k\in I}(1-\mathcal{H}_{jk})]) \\ &\leq \sum_{j\neq i}(\eta_j^2/2 + \eta_j^3/3! + \cdots) \\ &\leq \sum_{j\neq i}\eta_j^2 \leq \psi_u u. \end{aligned} \quad \text{(A.16)}$$

So combing the two inequality above gives

$$P\left(\sum_{j\neq i} Y_j > u\psi_u + \frac{4}{3}u\log\frac{8n}{\epsilon}\middle| I\right) \leq \frac{\epsilon}{8n} \quad \text{(A.17)}$$

Since $Y_i = X_i - \mathtt{t}(X_i) = u - 1$, we have, for any $u \leq 2t_i$

$$P\left(\sum_{j=1}^{n} Y_j > u + u\psi_u + \frac{4}{3}u\log\frac{8n}{\epsilon}\middle| |I| = u\right) \leq \frac{\epsilon}{8n} \quad \text{(A.18)}$$

Therefore, for i, under the assumption that (i) $t_i \geq \delta_1 \geq 3\log(8n/\epsilon)$, (ii) and (iii), we combine result in (A.11) and (A.18) to get the unconditional probability,

$$P\left(\sum_{j=1}^{n}[H_iH_j^T - \mathtt{t}(H_iH_j^T)] > 2t_i + 2t_i\psi_{2t_i} + \frac{8}{3}t_i\log\frac{8n}{\epsilon}\right) \leq \frac{\epsilon}{8n} + \frac{\epsilon}{8n} \leq \frac{\epsilon}{4n} \quad \text{(A.19)}$$

Furthermore, under the assumptions (i), (ii) and (iii), the union bound gives us that for all i's,

$$P\left(\max_i \sum_{j=1}^{n}[H_iH_j^T - \mathtt{t}(H_iH_j^T)] > \Delta_1(2 + 4\Delta_1\psi_{2\Delta_1} + \frac{8}{3}\log\frac{8n}{\epsilon})\right) \leq \epsilon/4 \quad \text{(A.20)}$$

Since under assumption (iii), we have $\Delta_1\psi_{2\Delta_1} \leq \frac{2}{9}\log(8n/\epsilon)$. Therefore, for sufficiently large n, we have with probability at least $1 - \epsilon/4$,

$$\max_i \sum_{j=1}^{n}[H_iH_j^T - \mathtt{t}(H_iH_j^T)] \leq 4\Delta_1\log(8n/\epsilon). \quad \text{(A.21)}$$

Proof of Theorem 1.14. First, we will bound the following term in the second term above,

$$\|\mathtt{t}(HH^T) - \mathcal{H}\mathcal{H}\| \leq \|\mathtt{t}(HH^T) - HH^T\| + \|HH^T - \mathcal{H}\mathcal{H}\| \quad \text{(A.22)}$$

Under Assumption (2), Lemma A.2 gives with probability $> 1 - \epsilon/4$

$$\|H - \mathcal{H}\| \leq \sqrt{4\max\{\Delta_1, \Delta_2\}\log(8N/\epsilon)}. \quad \text{(A.23)}$$

Also, we have

$$\|\mathcal{H}\| \leq \sqrt{\|\mathcal{H}\|_1\|\mathcal{H}\|_\infty} \leq \sqrt{\Delta_1\Delta_2}, \quad \text{(A.24)}$$

Therefore, with probability at least $1-\epsilon/4$, we have

$$\begin{aligned}\|HH^T-\mathcal{H}\mathcal{H}\| &= \|HH^T-\mathcal{H}H^T\|+\|\mathcal{H}H^T-\mathcal{H}\mathcal{H}\| = \|H-\mathcal{H}\|\,\|H^T\|+\|\mathcal{H}\|\,\|H^T-\mathcal{H}\| \\ &\le 2\|H-\mathcal{H}\|\,\|\mathcal{H}^T\|+\|H-\mathcal{H}\|^2 \\ &\le 4\sqrt{\max\{\Delta_1,\Delta_2\}\log(8N/\epsilon)}\,\sqrt{\Delta_1\Delta_2}+4\max\{\Delta_1,\Delta_2\}\log(8N/\epsilon) \quad \text{(A.25)}\end{aligned}$$

Furthermore, combining result from Lemma A.4 and results above, we have with probability $1-\epsilon/2$,

$$\begin{aligned}\|\mathtt{t}(HH^T)-\mathcal{H}\mathcal{H}\| &= \|HH^T-\mathcal{H}\mathcal{H}\|+\|\mathtt{t}(HH^T)-HH^T\| \\ &\le 4\sqrt{\max\{\Delta_1,\Delta_2\}\log(8N/\epsilon)}\,\sqrt{\Delta_1\Delta_2}+4\max\{\Delta_1,\Delta_2\}\log(8N/\epsilon) \\ &\quad +4\Delta_1\log(8n/\epsilon) \\ &= \Delta_1\Delta_2\left(4\sqrt{\frac{\log(8N/\epsilon)}{\min\{\Delta_1,\Delta_2\}}}+\frac{4\log(8N/\epsilon)}{\min\{\Delta_1,\Delta_2\}}+\frac{4\log(8n/\epsilon)}{\Delta_2}\right) \\ &\le 6\Delta_1\Delta_2\sqrt{\frac{\log(8N/\epsilon)}{\min\{\Delta_1,\Delta_2\}}} \quad \text{(A.26)}\end{aligned}$$

Under the assumption that $\min\{\Delta_1,\Delta_2\} > 16\log(8N/\epsilon)$, we have $4\sqrt{\frac{\log(8N/\epsilon)}{\min\{\Delta_1,\Delta_2\}}} \le 1$. Finally we note the fact that

$$Z\mathbf{B}Z^T = \sum_{s,t\in[K]} \mathbf{B}_{st} Z_{\cdot s} Z_{\cdot t}^T$$

where $Z_{\cdot s}, Z_{\cdot t} \in \{0,1\}^{n\times 1}$ are the s-th, t-th columns of Z. Lemma A.3 implies that

$$\|A\cdot q_n Z\mathbf{B}Z^T\|_2 \le q_n \sum_{s,t}\mathbf{B}_{st}\|A\| \le q_n K^2\|A\|. \quad \text{(A.27)}$$

If the off-diagonal of $\mathbf{B}$ are the same, say equal to $0 < \rho < 1$ and $\min_i \mathbf{B}_{ii} > \rho$, then $\mathbf{B} = \rho J_K + (\mathbf{B} - \rho J_K)$. $(\mathbf{B} - \rho J_K)$ is block-diagonal and $\|A \cdot Z(\mathbf{B} - \rho J_K)Z^T\| \leq (1-\rho)\|A\|$. Combinding with $\|A \cdot Z(\rho J_K)Z^T\| \leq \rho\|A\|$, we have that $\|A \cdot q_n Z\mathbf{B}Z^T\|_2 \leq q_n\|A\|$.

Substituting A with $\mathtt{t}(HH^T) - \mathcal{H}\mathcal{H}$ and combining (A.26), we have with probability at least $1 - \epsilon/2$,

$$\|\mathcal{A}_H - \mathcal{A}\| \leq 6K^2 q_n \sqrt{\Delta_1 \Delta_2} \sqrt{\max\{\Delta_1, \Delta_2\} \log(8N/\epsilon)}. \tag{A.28}$$

If the off-diagonal of $\mathbf{B}$ are the same, say equal to $0 < \rho < 1$ and $\min_i \mathbf{B}_{ii} > \rho$, the K^2 term can be removed. This completes the proof. ■ □

Lemma A.5. *Assume that the assumptions (1) - (3) in Theorem 1.14 hold. Then we have that with probability* $1 - \epsilon/4$,

$$d_{\max}(\mathcal{A}_H) \leq 3\Delta_1 \Delta_2 q_n. \tag{A.29}$$

Proof of Lemma A.5:

Next, we will prove the upper bound with the similar technique as used in the proof of Lemma A.4. First, for a given i, let $d_i = \sum_j H_{ij}$ and $t_i = \sum_j \mathcal{H}_{ij}$. We will show that d_i will concentrate around t_i. We can apply one-sided concentration inequality on sum of independent Bernoulli variables, and have

$$P(|d_i - t_i| > bt_i) \leq 2e^{-\frac{b^2 t_i^2}{2(t_i + bt_i/3)}}, \text{ for any } b > 0. \tag{A.30}$$

Let $b = \sqrt{\frac{3\log(16n/\epsilon)}{t_i}}$, then $b < 1/2$ since $t_i > \delta_1 \geq 12\log(16n/\epsilon)$, we have $2e^{-\frac{b^2 t_i^2}{3t_i}} \leq \epsilon/(8n)$. That is,

$$P(t_i/2 < d_i < 2t_i) \geq 1 - \frac{\epsilon}{8n}. \tag{A.31}$$

Second, let $Y_j \triangleq H_i H_j^T$, where $H_i, H_j \in \{0,1\}^{1\times m}$ are the i-th and j-th rows in the H. $I = \{k \in [m] : H_{ik} = 1\}$ and $|I| = u$. Conditional on I, $\sum_{j\neq i} Y_j$ is a summand of $n-1$ independent variables. The Bernstein concentration for sum of bounded independent variables has the following expression:

$$P\left(\sum_{j\neq i} Y_j - \mathbf{E}\sum_{j\neq i} Y_j > \lambda \middle| I\right) \leq e^{-\frac{\lambda^2}{2(v^2 + M\lambda/3)}}, \text{ for any } \lambda > 0. \tag{A.32}$$

where $|Y_j| \leq M = u$ and

$$\begin{aligned} v^2 &= \sum_{j\neq i} Var(Y_j) = \sum_{j\neq i}\sum_{k\in I} \mathcal{H}_{jk}(1 - \mathcal{H}_{jk}) \\ &\leq \sum_{j\neq i}\sum_{k\in I} \mathcal{H}_{jk} \leq \sum_{k\in I}\sum_{j\neq i} \mathcal{H}_{jk} \leq u\Delta_2, \end{aligned}$$

where $\Delta_2 = \max_j \sum_k \mathcal{H}_{jk}$.

$$\begin{aligned} & P\left(\sum_{j\neq i} Y_j - \mathbf{E}\sum_{j\neq i} Y_j > \sqrt{4u\Delta_2 \log(8n/\epsilon)} \middle| I\right) \\ \leq\ & e^{-\frac{\lambda^2}{2(v^2 + u\lambda/3)}} \leq e^{-\frac{\lambda^2}{2(u\Delta_2 + u\lambda/3)}} \\ \leq\ & e^{-\frac{\lambda^2}{2(2t_i\Delta_2 + 2t_i\lambda/3)}} \end{aligned} \tag{A.33}$$

Find λ through finding the solution $\frac{\lambda^2}{2(2t_i\Delta_2+2t_i\lambda/3)} = \log(8n/\epsilon)$. We the positive solution

$$\begin{aligned}\lambda &= \frac{4t_i/3\log(8n/\epsilon) + \sqrt{(4t_i/3\log(8n/\epsilon))^2 + 16t_i\Delta_2\log(8n/\epsilon)}}{2} \\ &\leq 4/3t_i\log(8n/\epsilon) + 2\sqrt{t_i\Delta_2\log(8n/\epsilon)}\end{aligned}$$

From (A.32), we have

$$\mathbf{E}\left[\sum_{j\neq i} Y_j \middle| I\right] = \sum_{k\in I}\sum_{j\neq i}\mathcal{H}_{jk} \leq u\Delta_2 \tag{A.34}$$

So combing the two inequality above gives that when $u \leq 2t_i$,

$$P\left(\sum_{j\neq i} Y_j > u\Delta_2 + 4/3t_i\log(8n/\epsilon) + 2\sqrt{t_i\Delta_2\log(8n/\epsilon)}\middle| I\right) \leq \frac{\epsilon}{8n} \tag{A.35}$$

Since $Y_i = X_i - \mathtt{t}(X_i) = u - 1$, we have

$$P\left(\sum_{j=1}^{n} Y_j > u + u\Delta_2 + 4/3t_i\log(8n/\epsilon) + 2\sqrt{t_i\Delta_2\log(8n/\epsilon)}\middle| I\right) \leq \frac{\epsilon}{8n} \tag{A.36}$$

Therefore, for a fixed i, if $t_i > 3\log(8n/\epsilon)$, we combine result in (A.31) and (A.36) to get the unconditional probability,

$$P\left(\sum_{j=1}^{n}[H_iH_j^T - \mathtt{t}(H_iH_j^T)] > 2t_i + 2t_i\Delta_2 + 4/3t_i\log(8n/\epsilon) + 2\sqrt{t_i\Delta_2\log(8n/\epsilon)}\right) \leq \frac{\epsilon}{8n} + \frac{\epsilon}{8n} \leq \frac{\epsilon}{4n} \tag{A.37}$$

Furthermore, if i) $\delta_1 > 12\log(16n/\epsilon)$ the equation above will holds for every i. Using

the union bound to get bound universal to all i's at the same time,

$$P\left(\max_i \sum_{j=1}^{n} H_i H_j^T > 2\Delta_1 + 2\Delta_1\Delta_2 + 4/3\Delta_1 \log(8n/\epsilon) + 2\sqrt{\Delta_1\Delta_2 \log(8n/\epsilon)}\right) \leq \epsilon/4 \quad \text{(A.38)}$$

Additionally, we have $2\Delta_1 + 4/3\Delta_1 \log(8n/\epsilon) + 2\sqrt{\Delta_1\Delta_2 \log(8n/\epsilon)} \leq \Delta_1\Delta_2$ from assumption that $\min\{\Delta_1, \Delta_2\} > 16 \log(8n/\epsilon)$ in Assumption (3). Thus we have, with probability at least $1 - \epsilon/4$, that

$$\sum_{j=1}^{n} H_i H_j^T \leq 3\Delta_1\Delta_2, \forall i. \quad \text{(A.39)}$$

$$\max_i \sum_{j=1}^{n} (\mathcal{A}_H)_{ij} = \max_i \sum_j \mathtt{t}(H_i H_j^T) q_n \mathbf{B}_{z_i z_j} \leq q_n \max_i \sum_j H_i H_j^T \leq 3 q_n \Delta_1 \Delta_2. \quad \text{(A.40)}$$

We complete the proof. ■

Proof of Theorem 1.17: Lemma A.13 gives

$$Pr\left(\|A - \mathcal{A}_H\| > \sqrt{4 \max\{d_{\max}(\mathcal{A}_H), \frac{4}{9}\log(8N/\epsilon)\} \log(8N/\epsilon)} \,\middle|\, H\right) \leq \epsilon/4 \quad \text{(A.41)}$$

Lemma A.5 above states that, under Assumption (4) that $3\Delta_1\Delta_2 q_n > \frac{4}{9}\log(8n/\epsilon)$, we have

$$Pr\left(\|A - \mathcal{A}_H\| > \sqrt{12\Delta_1\Delta_2 q_n \log(8N/\epsilon)}\right) \leq \epsilon/2 \quad \text{(A.42)}$$

Combining results in Theorem 1.14 and Equation (A.42) we have with probability

$1-\epsilon$,

$$\begin{aligned}\|A-\mathcal{A}\| &\leq \|\mathcal{A}_H-\mathcal{A}\|+\|A-\mathcal{A}_H\| \\ &\leq 6K^2 q_n\sqrt{\Delta_1\Delta_2}\sqrt{\max\{\Delta_1,\Delta_2\}\,\log(8N/\epsilon)}+\sqrt{12\,\Delta_1\Delta_2 q_n\log(2N/\epsilon)}\end{aligned}$$

Here we finish the proof. ■ □

Lemma A.6. *For two positive semidefinite matrices $A, B \in \mathbb{R}^{d\times d}$, we have $A\cdot B$ is positive semidefinite.*

Lemma A.7. *For two non-zero vectors u,v of the same dimension, we have*

$$\|\frac{u}{\|u\|}-\frac{v}{\|v\|}\|\leq 2\frac{\|u-v\|}{\max(\|u\|,\|v\|)}. \tag{A.43}$$

Proof of Theorem 1.18. Use Davis-Kahan's lemma, we have that there exists a $K\times K$ orthonormal matrix $\mathcal{O}$ such that

$$\|X-\mathcal{X}\mathcal{O}\|_F\leq\frac{2\sqrt{2K}}{\lambda_K}\|A-\mathcal{A}\| \tag{A.44}$$

Applying Lemma A.7 gives us that

$$\begin{aligned}\|X^*-\mathcal{X}^*\mathcal{O}\|_F^2 &= \sum_{i=1}^{n}\|X_i-\mathcal{X}_i\|_F^2 \\ &\leq \sum_{i=1}^{n}\frac{4}{r^2}\|X_i-\mathcal{X}_i\|_F^2\leq\frac{4}{r^2}\|X-\mathcal{X}\|_F^2.\end{aligned} \tag{A.45}$$

Therefore, we have

$$\|X^* - \mathcal{X}^*\mathcal{O}\|_F \leq \frac{4\sqrt{2K}}{r\,\lambda_K}\|A - \mathcal{A}\| \tag{A.46}$$

□

Proof of Theorem 1.20. Recall the eigen-structure of $\mathcal{A}$,

$$\mathcal{X} = \Theta Z(Z^T\Theta^2 Z)^{-1/2}U, \quad \text{and } \mathcal{X}^* = ZU. \tag{A.47}$$

So the i-row of $\mathcal{X}$, $\mathcal{X}_i = \frac{\theta_i}{\sqrt{\sum_i \theta_i^2}}U_{z_i}$, and the i-th row of $\mathcal{X}^*$, $\mathcal{X}_i^* = U_{z_i}$.
Based on the Theorem 1.18, we have

$$\|X^* - \mathcal{X}^*\mathcal{O}\|_F^2 \leq \frac{32K\|A - \mathcal{A}\|^2}{r^2\lambda_K^2}. \tag{A.48}$$

Let $\mathcal{U} = \{i : \|c_i - Z_iU\mathcal{O}\| \geq \frac{1}{\sqrt{2}}\}$. Since U is an orthonormal matrix, the two distinct rows in U will have distance equal to $\sqrt{2}$. If node i is mis-clustered, i.e. $i \in \mathcal{M}$, then $\|c_i - Z_iU\mathcal{O}\| > \frac{1}{\sqrt{2}}$. Therefore, $\mathcal{M} \subset \mathcal{U}$.

$$\begin{aligned}
\frac{|\mathcal{M}|}{n} \leq \frac{|\mathcal{U}|}{n} &\leq \frac{1}{n}\sum_{i=1}^{n} 1_{\left(\{i:\|c_i - Z_iU\mathcal{O}\| \geq \frac{1}{\sqrt{2}}\}\right)} \leq \frac{1}{n}\sum_{i:\|c_i - Z_iU\mathcal{O}\| \geq \frac{1}{\sqrt{2}}} \|c_i - Z_iU\mathcal{O}\|_2^2 * 2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|c_i - Z_iU\mathcal{O}\|_2^2 * 2 = \frac{1}{n}\|\widehat{C} - \mathcal{X}^*\mathcal{O}\|_F^2 * 2, \quad \widehat{C} \in R^{n\times K} \text{ with } \widehat{C}_i = c_i \\
&\leq \frac{4}{n}(\|\widehat{C} - X^*\mathcal{O}\|_F^2 + \|X^* - \mathcal{X}^*\mathcal{O}\|_F^2) \leq \frac{8}{n}\|X^* - \mathcal{X}^*\mathcal{O}\|_F^2 \quad (*) \\
&\leq \frac{256K\|A - \mathcal{A}\|^2}{nr^2\lambda_K^2} \quad (**)
\end{aligned}$$

$$\leq \frac{256K}{nr^2}\frac{1}{\lambda_K^2}\left(6K^2 q_n\sqrt{\Delta_1\Delta_2}\sqrt{\max\{\Delta_1,\Delta_2\}\,\log(8N/\epsilon)}+\sqrt{12\,\Delta_1\Delta_2 q_n\log(2N/\epsilon)}\right)^2$$
$$\leq c\frac{K}{nr^2}\frac{1}{\lambda_K^2}\left(K^4 q_n^2\,\Delta_1\Delta_2\max\{\Delta_1,\Delta_2\}\,\log(8N/\epsilon)+\Delta_1\Delta_2 q_n\log(2N/\epsilon)\right) \tag{A.49}$$

The inequality in (*) above comes from that both $\widehat{C}$ and $\mathcal{X}^*$ has K distinct rows and $\widehat{C}$ assumed to be the global minimizer of the following k-mean problem.

$$\widehat{C} = \underset{M\in R^{n\times K}, M \text{ has } K \text{ distinct rows}}{\operatorname{argmin}} \|X^* - M\|_F^2 \tag{A.50}$$

The inequality in (**) comes from the Theorem 1.17. So we complete the proof. ■ □

Proof of Corollary 1.22. Under this simplified case, $\mathcal{A} = Z\widetilde{\mathbf{B}}Z^T$ with $\widetilde{\mathbf{B}}$ as follows:

$$\widetilde{\mathbf{B}} = mp_n^2 q_n \begin{bmatrix} a & b & \cdots & b \\ b & a & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & a \end{bmatrix}. \tag{A.51}$$

where $a = \frac{1}{K} + \frac{K-1}{K}\eta^2$ and $b = (\frac{2}{K}\eta + \frac{K-2}{K}\eta^2)\rho$. The eigenvalues of $\mathcal{A}$ are the same as those of $(Z^TZ)^{1/2}\widetilde{\mathbf{B}}(Z^TZ)^{1/2} = \frac{n}{K}\widetilde{\mathbf{B}}$. Through some computation, we have $\lambda_1 = mnp_n^2q_n(\frac{1}{K}a + \frac{K-1}{K}b)$ and $\lambda_2 = \lambda_3 = \cdots = \lambda_K = \frac{mnp_n^2q_n}{K}(a-b)$. Also we have $\Delta_1 = np_n(\frac{1}{K} + \frac{K-1}{K}\eta), \Delta_2 = mp_n(\frac{1}{K} + \frac{K-1}{K}\eta), nr^2 = K$. With probability at least $1-\epsilon$, we have

$$\frac{|\mathcal{M}|}{n} \leq c_1\left(\frac{K^2\log(16N/\epsilon)}{\min\{m,n\}\,p_n} + \frac{K^2\log(16N/\epsilon)}{mnp_n^2q_n}\right) \tag{A.52}$$

where c_1 is a constant depend on η, ρ, and c_0 in (1.18). ■

Some Lemmas

Lemma A.8 (Davis-Kahan lemma, Lemma 5.1 from Lei and Rinaldo (2013)). *Assume that A and $\mathcal{A}$ are symmetric. Let the rank of $\mathcal{A}$ be K and λ_K be the K-th leading singular value of $\mathcal{A}$. Let $X, \mathcal{X} \in R^{n\times K}$ be the matrices with the leading K singular vectors of A and $\mathcal{A}$ as the columns. Then there exists a $K \times K$ orthonormal matrix $\mathcal{O}$ such that*

$$\|X - \mathcal{X}\mathcal{O}\|_F \leq \frac{2\sqrt{2K}}{\lambda_K}\|A - \mathcal{A}\| \tag{A.53}$$

Lemma A.9 (Chung' theorem 8). *Suppose X_i are independent random variables, satisfying $X_i \leq M$, $i = 1, 2, \ldots \cdots n$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathbf{E}(X_i^2)}$. Then we have*

$$P(X \geq EX + \lambda) \leq e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}$$

Lemma A.10 (Chung' theorem 9). *Suppose X_i are independent random variables, satisfying $X_i \geq -M$, $i = 1, 2, \ldots \cdots n$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathbf{E}(X_i^2)}$. Then we have*

$$P(X \leq EX - \lambda) \leq e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}$$

Lemma A.11 (Chung' theorem 8, 9). *Suppose X_i are independent random variables, satisfying $|X_i| \leq M$, $i = 1, 2, \ldots \cdots n$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathbf{E}(X_i^2)}$. Then we have*

$$P(|X - EX| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}$$

Assume that $\|X\|^2 \geq \frac{4}{9}\log(2n/\epsilon)$. *Let* $\lambda = \sqrt{4\|X\|^2 \log(2n/\epsilon)}$. *Then,*

$$P(|X - \mathbf{E}X| \geq \sqrt{4\|X\|^2 \log(2n/\epsilon)}) \leq \frac{\epsilon}{n}$$

Lemma A.12 (Matrix Concentration, see Theorem 5 in Chung and Radcliffe (2011))**.** *Let* X_1 , X_2, ... X_m *be independent random* $N \times N$ *Hermitian matrices. Assume that* $\|X_i - \mathbf{E}(X_i)\| \leq M$ *for all* i, *and* $v^2 = \|var(\sum_{i=1}^m X_i)\|$. *For* $a > 0$, *we have*

$$Pr(\|\sum_{i=1}^m X_i - \mathbf{E}\sum_{i=1}^m X_i\| > a) \leq 2N \exp\left(-\frac{a^2}{2v^2 + 2Ma/3}\right). \tag{A.54}$$

Based on this lemma, it is easy to derive the following result on random graphs as in Chung and Radcliffe (2011).

Lemma A.13 (Concentration on Random Graphs)**.** *Let* $\{A_{ij}, 0 \leq i < j \leq n\}$ *be independent Bernoulli random variables with success probability* p_{ij} *and* $A_{ji} = A_{ij}, j > i; A_{ij} = 0, i = j$. *Let* $\Delta = \max_i \sum_j p_{ij}$ *be the highest expected degree. If* $\Delta > \frac{4}{9}\log\frac{2N}{\epsilon}$, *the following holds with probability* $> 1 - \epsilon$,

$$\|A - \mathbf{E}A\| \leq \sqrt{4\Delta \log\frac{2N}{\epsilon}} \tag{A.55}$$

Even if $\Delta \leq \frac{4}{9}\log\frac{2N}{\epsilon}$, we have with probability $> 1 - \epsilon$,

$$\|A - \mathbf{E}A\| \leq \frac{4}{3}\log\frac{2N}{\epsilon} \tag{A.56}$$

Let $\Delta' = \max\{\Delta, \frac{4}{9}\log\frac{2N}{\epsilon}\}$, we always have, with probability $> 1-\epsilon$,

$$\|A - \mathbf{E}A\| \leq \sqrt{4\Delta' \log\frac{2N}{\epsilon}} \tag{A.57}$$

Appendix B

Appendix for Chapter 2

EM algorithm under SBM

The log likelihood is:

$$\begin{aligned}\ell(\mu,\pi|\widehat{X}) &= \sum_{i=1}^{n}\log\sum_{z_i}\pi_{z_i}N(\widehat{X}_i;\mu_{z_i},\mathbf{\Sigma}_{n,z_i}) = \sum_{i=1}^{n}\log\sum_{z_i}q_i(z_i)\frac{\pi_{z_i}N(\widehat{X}_i;\mu_{z_i},\mathbf{\Sigma}_{n,z_i})}{q_i(z_i)}\\ &\geq \sum_{i=1}^{n}\sum_{z_i}q_i(z_i)\log\frac{\pi_{z_i}N(\widehat{X}_i;\mu_{z_i},\mathbf{\Sigma}_{n,z_i})}{q_i(z_i)} \qquad \text{(B.1)}\end{aligned}$$

As in (2.16),

$$\begin{aligned}Q &\triangleq \sum_{i=1}^{n}\sum_{k=1}^{K}w_{ik}\log\frac{\pi_k N(\widehat{X}_i;\mu_k,\mathbf{\Sigma}_{n,k})}{w_{ik}}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}w_{ik}\Big(\log\pi_k-\frac{K}{2}\log(2\pi)-\frac{1}{2}\log(|\Sigma_{n,k}|)\\ &\quad -\frac{1}{2}(\widehat{X}_i-\mu_k)\Sigma_{n,k}^{-1}(\widehat{X}_i-\mu_k)-\log w_{ik}\Big) \qquad \text{(B.2)}\end{aligned}$$

Here $w_{ik} = q_i(k), k = 1, 2, \cdots, K$ with $\sum_{k=1}^{K} w_{ik} = 1$ are the intermediate variables.

EM algorithm for the original likelihood is equivalent to the coordinate ascent algorithm for J w.r.t to (w_{ik}) and (μ_k, π_k).

It is easy to tell that, given (π, μ), we will have following update:

$$w_{ik} = P(z_i = k|\alpha_n, \pi, \mu) \tag{B.3}$$

In terms of computation, for $i = 1, 2, \cdots, n$,

$$\begin{aligned} w_{ik} &\leftarrow \pi_k |\Sigma_{n,k}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\widehat{X}_i - \mu_k)\Sigma_{n,k}^{-1}(\widehat{X}_i - \mu_k)\right); \\ w_{ik} &\leftarrow \frac{w_{ik}}{\sum_{k=1}^{K} w_{ik}}, \quad k = 1, 2, \cdots, K. \end{aligned} \tag{B.4}$$

Add a Lagrange multiplier for the constraint that $\sum_{k=1}^{K} \pi_k = 1$ and take derivative w.r.t. π_k. We have the update for π_k as follows:

$$\pi_k \leftarrow \frac{1}{n}\sum_{i=1}^{n} w_{ik}, \quad k = 1, 2, \cdots, K. \tag{B.5}$$

The dependence relationship among Σ_k and μ_k is quite complicated. For simplification, assume that $\Sigma_{n,k}$ is free of μ_k when taking derivative of Q w.r.t. μ_k.

$$\frac{\partial Q}{\partial \mu_k} = \sum_{i=1}^{n} w_{ik}(\Sigma_{n,k}^{-1}\widehat{X}_i - \mu_k) = 0 \quad \Rightarrow \quad \mu_k \leftarrow \frac{\sum_{i=1}^{n} w_{ik}\widehat{X}_i}{\sum_{i=1}^{n} w_{ik}} \tag{B.6}$$

This should be a consistent estimator when the w_{ik} are accurate since they are essentially sample means.

We will maintain the dependence by computing $\Sigma_{n,k}$ using the updated μ_k and then use this to determine the posterior probabilities, i.e., w_{ik}'s. That is,

$$\Sigma_{n,k} \leftarrow \frac{1}{n}\Delta^{-1}\left(\sum_j \pi_j(\mu_k^T\mu_j)\mu_j\mu_j^T\right)\Delta^{-1}, \text{where } \Delta = \sum_j \pi_j\mu_j\mu_j^T. \tag{B.7}$$

Here we finish the derivation for Algorithm in Section 3.2.

The derivation for variational MLE for random degree

Recall the objective function expressed in (**??**). We have

$$\begin{aligned} Q(\alpha_i,\beta_i,w_i,\lambda,\boldsymbol{\mu},\pi) \;=\; & -n(\lambda\log(\lambda)-\log\Gamma(\lambda))+(\lambda-1)(\psi(\alpha_i)-\log(\beta_i))-\lambda\sum_{i=1}^n\frac{\alpha_i}{\beta_i} \\ & +\sum_{i=1}^n\sum_{k=1}^K w_{ik}\log\pi_k-\frac{K}{2}\sum_{i=1}^n(\psi(\alpha_i)-\log(\beta_i)) \\ & -\frac{1}{2}\sum_{i=1}^n\sum_{k=1}^K w_{ik}\log|\boldsymbol{\Sigma}_{n,k}|-\frac{1}{2}\sum_{i=1}^n\frac{\beta_i}{\alpha_i-1}\widehat{X}_i^T\sum_{k=1}^K w_{ik}\Sigma_{n,k}^{-1}\widehat{X}_i \\ & +\sum_{i=1}^n\sum_{k=1}^K w_{ik}\,\widehat{X}_i^T\Sigma_{n,k}^{-1}\mu_k-\frac{1}{2}\sum_{i=1}^n\frac{\alpha_i}{\beta_i}\sum_{k=1}^K w_{ik}\,\mu_k^T\boldsymbol{\Sigma}_{n,k}^{-1}\mu_k \\ & +\sum_{i=1}^n[\alpha_i-\log(\beta_i)+\log(\Gamma(\alpha_i))+(1-\alpha_i)\psi(\alpha_i)] \\ & -\sum_{i=1}^n\sum_{k=1}^K w_{ik}\log w_{ik}. \end{aligned} \tag{B.8}$$

We use the coordinate accent algorithm to update $(\alpha_i,\beta_i,w_i)_{i=1}^n$ and $(\lambda,\boldsymbol{\mu},\pi)$. For simplification and reduced overfitting we assume α_i's are given (default is set to 2).

$$\begin{aligned}\frac{\partial Q}{\partial \beta_i} &= \alpha_i(\lambda + \frac{1}{2}\sum_{k=1}^{K} w_{ik}\mu_k^T\Sigma_{n,k}^{-1}\mu_k)\frac{1}{\beta_i^2} + (\frac{K}{2}-\lambda)\frac{1}{\beta_i} - \frac{1}{2(\alpha_i-1)}\widehat{X}_i\sum_{k=1}^{K} w_{ik}\Sigma_{n,k}^{-1}\widehat{X}_i \\ &= -\frac{1}{\beta_i^2}(A_i\beta_i^2 - B_i\beta_i - C_i) = 0 \\ &\Rightarrow \beta_i \leftarrow \frac{B_i + \sqrt{B_i^2 + 4A_iC_i}}{2A_i}. \end{aligned} \tag{B.9}$$

where $A_i = \frac{1}{2(\alpha_i-1)}\widehat{X}_i\sum_{k=1}^{K} w_{ik}\Sigma_{n,k}^{-1}\widehat{X}_i$, $B_i = (\frac{K}{2}-\lambda)$ and $C_i = \alpha_i(\lambda + \frac{1}{2}\sum_{k=1}^{K} w_{ik}\mu_k^T\Sigma_{n,k}^{-1}\mu_k)$. A_i, C_i will be positive and B_i will be positive only when $\lambda < \frac{K}{2}$.

For w_{ik}, and for a fixed i, to maximize the following quantity over $w_i = [w_{i1}, \cdots, w_{iK}]$,

$$\sum_k w_{ik}\log\frac{\pi_k|\Sigma_{n,k}|^{-\frac{1}{2}}\exp(-\frac{1}{2}\frac{\beta_i}{\alpha_i-1}\widehat{X}_i\Sigma_{n,k}^{-1}\widehat{X}_i + \widehat{X}_i\Sigma_{n,k}^{-1}\mu_k - \frac{1}{2}\frac{\alpha_i}{\beta_i}\mu_k\Sigma_{n,k}^{-1}\mu_k)}{w_{ik}}$$

we will set

$$w_{ik} \propto \pi_k|\Sigma_{n,k}|^{-\frac{1}{2}}\exp(-\frac{1}{2}\frac{\beta_i}{\alpha_i-1}\widehat{X}_i\Sigma_{n,k}^{-1}\widehat{X}_i + \widehat{X}_i\Sigma_{n,k}^{-1}\mu_k - \frac{1}{2}\frac{\alpha_i}{\beta_i}\mu_k\Sigma_{n,k}^{-1}\mu_k). \tag{B.10}$$

Here comes the the algorithm: **E-step:** given $(\lambda, \pi, \boldsymbol{\mu})$, update (β_i, w_i). For simplicity, we assume that $\alpha_i > 1$ is a given constant, say 2).

$$\begin{aligned} A_i &= \frac{1}{2(\alpha_i-1)}\widehat{X}_i\sum_{k=1}^{K} w_{ik}\Sigma_{n,k}^{-1}\widehat{X}_i, \quad B_i = (\frac{K}{2}-\lambda); \\ C_i &= \alpha_i(\lambda + \frac{1}{2}\sum_{k=1}^{K} w_{ik}\mu_k^T\Sigma_{n,k}^{-1}\mu_k); \end{aligned}$$

$$\begin{aligned}
\beta_i &\leftarrow \frac{B_i + \sqrt{B_i^2 + 4A_iC_i}}{2A_i}. \\
w_{ik} &\leftarrow \pi_k |\Sigma_{n,k}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{\beta_i}{\alpha_i - 1}\widehat{X}_i\Sigma_{n,k}^{-1}\widehat{X}_i + \widehat{X}_i\Sigma_{n,k}^{-1}\mu_k - \frac{1}{2}\frac{\alpha_i}{\beta_i}\mu_k\Sigma_{n,k}^{-1}\mu_k \right). \\
w_{ik} &\leftarrow \frac{w_{ik}}{\sum_{k=1}^{K} w_{ik}}
\end{aligned}$$

Similarly, given $(\beta_i, w_i)_{i=1}^n$, we will update $(\lambda, \pi, \boldsymbol{\mu})$.

$$\frac{\partial Q}{\partial \lambda} = n\left(\log\lambda + 1 - \frac{\psi(\lambda)}{\Gamma(\lambda)}\right) + \sum_i (\psi(\alpha_i) - \log(\beta_i)) - \sum_{i=1}^{n}\frac{\alpha_i}{\beta_i} = 0. \quad \text{(B.11)}$$

One way is to update λ by solving this non-linear eqution of λ. Alternatively, use the fact that $Var(\theta_i) = \frac{1}{\lambda}$ and $\mathbf{E}(\theta_i|\alpha_i, \beta_i) = \frac{\alpha_i}{\beta_i}$, we use this updating procedure:

$$\lambda^{-1} \leftarrow \frac{1}{n}\sum_{i=1}^{n}(\frac{\alpha_i}{\beta_i})^2 - (\frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_i}{\beta_i})^2. \quad \text{(B.12)}$$

The update for π_k, μ_k are the similar to those in SBM.

$$\begin{aligned}
\pi_k &\leftarrow \frac{\sum_{i=1} w_{ik}}{n}; \\
\mu_k &\leftarrow \frac{\sum_{i=1}^{n} w_{ik}\widehat{X}_i}{\sum_{i=1}^{n}\frac{\alpha_i}{\beta_i}w_{ik}}.
\end{aligned} \quad \text{(B.13)}$$

Update $\Sigma_{n,k}$ by using the following formula as in the CLT result:

$$\boldsymbol{\Sigma}_{n,k} = \frac{1}{n}\Delta^{-1}\mathbf{E}[X_1X_1^T(\mu_k^T X_1)]\Delta^{-1}, \text{ where } \Delta = \mathbf{E}(X_1X_1^T). \quad \text{(B.14)}$$

One approach is to update $\boldsymbol{\Sigma}_{n,k}$ by substituting the population quantities above

with their sample versions:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i^T, \quad \mathbf{E}[X_1 X_1^T (\mu_k^T X_1)] = \frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i^T (\mu_k^T \widehat{X}_i). \tag{B.15}$$

Another approach is to represent (B.14) in terms of the (λ, μ, π) and then plug in the estimates. That is,

$$\Delta = \psi_2 \sum_{k} \pi_k \mu_k \mu_k^T, \quad \mathbf{E}[X_1 X_1^T (\mu_k^T X_1)] = \psi_3 \sum_{j} \pi_j (\mu_j^T \mu_k) \mu_j \mu_j^T. \tag{B.16}$$

where $\psi_2 = \mathbf{E}\theta^2 = \frac{\lambda+1}{\lambda}, \psi_3 = \mathbf{E}\theta^3 = \frac{(\lambda+1)(\lambda+2)}{\lambda^2}$, since $\theta \sim Gamma(\lambda, \lambda)$. Updates in (B.16) is preferred due to computation efficiency.

The derivations for eMLE for fixed degree

the probability to observe $\widehat{X}$ and the latent variables $\{z_i, i = 1, 2, \cdots, n\}$ can be written as follows.

$$\begin{aligned} P(\widehat{X}, \mathbf{Z}|\theta, \mu, \pi) &= \prod_{i=1}^{n} N(\widehat{X}_i; \theta_i \mu_{z_i}, \theta_i \mathbf{\Sigma}_{n,z_i})\, P(z_i) \\ &= \prod_{i=1}^{n} \pi_{z_i} (2\pi)^{-\frac{K}{2}} |\theta_i \mathbf{\Sigma}_{n,z_i}|^{-\frac{1}{2}} e^{-\frac{1}{2\theta_i}(\widehat{X}_i - \theta_i \mu_{z_i})^T \mathbf{\Sigma}_{n,z_i}^{-1} (\widehat{X}_i - \theta_i \mu_{z_i})} \end{aligned} \tag{B.17}$$

The probability to observe $\widehat{X}$ is (integrating over hidden variables z_i):

$$P(\widehat{X}|\theta, \boldsymbol{\mu}, \pi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(\widehat{X}_i; \theta_i \mu_k, \theta_i \mathbf{\Sigma}_{n,k}). \tag{B.18}$$

Let $\mathbf{W} = (w_{ik})$, and

$$\begin{aligned} L(\theta, \pi, \boldsymbol{\mu}|\mathbf{W}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log \frac{\pi_k P(\widehat{X}_i|\theta_i\mu_k, \theta_i\boldsymbol{\Sigma}_{n,k})}{w_{ik}} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \left(w_{ik} \log \left\{ \pi_k \left(2\pi\right)^{-K/2} |\theta_i\boldsymbol{\Sigma}_{n,k}|^{-\frac{1}{2}} e^{-\frac{1}{2\theta_i}(\widehat{X}_i - \theta_i\mu_k)^T \boldsymbol{\Sigma}_{n,k}^{-1}(\widehat{X}_i - \theta_i\mu_k)} \right\} \right. \\ &\quad \left. - w_{ik} \log w_{ik} \right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} [\log \pi_k - \frac{K}{2} \log \theta_i - \frac{1}{2} \log |\boldsymbol{\Sigma}_{n,k}| - \frac{1}{2\theta_i} (\widehat{X}_i \\ &\quad - \theta_i\mu_k)^T \boldsymbol{\Sigma}_{n,k}^{-1} (\widehat{X}_i - \theta_i\mu_k)] - \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log w_{ik} + C \end{aligned} \tag{B.19}$$

For π_k. Add the constraint $\sum_k \pi_k = 1$ to the objective with a Lagrange multiplier. Taking derivative w.r.t. π_k and setting the result to 0.

$$\pi_k = \frac{1}{n} \sum_i w_{ik}$$

This is the same as in (2.40).

For μ_k. Taking derivative w.r.t. μ_k and setting the result to 0 (here for simplicity, assuming $\boldsymbol{\Sigma}_{n,k}$ is free of μ_k):

$$\sum_{i=1}^{n} w_{ik} \boldsymbol{\Sigma}_{n,k}^{-1} (\widehat{X}_i - \theta_i\mu_k) = 0. \tag{B.20}$$

Therefore $\mu_k = \frac{\sum_i w_{ik}\widehat{X}_i}{\sum_i w_{ik}\theta_i}$ as in (2.40) in the algorithm. Since we enforce identifiability

through putting constraints on $\boldsymbol{\mu}$, we can standardize the norm of the above estimate.

$$\mu_k \leftarrow \frac{\mu_k}{\|\mu_k\|}, \forall k.$$

Note, the optimization here is not exact 1) not $\Sigma_{n,k}$ is assumed to be free of μ_k; The optimization with constraint is addressed by solving the non-constraint first and then projecting back to the unit sphere.

For θ_i. We only keep the term related to θ_i and have

$$\begin{aligned} L_i &= -\frac{1}{2}\sum_{k=1}^{K} w_{ik}\mu_k^T \mathbf{\Sigma}_{n,k}^{-1}\mu_k\theta_i - \frac{K}{2}\log\theta_i \\ &\quad + \sum_k w_{ik}\mu_k^T\mathbf{\Sigma}_{n,k}^{-1}\widehat{X}_i - \frac{1}{2\theta_i}\sum_k w_{ik}\widehat{X}_i^T\mathbf{\Sigma}_{n,k}^{-1}\widehat{X}_i + C \\ &= -\frac{1}{2}(A_i\,\theta_i + K\log\theta_i + \frac{1}{\theta_i}B_i) + C \end{aligned} \tag{B.21}$$

where $A_i = \sum_{k=1}^{K} w_{ik}\mu_k^T\mathbf{\Sigma}_{n,k}^{-1}\mu_k$, $B_i = \sum_k w_{ik}\widehat{X}_i^T\mathbf{\Sigma}_{n,k}^{-1}\widehat{X}_i$.

Taking derivative w.r.t. θ_i gives that $-\frac{1}{2}(A_i + K\,\frac{1}{\theta_i} - B_i\frac{1}{\theta_i^2}) = 0$. Therefore, the stationary point is

$$\theta_i = \frac{-K + \sqrt{K^2 + 4A_iB_i}}{2A_i} \tag{B.22}$$

For $\mathbf{\Sigma}_{n,k}$. Taking derivative w.r.t. $\mathbf{\Sigma}_{n,k}^{-1}$ and setting the result to 0 (here for simplicity, assuming $\mathbf{\Sigma}_{n,k}$ is free of μ_k):

$$\sum_{i=1}^{n} w_{ik}[\mathbf{\Sigma}_{n,k} - \frac{1}{\theta_i}(\widehat{X}_i - \theta_i\mu_k)(\widehat{X}_i - \theta_i\mu_k)^T] = 0. \tag{B.23}$$

Therefore

$$\widehat{\mathbf{\Sigma}}_{n,k} \leftarrow \frac{\sum_i w_{ik} \frac{1}{\theta_i}(\widehat{X}_i - \theta_i \mu_k)(\widehat{X}_i - \theta_i \mu_k)^T}{\sum_i w_{ik}}$$

This is the formula in (2.37).

From the CLT result,

$$\mathbf{\Sigma}_{n,k} = \frac{1}{n}\Delta^{-1}\mathbf{E}[X_1 X_1^T(\mu_k^T X_1)]\Delta^{-1}.$$

Alternative update for $\mathbf{\Sigma}_{n,k}$ is to plug μ_k into expression of $\mathbf{\Sigma}_{n,k}$ above and substitute the quantity by their sample versions.

$$\mathbf{\Sigma}_{n,k} \leftarrow \frac{1}{n}\widehat{\Delta}^{-1}[\frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i^T(\mu_k^T\widehat{X}_i)]\widehat{\Delta}^{-1}. \tag{B.24}$$

where $\widehat{\Delta}^{-1} = \frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i^T$.

In the original parameters, Δ and $\mathbf{E}[X_1 X_1^T(\mu_k^T X_1)]$ can be represented in terms of μ, θ, π. Thus we can arrive at another more efficient estimate for $\Sigma_{n,k}$. $\widehat{\Delta}^{-1} = \sum_{k=1}^{K}\psi_k\pi_k\mu_k\mu_k^T$, where $\psi_k = \sum_{z_i=k}\theta_i^2$. $\mathbf{E}[X_1 X_1^T(\mu_k^T X_1)] = \sum_{k'}\phi_{k'}\pi_{k'}\mu_{k'}\mu_{k'}^T(\mu_k^T\mu_{k'})$, where $\phi_{k'} = \sum_{z_i=k'}\theta_i^3$.

Therefore, we have

$$\Sigma_{n,k} \leftarrow \widehat{\Delta}^{-1}\left(\sum_{k'}\phi_{k'}\pi_{k'}\mu_{k'}\mu_{k'}^T(\mu_k^T\mu_{k'})\right)\widehat{\Delta}^{-1} \tag{B.25}$$

Proof for the Proposition 2.19:

Proof : There are two common approaches to get the MLE for parameters of this model.

(**Restricted MLE**):The log-likelihood with Lagrange term for constraint $\|\mu\| = 1$ is :

$$\begin{aligned}\ell(\mu, \Sigma, \theta_i|\widehat{X}) &= -\sum_{i=1}^{n} \frac{1}{2\theta_i}(\widehat{X}_i - \theta_i\mu)^T\Sigma^{-1}(\widehat{X}_i - \theta_i\mu) - \frac{K}{2}\sum_{i=1}^{n}\log\theta_i - \frac{n}{2}\log|\Sigma| - \frac{nK}{2}\log(2\pi) \\ &\quad + \frac{\delta}{2}(\mu^T\mu - 1) \qquad \text{(B.26)}\end{aligned}$$

Taking derivative w.r.t. μ, Σ, θ_i respectively and setting them to zeros, we have:

$$\widehat{\mu} = (\sum_{i=1}^{n}\theta_i I_K - \delta\Sigma)^{-1}\sum_{i=1}^{n}\widehat{X}_i = (\bar{\theta}I_K - \frac{\delta}{n}\Sigma)^{-1}\bar{\widehat{X}}, \qquad \text{(B.27)}$$

(using search algorithm to find the $\delta\, s.t.\, \widehat{\mu}^T\widehat{\mu} = 1$)

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i}\frac{(\widehat{X}_i - \theta_i\mu)(\widehat{X}_i - \theta_i\mu)^T}{\theta_i} \qquad \text{(B.28)}$$

$$\widehat{\theta}_i = \frac{-K + \sqrt{K^2 + 4B_iA}}{2A}, \text{ for } i = 1, 2, \cdots n \qquad \text{(B.29)}$$

where in (B.29), $A = \mu^T\Sigma^{-1}\mu$, $B_i = \widehat{X}_i^T\Sigma^{-1}\widehat{X}_i$. These parameters are inter-dependent, and we have to update them iteratively as in the traditional coordinate ascent algorithms. Otherwise, the updating μ will depend on Σ and δ. If $\Sigma = \sigma^2 I_K$ is given, we can get $\widehat{\mu} = \frac{\sum_i \widehat{X}_i}{\|\sum_i \widehat{X}_i\|}$, the estimator $\widehat{\mu}$ will be free of the updates on σ, θ_i.

Appendix C

Appendix for Chapter 3

Proof. (**Proof of Theorem 3.2**) Using triangle inequality, we have

$$\begin{aligned} \|A - \mathcal{A}\|_2 &\leq \|A_1 - \bar{A}_1\|_2 + \|A_2 - \bar{D}_{A_1}^{-1/2} \bar{A}_1^2 \bar{D}_{A_1}^{-1/2}\|_2 + \|A_1 \circ A_2\|_2 \\ &\leq \|A_1 - \bar{A}_1\|_2 + \|A_2 - P_{A_1}\|_2 + \|P_{A_1} - P_{\bar{A}_1}\|_2 + \|A_1 \circ A_2\|_2 \quad \text{(C.1)} \end{aligned}$$

1. For term $\|A_1 - \bar{A}_1\|_2$ in Equation (C.1) , $\|A_1 - \bar{A}_1\|_2 \leq \sqrt{4\Delta \log \frac{2n}{2\epsilon}}$.

2. For term $\|A_2 - P_{A_1}\|$ in Equation (C.1), where $P_{A_1} = D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2}$.

   Conditional on A_1, using Lemma 1, if $d_{max}(P_{A_1}) > \frac{4}{9} \log \frac{2n}{\epsilon}$

$$Pr \quad \left(\|A_2 - P_{A_1}\| > \sqrt{4 d_{max}(P_{A_1}) \log \tfrac{2n}{\epsilon}} \, | A_1 \right) \leq \epsilon$$

   NOTE: The probability here is just conditional probability and the bound is

dependent on A_1.

$$\begin{aligned} Pr & \left(\|A_2 - \mathbf{E}(A_2|A_1)\| > \sqrt{4d_{max}(P_{A_1}) \log \frac{2n}{\epsilon}} \right) \\ = & \mathbf{E}_{A_1} Pr \left(\|A_2 - \mathbf{E}(A_2|A_1)\| > \sqrt{4d_{max}(P_{A_1}) \log \frac{2n}{\epsilon}} | A_1 \right) \\ = & P(\text{good } A_1) \times \epsilon + 1 - P(\text{good } A_1). \end{aligned} \tag{C.2}$$

Here we say A_1 is "good" when A_1 satisfies that $d_{\max}(P_{A_1}) > \frac{4}{9} \log \frac{2n}{\epsilon}$. Let: $\mathcal{T}_1$ denote the set consisting of such A_1's

$$\mathcal{T}_1 = \left\{ A_1 \in R^{n \times n} : d_{max}(P_{A_1}) > \frac{4}{9} \log \frac{2n}{\epsilon} \right\}.$$

For A_1 generated from $\bar{A}_1$, using chernoff bound (as in Theorem 4 in Chung and Lu (2006)), we can prove that if $d_i(\bar{A}_1) > 3 \log(2n/\epsilon)$, then

$$Pr(|d_i(A_1) - d_i(\bar{A}_1)| > b\, d_i(\bar{A}_1)) \leq \frac{\epsilon}{n}, \text{ if } b > \sqrt{\frac{3 \log(2n/\epsilon)}{d_i(\bar{A}_1)}}. \tag{C.3}$$

Take $b = \sqrt{\frac{3ln(2n/\epsilon)}{\delta}}$ so for all i we have $P\left(d_i(A) - d_i(\bar{A})| \geq b\, d_i(\bar{A})\right) < \frac{\epsilon}{n}$. Therefore, if $\delta > 3 \log \frac{2n}{\epsilon}$, we have with probability at least $1 - \epsilon$, we have

$$\|\bar{D}_{A_1}^{-1} D_{A_1} - \mathbf{I}\|_2 = \max_i |\frac{d_i(A_1)}{d_i(\bar{A}_1)} - 1| \leq b.$$

Define the two sets of A_1:

$$\mathcal{T}_0 = \left\{ A_1 \in R^{n\times n} : \|A_1 - \bar{A}_1\| \leq \sqrt{4\Delta \log(2n/\epsilon)} \right\}, \text{ and}$$
$$\mathcal{T}_2 = \{A_1 \in R^{n\times n} : \|\bar{D}_{A_1}^{-1} D_{A_1} - \mathbf{I}\|_2 \leq b\}.$$

As we have proved that if $\delta > 3\log\frac{2n}{\epsilon}$, then $P(\mathcal{T}_2) \geq 1-\epsilon$; if $\Delta \geq \frac{4}{9}\log\frac{2n}{\epsilon}$, then $P(\mathcal{T}_0) \geq 1-\epsilon$. Especially if $\delta > 3\log\frac{2n}{\epsilon}$, then $\Delta > \frac{4}{9}\log\frac{2n}{\epsilon}$ holds automatically. Therefore $\delta > 3\log\frac{2n}{\epsilon}$, then $P(\mathcal{T}_0 \cap \mathcal{T}_2) \geq 1-2\epsilon$.

Claim 1: For $A_1 \in \mathcal{T}_2$, we will have the following inequalities.

$$d_{max}(P_{A_1}) \geq d_{mean}(P_{A_1}) \geq \frac{d_{mean}(A_1)^2}{d_{max}(A_1)} \geq \frac{d_{mean}(\bar{A}_1)^2(1-b)^2}{d_{max}(\bar{A}_1)(1+b)}, \text{ and}$$
$$d_{max}(P_{A_1}) \leq \frac{d_{max}(A_1)^{3/2}}{d_{min}(A_1)^{1/2}} \leq \frac{d_{max}(\bar{A}_1)^{3/2}(1+b)^{3/2}}{d_{min}(\bar{A}_1)^{1/2}(1-b)^{1/2}}$$

Proof : From definition and basic algebra, we have

$$\begin{aligned} d_{max}(P_{A_1}) &= \max_i \sum_j (d_i d_j)^{-1/2} \sum_k A_{ik}A_{jk} \\ &\leq \max_i (d_i d_{\min})^{-1/2} \sum_j \sum_k A_{ik}A_{jk} \\ &\leq \max_i (d_i d_{\min})^{-1/2} \sum_k A_{ik} d_k \\ &\leq \max_i (d_i d_{\min})^{-1/2} d_i d_{\max} \leq \frac{d_{max}(A_1)^{3/2}}{d_{min}(A_1)^{1/2}}. \end{aligned}$$

Claim 2: Under the condition that $\frac{d_{mean}(\bar{A}_1)^2(1-b)^2}{d_{max}(\bar{A}_1)(1+b)} > \frac{4}{9}\log(\frac{2n}{\epsilon})$. Then $A_1 \in \mathcal{T}_2$, implies that $A_1 \in \mathcal{T}_1$. Let $R = \frac{d_{max}(\bar{A}_1)^{3/2}(1+b)^{3/2}}{d_{min}(\bar{A}_1)^{1/2}(1-b)^{1/2}}$, where $b = \sqrt{\frac{3\log(2n/\epsilon)}{d_{min}(\bar{A}_1)}}$.

$$
\begin{aligned}
& Pr\left(\|A_2 - \mathbf{E}(A_2|A_1)\| > \sqrt{4R\ \log\frac{2n}{\epsilon}}\right) \\
= \ & \mathbf{E}_{A_1} Pr\left(\|A_2 - \mathbf{E}(A_2|A_1)\| > \sqrt{4R\ \log\frac{2n}{\epsilon}}\middle| A_1\right) \\
\leq \ & \mathbf{E}_{A_1} Pr\left(\|A_2 - \mathbf{E}(A_2|A_1)\| > \sqrt{4d_{max}(P_{A_1})\ \log\frac{2n}{\epsilon}}\middle| A_1\right) \\
\leq \ & P(A_1 \in \mathcal{T}_2) \times \epsilon + P(A_1 \notin \mathcal{T}_2) \times 1 \\
\leq \ & 2\epsilon
\end{aligned}
$$

3. For 3rd term $\|P_{A_1} - \mathbf{E}P_{A_1}\|$ in Equation (C.1)

$$
\begin{aligned}
\|P_{A_1} - \bar{P}_{A_1}\| &= \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2} - \bar{D}_{A_1}^{-1/2} \bar{A}_1^2 \bar{D}_{\bar{A}_1}^{-1/2}\| \\
&\leq \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2} - \bar{D}_{A_1}^{-1/2} A_1^2 \bar{D}_{A_1}^{-1/2}\| \\
&\quad + \|\bar{D}_{A_1}^{-1/2} A_1^2 \bar{D}_{A_1}^{-1/2} - \bar{D}_{A_1}^{-1/2} \bar{A}_1^2 \bar{D}_{A_1}^{-1/2}\| \\
&= \#Term1 + \#Term2
\end{aligned}
$$

If $A_1 \in \mathcal{T}_2$, and let b same as before, we have

$$
\|\bar{D}_{A_1}^{-1/2} D_{A_1}^{1/2} - \mathbf{I}\| = \max_{i=1,\ldots,n} |\sqrt{\frac{d_i(A_1)}{d_i(\bar{A}_1)}} - 1| \leq \max_{i=1,2,\ldots,n} |\frac{d_i(A_1)}{d_i(\bar{A}_1)} - 1| \leq b.
$$

Therefore,

$$
\begin{aligned}
\#Term1 &= \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2} - \bar{D}_{A_1}^{-1/2} {A_1}^2 \bar{D}_{A_1}^{-1/2}\| \\
&\leq \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2} - \bar{D}_{A_1}^{-1/2} D_{A_1}^{1/2} D_{A_1}^{-1/2} {A_1}^2 D_{A_1}^{-1/2} D_{A_1}^{1/2} \bar{D}_{A_1}^{-1/2}\| \\
&\leq \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2} (\mathbf{I} - \bar{D}_{A_1}^{-1/2} D_{A_1}^{1/2})\| \\
&\quad + \|(\bar{D}_{A_1}^{-1/2} D_{A_1}^{1/2} - \mathbf{I}) D_{A_1}^{-1/2} {A_1}^2 D_{A_1}^{-1/2} D_{A_1}^{1/2} \bar{D}_{A_1}^{-1/2}\| \\
&\leq \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2}\|\, b + b\, \|D_{A_1}^{-1/2} {A_1}^2 D_{A_1}^{-1/2}\|(1+b) \\
&\leq \|D_{A_1}^{-1/2} A_1^2 D_{A_1}^{-1/2}\|(b^2 + 2b) \\
&\leq \|D_{A_1}^{-1/2} A_1 D_{A_1}^{1/2}\|\, \|D_{A_1}^{-1/2} A_1 D_{A_1}^{-1/2}\|\, (b^2 + 2b) \\
&= \|D_{A_1}^{-1/2} A_1 D_{A_1}^{1/2}\|\, (b^2 + 2b) \\
&= \|A_1\|(b^2 + 2b) \\
&= d_{max}(\bar{A}_1)(1+b)(b^2 + 2b) \\
\text{For } A_1 &\in \mathcal{T}_1 \cap \mathcal{T}_2, \\
\#Term2 &= \|\bar{D}_{A_1}^{-1/2} (A_1^2 - (\bar{A}_1)^2) \bar{D}_{A_1}^{-1/2}\| \\
&\leq \frac{1}{d_{min}(\bar{A}_1)} \|A_1^2 - (\bar{A}_1)^2\| \\
&\leq \frac{d_{max}(\bar{A}_1)}{d_{min}(\bar{A}_1)} (2+b) \|A_1 - \bar{A}_1\| \\
&\leq \frac{d_{max}(\bar{A}_1)}{d_{min}(\bar{A}_1)} (2+b) \sqrt{4 d_{\max}(\bar{A}_1) \log(2n/\epsilon)}
\end{aligned}
$$

In the second inequality above, we use the following result:

$$\begin{aligned}
\|A_1^2 - (\bar{A}_1)^2\| &\leq \|(A_1 - \bar{A}_1)A_1\| + \|\bar{A}_1(A_1 - \bar{A}_1)\| \\
&\leq \|(A_1 - \bar{A}_1)\|\,\|A_1\| + \|\bar{A}_1\|\,\|(A_1 - \bar{A}_1)\| \\
&\leq (d_{max}(A_1) + d_{max}(\bar{A}_1))\|A_1 - \bar{A}_1\| \\
&\leq d_{max}(\bar{A}_1)(2+b)\|A_1 - \bar{A}_1\|
\end{aligned}$$

Also, the result that $\|A_1 - \bar{A}_1\| = \sqrt{4d_{max}(\bar{A}_1)\log\frac{2n}{\epsilon}}$ hold with for $A_1 \in \mathcal{T}_0$.

To sum up, for $A_1 \in \mathcal{T}_0 \cap \mathcal{T}_2$,

$$\begin{aligned}
\|P_{A_1} - \mathbf{E}P_{A_1}\| &\leq \frac{d_{max}(\bar{A}_1)}{d_{min}(\bar{A}_1)}(2+b)\|A_1 - \bar{A}_1\| + d_{max}(\bar{A}_1)(1+b)(b^2+2b) \\
&\leq \frac{d_{max}(\bar{A}_1)}{d_{min}(\bar{A}_1)}(2+b)\sqrt{4\,d_{max}(A)_1)\log\frac{2n}{\epsilon}} + d_{max}(\bar{A}_1)(2b+3b^2+b^3)
\end{aligned}$$

(4) : For 4th term $A_1 \circ A_2$ in Equation (C.1).

First of all $A_2 \circ A_1$ can be decompsed into two parts:

$$A_2 \circ A_1 = [A_2 - D_{A_1}^{-1/2}A_1^2D_{A_1}^{-1/2}] \circ A_1 + D_{A_1}^{-1/2}A_1^2D_{A_1}^{-1/2} \circ A_1 \tag{C.4}$$

We are going to apply Lemma C.1 and Lemma C.3 to bound $A_2 \circ A_1$. When given A_1, $[A_2 - D_{A_1}^{-1/2}A_1^2D_{A_1}^{-1/2}] \circ A_1$ can be expressed in terms of sum of independent matrices. Let $\mathbf{E}X_{ij} = \sum_k [A_1]_{ik}[A_1]_{jk}[A_1]_{ij}E^{ij} = [P_{A_1} \circ A_1]_{ij}E^{ij}$, for $i < j$, where $E^{ij} \in \mathbb{R}^{n\times n}$ is a symmetric matrix with (i,j) and (j,i) cell taking value 1, others 0.

To match the Lemma C.1, $\|X_{ij}\| \leq 1$, i.e. $M = 1$.

$$\begin{aligned} v^2 : &= \|\sum_{i<j} var(X_{ij})\|_2 \\ &= \|\sum_{i<j}([P_{A_1} \circ A_1]_{ij}(1 - [P_{A_1} \circ A_1]_{ij})(E^{ii} + E^{jj})\| \\ &\leq \max_i \sum_{j=1}^{n} [P_{A_1} \circ A_1]_{ij} \\ &= d_{\max}(P_{A_1} \circ A_1) \\ &\leq d_{\max}(A_1)(\text{ since element of } P_{A_1} \text{ is smaller than 1}) \end{aligned}$$

Applying the Lemma C.1, we have

$$\begin{aligned} P\left(\|A_2 \circ A_1 - P_{A_1} \circ A_1\|_2 > t | A_1\right) &\leq 2n \exp\left(-\frac{t^2/2}{d_{\max}(P_{A_1} \circ A_1) + t/3}\right) \\ &\leq 2n \exp\left(-\frac{t^2/2}{d_{\max}(A_1) + t/3}\right) \end{aligned} \tag{C.5}$$

As we know, when $\delta > 3 \log \frac{2n}{\epsilon}$, $d_{\max}(A_1) < (1+b) d_{\max}(\bar{A}_1)$ hold with probability $> 1 - \epsilon$. Use the fact $d_{\max}(\bar{A}_1) \leq n\alpha_n$, and let $t = \sqrt{4n\alpha_n \log(2n/\epsilon)}$, we have

$$
\begin{aligned}
& P\left(\|A_2 \circ A_1 - P_{A_1} \circ A_1\|_2 > \sqrt{4(1+b)n\alpha_n \log(2n/\epsilon)}\right) \\
&= \mathbf{E}P\left(\|A_2 \circ A_1 - P_{A_1} \circ A_1\|_2 > \sqrt{4(1+b)n\alpha_n \log(2n/\epsilon)}|A_1\right) \\
&= P\left(\|A_2 \circ A_1 - P_{A_1} \circ A_1\|_2 > \sqrt{4(1+b)n\alpha_n \log(2n/\epsilon)}|A_1\right) P(\text{good } A_1) \\
&\quad + (1 - P(\text{good } A_1)) \\
&\leq 2n \exp\left(-\frac{t^2/2}{d_{\max}(A_1) + t/3}\right)(1-\epsilon) + \epsilon \\
&\leq 2n \exp\left(-\frac{t^2/2}{d_{\max}(\bar{A}_1)(1+b) + t/3}\right)(1-\epsilon) + \epsilon \\
&\leq 2n \exp\left(-\frac{2(1+b)n\alpha_n \log(2n/\epsilon)}{n\alpha_n(1+b) + \sqrt{4(1+b)n\alpha_n \log(2n/\epsilon)}/3}\right)(1-\epsilon) + \epsilon \\
&\leq 2\epsilon, \text{ if } \quad \frac{1}{3}\sqrt{\frac{4\log(2n/\epsilon)}{(1+b)n\alpha_n}} < 1
\end{aligned} \tag{C.6}
$$

Therefore, under the assumptions: $\delta > 3\log(2n/\epsilon)$, thus $b = \sqrt{\frac{3\log(2n/\epsilon)}{\delta}} < 1$, with probability at least $1 - 2\epsilon$, we have

$$\|A_2 \circ A_1 - P_{A_1} \circ A_1\|_2 \leq \sqrt{4(1+b)n\alpha_n \log(2n/\epsilon)}.$$

For the second term in Equation (C.4), Assume that $A_1 \in \mathcal{T}_2$, then the degree matrix D_{A_1} concentrates on its population degree $D_{\bar{A}_1}$, with $|[D_{A_1}]_{ii} - [D_{\bar{A}_1}]_{ii}| \leq$

$b\,[D_{\bar{A}_1}]_{ii}$. Use the Lemma C.3 below, we can have with $> 1-\epsilon$,

$$\begin{aligned}\|D_{A_1}^{-1/2}A_1^2D_{A_1}^{-1/2}\circ A_1\| &\le \|D_{A_1}^{-1/2}(A_1^2\circ A_1)D_{A_1}^{-1/2}\|_F\\ &\le \frac{1}{\delta(1-b)}\|A^2\circ A\|_F\\ &\le \frac{C}{\delta_0(1-b)}(n\alpha_n/\epsilon)^{1/2} \qquad \text{(C.7)}\end{aligned}$$

(5) Last step, we incorporate all those bounds.

Therefore, $A = A_1 + A_2 - A_1\circ A_2$, $\|A-\mathcal{A}\| \le d_{\max}(\bar{A}_1)b = O(n\alpha_n b)$, where $b = \sqrt{\frac{\log(2n/\epsilon)}{d_{\min}(\bar{A}_1)}}$ Under the conditions C1, C2, with probability at least $1-3\epsilon$, we have following bound:

$$\begin{aligned}\|A-\mathcal{A}\| &\le \|A_1-\bar{A}_1\|_2 + \|A_2-P_{A_1}\|_2 + \|P_{A_1}-P_{\bar{A}_1}\|_2 + \|A_1\circ A_2\|\\ &\le \|\sqrt{4d_{max}(\bar{A}_1)\log\frac{2n}{2\epsilon}} + \sqrt{\frac{d_{max}(\bar{A}_1)^{3/2}(1+b)^{3/2}}{d_{min}(\bar{A}_1)^{1/2}(1-b)^{1/2}}\log\frac{2n}{\epsilon}}\\ &\quad + \frac{d_{max}(\bar{A}_1)}{d_{min}(\bar{A}_1)}(2+b)\sqrt{4\,d_{max}(A)_1)\log\frac{2n}{\epsilon}} + d_{max}(\bar{A}_1)(2b+3b^2+b^3)\\ &\quad + \sqrt{4(1+b)n\alpha_n\log(2n/\epsilon)} + \frac{C}{\delta_0(1-b)}(n\alpha_n/\epsilon)^{1/2}\\ &\le C(n\alpha_n)b \qquad \text{(C.8)}\end{aligned}$$

□

Proof. (**Proof of Theorem 3.5**)

Based on the definition, there is a sufficient condition for node i to be correctly clustered. From the fact that $\mathcal{X} = Z(Z^TZ)^{-1/2}U$, we have $\mathcal{C}_i$ and $\mathcal{C}_j$ are orthogonal

if $Z_i \neq Z_j$, and the $\|\mathcal{C}_i\|_2 = \frac{1}{\sqrt{n_i}}$, where n_i is the block size of the cluster to which node i belongs. On the other hand, $||\mathcal{C}_i - \mathcal{C}_j|| \geq \sqrt{2/n_{max}}$, for $Z_i \neq Z_j$. Therefore the condition that $\|C_iO - \mathcal{C}_i\| \leq \frac{1}{2}\sqrt{2/n_{max}}$ will imply that $\mathcal{C}_i$ is the closest population center to C_iO. On the other hand, if node i is mis-clustered, then $\|C_iO - \mathcal{C}_i\| \geq \frac{1}{2}\sqrt{2/n_{max}}$.

Now we can bound the mis-clustering rate as below.

$$\begin{aligned}
|\mathcal{U}|/n &\leq \frac{1}{n}\sum_{i=1}^{n} 1_{\left(\{i:\|C_iO-\mathcal{C}_i\|\geq\sqrt{1/(2n_{max})}\}\right)} \\
&\leq \frac{1}{n}\sum_{i:\|C_iO-\mathcal{C}_i\|\geq\sqrt{1/(2n_{max})}} \|C_iO - \mathcal{C}_i\|_2^2 * 2n_{max} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|C_iO - \mathcal{C}_i\|_2^2 * 2n_{max} \\
&= \frac{1}{n}\|CO - \mathcal{C}\|_F^2 * 2n_{max} \\
&\leq \frac{1}{n}2(\|CO - XO\|_F^2 + \|XO - \mathcal{X}\|_F^2) * 2n_{max} \quad (*) \\
&\leq \frac{4}{n}\|XO - \mathcal{X}\|_F^2 * n_{max}.
\end{aligned}$$

Continued from above,

$$\begin{aligned}
|\mathcal{U}|/n &\leq \frac{8}{n}(\frac{2\sqrt{2K}}{\lambda_K}\|A - \mathcal{A}\|)^2 * n_{max} \\
&= \frac{64Kn_{max}}{n\lambda_K^2}\|A - \mathcal{A}\|^2 \\
&= O(\frac{64Kn_{max}\Delta^2}{n\lambda_K^2}b^2).
\end{aligned}$$

Here the inequality (*) follows from the fact that C is a $n \times K$ matrix consisting of centers obtained by k-means from X, and we assume that k-means achieves the global minimum ,thus satisfying that $\|XO - CO\|_F \leq \|XO - \mathcal{X}\|_F$. □

Lemma C.1 (Matrix Concentration, see Theorem 5 in Chung and Radcliffe (2011))**.** *Let X_1 , X_2, ... X_m be independent random Hermitian matrices. Assume that $\|X_i\| \leq M$ and $v^2 = \|var(\sum_{i=1}^m X_i)\|$, Then*

$$P(\|\sum_{i=1}^m X_i - \mathbf{E}\sum_{i=1}^m X_i\| > a) \leq 2n \exp\left(-\frac{a^2}{2v^2 + 2Ma/3}\right).$$

Results on random graphs: Δ be the highest expected degree and δ be the lowest expected degree. With probability $> 1 - \epsilon$, we have

$$\|A - \mathbf{E}A\| \leq \sqrt{4\Delta \log \frac{2n}{\epsilon}}, \text{ and}$$

$$\|L - \mathbf{E}L\| \leq 4\sqrt{\frac{3\log(2n/\epsilon)}{\delta}}.$$

Lemma C.2 (Dave-Kahan's Theorem, Lemma 5.1 in Lei and Rinaldo (2013))**.** *Let X and $\mathcal{X}$ be the leading K-dimensional eigen-space of A and $\mathcal{A}$, and λ_K be K-th leading eigenvalue of $\mathcal{A}$. Then there exists a $K \times K$ orthonormal matrix O such that*

$$XO - \mathcal{X}\|_F \leq \frac{2\sqrt{2K}}{\lambda_K}\|A - \mathcal{A}\|_2$$

Lemma C.3. *Assume that A is sampled from an SBM model with with π_0-proper*

clusters, big enough eigengap, we have with probability $> 1 - \epsilon$,

$$\|A^2 \circ A\|_F \leq \sqrt{(n\alpha_n)^3/\epsilon}.$$

Proof. For item (i, j) in the matrix,

$$(A^2 \circ A)_{ij} = \begin{cases} \sum_{k \neq i,j} A_{ik}A_{jk}A_{ij} + (A_{ii} + A_{jj})A_{ij}, & i \neq j; \\ \sum_{k \neq i,j} A_{ik}A_{ii} + 2A_{ii} & i = j. \end{cases}$$

$$\mathbf{E}[(A^2 \circ A)_{ij}] = \begin{cases} \sum_{k \neq i,j} \bar{A}_{ik}\bar{A}_{jk}\bar{A}_{ij} + (\bar{A}_{ii} + \bar{A}_{jj})\bar{A}_{ij}, & i \neq j; \\ \sum_{k \neq i,j} \bar{A}_{ik}\bar{A}_{ii} + 2\bar{A}_{ii} & i = j. \end{cases}.$$

For $i < j$,

$$\begin{aligned}((A^2 \circ A)_{ij})^2 = &\sum_{k,l \neq i,j} A_{ij}A_{ik}A_{jk}A_{il}A_{jl} + \sum_{k \neq i,j} A_{ij}A_{ik}A_{jk} \\ &+ \sum_k A_{ij}A_{ik}A_{jk}(a_{ii} + A_{jj}) + A_{ij}(A_{ii} + A_{jj} + 2A_{ii}A_{jj}).\end{aligned}$$

For $i = j$,

$$((A^2 \circ A)_{ij})^2 = \sum_{k,l \neq i} A_{ii}A_{ik}A_{il} + \sum_{k \neq i} A_{ii}A_{ik} + 4A_{ii}.$$

In the expectation,there exists absolute constant C_1, C_2, such that

$$\mathbf{E}[(A^2 \circ A)_{ij}]^2 = O(n^2\alpha_n^5 + n\alpha_n^3 + n\alpha_n^4 + 2\alpha_n^2 + 2\alpha_n^3) \leq C_1 n\alpha_n^3$$
$$\mathbf{E}[(A^2 \circ A)_{ii}]^2 \leq C_2 n^2\alpha_n^3$$

Luckily, A_2 and A_1 will be foreced to be zero on the diagnoal, so

$$\mathbf{E}\|A^2 \circ A)\|_F^2 = 2\sum_{i<j} \mathbf{E}[(A^2 \circ A)_{ij}]^2 = C(n^3\alpha_n^3)$$

Using Markov inequality, we have

$$P\left(\|A^2 \circ A)\|_F > C[(n\alpha_n)^3/\epsilon]^{1/2}\right) \leq \epsilon.$$

□

Appendix D

Appendix for Chapter 4

Theorem 4.9

Denote B_i as the ith row in B and $\bar{B}_i$ as ith row in $\bar{B}$. Let $y_i = B_i X \Lambda^{-1}$, and denote its population counterpart as $\bar{y}_i = \bar{B}_i \bar{X} \bar{\Lambda}$. Next is trying to bound $\|y_i - \bar{y}_i \mathcal{O}_1\|$:

$$\begin{aligned} \|y_i - \bar{y}_i \mathcal{O}\| &= \|BX\Lambda^{-1} - \bar{B}_i \bar{X} \bar{\Lambda}^{-1} \mathcal{O}\| \\ &\leq \|B_i X \Lambda^{-1} - B_i \bar{X} \mathcal{O} \bar{\Lambda}^{-1}\| + \|(B_i - \bar{B}_i) \bar{X} \bar{\Lambda}^{-1} \mathcal{O}\| \\ &\leq \|(B_i - \bar{B}_i) \bar{X} \mathcal{O}_1 \bar{\Lambda}^{-1}\| + \|B_i (X - \bar{X}\mathcal{O}) \bar{\Lambda}^{-1}\| \\ &\quad + \|B_i X (\Lambda^{-1} - \bar{\Lambda}^{-1})\|. \end{aligned} \tag{D.1}$$

Looking at the first term in (D.1), i.e. $\|(B_i - \bar{B}_i) \bar{X} \mathcal{O}_1 \bar{\Lambda}^{-1}\|$:

Using the concentration inequality Lemma D.4 in Appendix, and let $M := \|\bar{X}\|_\infty$,

$v^2 := \sum_j \mathbf{v}ar[(B_{ij} - \bar{B}_{ij})\bar{X}_{jk}] = \sum_j X_{jk}^2 \bar{B}_{ij}(1 - \bar{B}_{ij}) \leq \max_j \bar{B}_{ij}$,

$$Pr(|(B_i - \bar{B}_i)^T \bar{X}_{*k}| > t) \quad \leq \quad 2\exp(-\frac{t^2/2}{v^2 + Mt/3})$$

Let $t = 2v\sqrt{\log(2n/\epsilon)}$, and if $\|\bar{X}_{*k}\|_\infty \leq \frac{3v}{2\sqrt{\log\frac{2n}{\epsilon}}}$, we have $Mt \leq 3v^2$, and futher that the probability above will be smaller than ϵ/n.

Plugging in the conditions ii) in Theorem 4.9 on eigenvalues that $\lambda_K = O(n\alpha_n)$ and $v \leq \sqrt{\alpha_n}$, the first term in (D.1) has the following bound with probability at least $1 - \epsilon$,

$$\|(B_i - \bar{B}_i)\bar{X}\bar{\Lambda}^{-1}\| \leq C\, 2v\sqrt{\log(2n/\epsilon)} \cdot \sqrt{K} \cdot \frac{1}{\bar{\lambda}_K} = O(n^{-1/2}\sqrt{\frac{K\log\frac{2n}{\epsilon}}{n\alpha_n}}) \qquad \text{(D.2)}$$

Remark D.1. *1,* $v^2 = \max_{ij} \bar{B}_{ij} = \max_{ij} \theta_i\theta_j Z_i P Z_j^T = \alpha_n$.

2, Here I used Bernstein inequality. using Hoeffding's inequality, with the intervals being $[-|X_{jk}|, |X_{jk}|]$*, we have* $|(B_i - \bar{B}_i)^T \bar{X}_{*k}| \leq \sqrt{2\log(2n/\epsilon)}$ *with probability* $1 - \frac{\epsilon}{n}$.

Looking at the second term in (D.1) i.e. $\|B_i(X - \bar{X}\mathcal{O})\bar{\Lambda}^{-1}\|$:

$\|B_i(X - \mathcal{X}\mathcal{O})\bar{\Lambda}^{-1}\| \leq \|B_i\|_2 \frac{1}{\bar{\lambda}_K} \frac{\sqrt{K\Delta\log\frac{2n}{\epsilon}}}{\bar{\lambda}_K}$, where Δ is the maximum expected degree in A or maximum row sum of $\bar{A}$. So, $\Delta \leq n\alpha_n$.

$\|B_i\|_2 = \sqrt{\sum_j B_{ij}^2} = \sqrt{\sum_j B_{ij}} \leq O(\sqrt{\sum_j \bar{B}_{ij}}) = O(\sqrt{n\alpha_n})$. This is because B_{ij}'s are Bernoulli random variables with probabilities bounded by α_n.

Plugging in $\lambda_K \geq c_0\,(n\alpha_n), \Delta = O(n\alpha_n), \|B_i\|_2 = O(n\alpha_n)$, we have

$$\|B_i(X - \mathcal{X}\mathcal{O})\bar{\Lambda}^{-1}\| \leq O(n^{-1/2}\sqrt{\frac{K\log(2n/\epsilon)}{n\alpha_n^2}}) \qquad \text{(D.3)}$$

Looking at the third term in (D.1), i.e. $\|B_i X(\Lambda^{-1} - \bar{\Lambda}^{-1})\|$:

First,

$$\begin{aligned} |\lambda_i^{-1} - \bar{\lambda}_i^{-1}| &\leq \frac{1}{\lambda_i \bar{\lambda}_i} |\lambda_i - \bar{\lambda}_i| \leq \frac{2}{\bar{\lambda}_K^2} \|A - \mathcal{A}\| \\ &= O\left((n\alpha_n)^{-2} \sqrt{n\alpha_n \log(2n/\epsilon)} \right) \end{aligned}$$

Conditional on X is given with $\|X_{*k}\|_2 = 1$,

$$B_i X_{*k} \leq \|B_i\|_2 \|X_{*k}\|_2 \leq O(\sqrt{n\alpha_n}) \tag{D.4}$$

. Therefore,

$$\begin{aligned} \|B_i^T X \mathcal{O}(\Lambda^{-1} - \bar{\Lambda}^{-1}) &\leq O(\sqrt{n\alpha_n} K^{1/2}) \, (n\alpha_n)^{-2} \sqrt{n\alpha_n \log(2n/\epsilon)} \\ &\leq O\left(n^{-1/2} \sqrt{\frac{K \log(2n/\epsilon)}{n\alpha_n^2}} \right). \end{aligned} \tag{D.5}$$

Combining all the results above, we have

$$\begin{aligned} \|y_i - \bar{y}_i \mathcal{O}\| &= \|BX\Lambda^{-1} - \bar{B}_i \bar{X} \bar{\Lambda}^{-1} \mathcal{O}\| \\ &\leq \|(B_i - \bar{B}_i) \bar{X} \mathcal{O}_1 \bar{\Lambda}^{-1}\| + \|B_i (X - \bar{X}\mathcal{O}) \bar{\Lambda}^{-1}\| + \|B_i X (\Lambda^{-1} - \bar{\Lambda}^{-1})\| \\ &\leq O(n^{-1/2} \sqrt{\frac{K \log \frac{2n_B}{\epsilon}}{n\alpha_n}}) + O(n^{-1/2} \sqrt{\frac{4 \log \frac{2n}{\epsilon}}{n\alpha_n^2}}) + O(n^{-1/2} \sqrt{\frac{K \log \frac{2n}{\epsilon}}{n\alpha_n^2}}) \\ &\leq O\left(n^{-1/2} \sqrt{\frac{K \log \frac{2n}{\epsilon}}{n\alpha_n^2}} \right). \end{aligned} \tag{D.6}$$

If $n\alpha_n^2 = \omega(\sqrt{K \log n})$, we claim that this is a concentration result, since

$$\begin{aligned} \|\bar{y}_i\| &= \|\bar{B}_i^T \mathcal{X} \mathcal{O}_1 \bar{\Lambda}^{-1}\| \\ &= \|\theta_i (Z_B)_i \alpha_n P Z_A^T \Theta^2 Z_A (Z_A^T \Theta^2 Z_A)^{-1/2} U \Lambda^{-1} \mathcal{O}_1\| \\ &= \|\theta_i (Z_B)_i (Z_A^T \Theta^2 Z_A)^{-1/2} (Z_A^T \Theta^2 Z_A)^{1/2} P (Z_A^T \Theta^2 Z_A)^{1/2} U \Lambda^{-1} \mathcal{O}_1\| \\ &= \|\theta_i (Z_B)_i (Z_A^T \Theta^2 Z_A)^{-1/2} U\| = O(n^{-1/2}). \end{aligned} \tag{D.7}$$

Proof of Main result

Step 1: As we have proved, except a smaller proportion of points, all the rows in Y are in r_n neighborhood of corresponding rows in $\bar{Y}$. Orginally, $\bar{Y}$ is a polytope and thus Y is a perturbed $\bar{Y}$.

Two properties about the pure points: First, in $\bar{Y}$, two rows from same single cluster will have same row representation, thus there are many points are just perturbed version of the pure point, thus there is high density around the pure points;

secondly, the rows in $\bar{Y}$ are vertex points in the polytope, thus pure points in Y also tends to be extreme/boundary points.

The projection rule to Keep all the points taking values $> (1 - 2r_n)Max$ or $< (1 + 2r_n)Min$ around the extreme points (maximum or minimum) will help exploit these two properties. Each such project will allow us to keep the points around a pure/vertex point, which will then include around $n/K \approx O(n)$ of pure points compared to the $O(nr_n)$ mixed points if the mixed nodes are uniform, or have no point mass.

Step 2: Try to bound $\|\widehat{V} - V\|$, where $\widehat{V}$ contains centers returned by K-means

on Y_S, while V contains centers from population $\bar{Y}_S$. From definition, we have

$$\|\widehat{V} - \bar{V}\|_F \quad \leq \quad \sqrt{K} \max_{0 \leq i \leq K} \|\widehat{V}_{i\cdot} - \bar{V}_{i\cdot}\| \tag{D.8}$$

Let $\widehat{V} := \arg\min_{V \in \mathbb{R}^{K \times K}} \sum_i \min_{0 \leq k \leq K} \|Y_i^* - v_k\|_2^2$, and

$\bar{V} := \arg\min_{V \in \mathbb{R}^{K \times K}} \sum_i \min_{0 \leq k \leq K} \|\mathcal{Y}_i^* - v_k\|_2^2$, where $\bar{V}$ is asymptotically close to $(Z_A^T Z_A)^{-1/2} U \mathcal{O}$.

Triangle inequality gives that:

$$\|Y_S - \bar{Y}_S\| = \|Y_S - Z_1\widehat{V} + Z_1\widehat{V} - Z_2\bar{V} + Z_2\bar{V} - \bar{Y}_S\| \tag{D.9}$$

$$n_{min} \max_i \|\widehat{V}_i - \bar{V}_i\| \leq \|Z_1\widehat{V} - Z_2\bar{V}\| \leq 3\|Y_S - \bar{Y}_S\| = \sqrt{n} O(n^{-1/2} r_n).$$

$$\|\widehat{V} - \bar{V}\|_F \leq O(\sqrt{K} n^{-1/2}\, r_n), \quad (\bar{V}^T \bar{V})^{1/2} = (Z_A^T Z_A)^{-1/2} = O(1/\sqrt{n}). \tag{D.10}$$

Step 3: try to bound the membership $\frac{1}{n}\|\widehat{Z} - Z\|$.

$$\begin{aligned}
& \| \begin{pmatrix} X \\ Y \end{pmatrix} V^{-1} - \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \bar{V}^{-1} \| \\
\leq & \begin{pmatrix} X \\ Y \end{pmatrix} \|V^{-1} - \bar{V}^{-1}\| + \| \begin{pmatrix} X \\ Y \end{pmatrix} - \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \| \|\bar{V}^{-1}\| \\
\leq & \begin{pmatrix} X \\ Y \end{pmatrix} \|V - \bar{V}\| \|\bar{V}^{-1}\widehat{V}^{-1}\| + \| \begin{pmatrix} X \\ Y \end{pmatrix} - \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \| \|\bar{V}^{-1}\| \\
\leq & 2\sqrt{K}\sqrt{K} O(n^{-1/2} r_n) O(n) + (\frac{\|A - \bar{A}\|}{\bar{\lambda}_K} + r_n) O(n^{1/2})
\end{aligned}$$

$$\frac{1}{\sqrt{n}}\|\widehat{Z} - Z\|_F = O(\frac{\|A - \bar{A}\|}{\bar{\lambda}_K} + r_n) = O(\sqrt{\frac{\log(2n/\epsilon)}{n\alpha_n^2}}) \tag{D.11}$$

since the $\frac{\|A-\bar{A}\|}{\bar{\lambda}_K} = \sqrt{\frac{\log(2n/\epsilon)}{n\alpha_n}}$.

$$\frac{1}{\sqrt{n}}\|\widehat{Z}^* - Z^*\|_F \leq \frac{2}{m}\frac{1}{\sqrt{n}}\|\widehat{Z} - Z\|_F = O(\sqrt{\frac{\log(2n/\epsilon)}{n\alpha_n^2}}) \tag{D.12}$$

where m is the minimum row norm of $\bar{Y} \cdot (\bar{V})^{-1} = Z$, which has row norm all equal to 1. $m = 1$ since $Z^* = Z$.

Some lemmas

Lemma D.2 (Sherman-Mrrison formula)**.** *Suppose A is an invertible square matrix and u, v are column vectors . Suppose that $1 + v^T A^{-1} u \neq 0$, then*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}. \tag{D.13}$$

Lemma D.3 (Hoeffding's inequality)**.** *Assume that X_i are random variables and are strictly bounded by the intervals $[a_i, b_i]$ and $S_n = \sum_{i=1}^n X_i$, then*

$$\mathbb{P}\left(S_n - \mathrm{E}\left[S_n\right] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{D.14}$$

$$\mathbb{P}\left(|S_n - \mathrm{E}\left[S_n\right]| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{D.15}$$

Lemma D.4 (Scalar Bernstein concentration in Chung and Radcliffe (2011))**.** *Let X_1 , X_2, ... X_m be independent random variables. Assume that $|X_i| \leq M$ and $v^2 = var(\sum_{i=1}^m X_i)$, Then*

$$P(|\sum_{i=1}^m X_i - \mathbf{E}\sum_{i=1}^m X_i| > a) \leq 2\exp\left(-\frac{a^2}{2v^2 + 2Ma/3}\right). \tag{D.16}$$

If $M \leq \frac{3v}{\sqrt{\log\frac{2}{\epsilon}}}$, let $a = 2v\sqrt{\log\frac{2}{\epsilon}}$, we have

$$P(|\sum_{i=1}^m X_i - \mathbf{E}\sum_{i=1}^m X_i| > a) \leq \epsilon. \tag{D.17}$$

Lemma D.5 (Matrix Bernstein Concentration in Chung and Radcliffe (2011))**.** *Let*

X_1 , X_2, ... X_m be independent random Hermitian matrices. Assume that $\|X_i\| \leq M$ and $v^2 = \|var(\sum_{i=1}^m X_i)\|$, Then

$$P(\|\sum_{i=1}^m X_i - \mathbf{E}\sum_{i=1}^m X_i\| > a) \leq 2n \exp\left(-\frac{a^2}{2v^2 + 2Ma/3}\right). \tag{D.18}$$

Results on random graphs: Δ be the highest expected degree and δ be the lowest expected degree. With probability $> 1 - \epsilon$, we have

$$\|A - \mathbf{E}A\| \leq \sqrt{4\Delta \log \frac{2n}{\epsilon}} \tag{D.19}$$

$$\|L - \mathbf{E}L\| \leq 4\sqrt{\frac{3\log(2n/\epsilon)}{\delta}} \tag{D.20}$$

Lemma D.6 (Dave-Kahan's Theorem, Lemma 5.1 in Lei and Rinaldo (2013))**.** *Let X and $\mathcal{X}$ be the leading K-dimensional eigen-space of A and $\mathcal{A}$, and λ_K be K-th leading eigenvalue of $\mathcal{A}$. Then there exists a $K \times K$ orthonormal matrix O such that*

$$\|X - \mathcal{X}O\|_F \leq \frac{2\sqrt{2K}}{\bar{\lambda}_K}\|A - \mathcal{A}\| \tag{D.21}$$

Lemma D.7 (Infinity norm inequality non-symmetric, Jianqing Fan and Zhong (2016))**.** *Let $A = \bar{A} + E$, where $\bar{A} \in \mathbb{R}^{n\times n}$, is a rank k and E is the perturbation symmetric matrix. Let the sigular value decomposition be $\bar{A} = \bar{X}\bar{\Lambda}\bar{X}^T$ and $A = X\Lambda X^T$. $\mu(\bar{X}) = \frac{n}{k}\max_i \sum_{j=1}^k \bar{X}_{ij}^2$. $\tau_0 = \|E\|_1$. Then there exists $\eta_1, \eta_2, ..., \eta_k$ that are either 1*

or -1*, and constants* $C_0 2$ *that depend on* r *and* μ*, such that*

$$\max_{1\le i\le k} \|X_{*i} - \eta_i \bar{X}_{*i}\|_{\max} \le C_0 \frac{\tau_0}{\gamma_0 \sqrt{n}} \tag{D.22}$$

Lemma D.8 (Infinity norm inequality Symmetric, Jianqing Fan and Zhong (2016))**.** *Let* $A = \bar{A} + E$*, where* $\bar{A} \in \mathbb{R}^{n\times n}$*, is a rank* r *and* E *is the perturbation symmetric matrix. Let the singular value decomposition be* $\bar{A} = \bar{X}\bar{\Lambda}\bar{X}^T$ *and* $A = X\Lambda X^T$*. Assume* $\bar{A}$ *satisfies the following conditions:*

1. *matrix coherence:* $\mu = \frac{n}{k} \max_i \sum_{j=1}^k \bar{X}_{ij}^2$.
2. $\tau = \max_{1\le i\le n} \sum_{j=1}^n |E_{ij}| \quad (= \|E\|_1)$
3. $\kappa := \sqrt{n}\|\bar{X}^T E\|_{\max} \quad (\le \tau\sqrt{r\mu})$
4. *eigengap :* $\gamma = \min\{\lambda_i - \lambda_{i+1} : 1 \le i \le r\} \wedge \min\{|\lambda_i| : 1 \le i \le r\}$ *with convention* $\lambda_{r+1} = -\infty$*. If* A *is positive definite,* $\gamma = \gamma_0$.
5. $\gamma > 5r\mu(\tau + 2r\kappa)$ *(May not be satisfied!!!)*

Then there exists $\eta_1, \eta_2, ..., \eta_r$ *that are either 1 or -1,*

$$\max_{1\le i\le r} \|X_{*i} - \eta_i \bar{X}_{*i}\|_{\max} \le C(r,\mu) \frac{\tau + \kappa}{\gamma\sqrt{n}} \tag{D.23}$$

where $C(r, \mu(\bar{X})) = 45r^{5/2}\sqrt{\mu(\bar{X})}(1 + r\mu(\bar{X}))$.

In particular, when r *and* μ *are bounded by a constant, we have*

$$\max_{1\le i\le r} \|X_{*i} - \eta_i \bar{X}_{*i}\|_{\infty} \le C' \frac{\tau}{\gamma\sqrt{n}} \tag{D.24}$$

References

Adamic, Lada A, and Natalie Glance. 2005. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on link discovery*, 36–43. ACM.

Airoldi, Edoardo M, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9(Sep):1981–2014.

Amini, Arash A, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. 2013. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* 41(4):2097–2122.

Anandkumar, Anima, Rong Ge, Daniel Hsu, and Sham M Kakade. 2013. A tensor approach to learning mixed membership community models. *arXiv preprint arXiv:1302.2684*.

Athreya, Avanti, Carey E Priebe, Minh Tang, Vince Lyzinski, David J Marchette, and Daniel L Sussman. 2016. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* 78(1):1–18.

Banerjee, Arindam, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research* 6(Sep):1345–1382.

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

Bloznelis, Mindaugas. 2013. Degree and clustering coefficient in sparse random intersection graphs. *The Annals of Applied Probability* 23(3):1254–1289.

Bloznelis, Mindaugas, Erhard Godehardt, Jerzy Jaworski, Valentas Kurauskas, and Katarzyna Rybarczyk. 2015a. Recent progress in complex network analysis: Models of random intersection graphs. In *Data science, learning by latent structures, and knowledge discovery*, 69–78. Springer.

———. 2015b. Recent progress in complex network analysis: Properties of random intersection graphs. In *Data science, learning by latent structures, and knowledge discovery*, 79–88. Springer.

Bloznelis, Mindaugas, and Valentas Kurauskas. 2016. Clustering coefficient of random intersection graphs with infinite degree variance. *Internet Mathematics* (just-accepted).

Chandrasekhar, Arun G, and Matthew O Jackson. 2014. Tractable and consistent random graph models. Tech. Rep., National Bureau of Economic Research.

Chatterjee, Sourav, Persi Diaconis, et al. 2013. Estimating and understanding exponential random graph models. *The Annals of Statistics* 41(5):2428–2461.

Chung, Fan, and Linyuan Lu. 2006. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics* 3(1):79–127.

Chung, Fan, and Mary Radcliffe. 2011. On the spectra of general random graphs. *the electronic journal of combinatorics* 18(1):P215.

Deijfen, Maria, and Willemien Kets. 2009. Random intersection graphs with tunable degree distribution and clustering. *Probability in the Engineering and Informational Sciences* 23(04):661–674.

Dhillon, Inderjit S, and Dharmendra S Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning* 42(1):143–175.

Gopalan, Prem K, and David M Blei. 2013. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences* 110(36):14534–14539.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5(2):109–137.

Hunter, David R, Steven M Goodreau, and Mark S Handcock. 2008. Goodness of fit of social network models. *Journal of the American Statistical Association* 103(481).

Ji, Pengsheng, and Jiashun Jin. 2014. Coauthorship and citation networks for statisticians. *arXiv preprint arXiv:1410.2840*.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. 2016. An eigenvector perturbation bound and its application to robust covariance estimation. *arXiv preprint http://arxiv.org/pdf/1603.03516.pdf*.

Jin, Jiashun. 2012. Fast network community detection by score. *arXiv preprint arXiv:1211.5803*.

Jin, Jiashun, et al. 2015. Fast community detection by score. *The Annals of Statistics* 43(1):57–89.

Karoński, Michał, Edward R Scheinerman, and Karen B Singer-Cohen. 1999. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing* 8(1-2):131–159.

Karrer, Brian, and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1):016107.

Kleinberg, Jon, Christos Papadimitriou, and Prabhakar Raghavan. 1998. A microeconomic view of data mining. *Data mining and knowledge discovery* 2(4): 311–324.

Kolaczyk, Eric D. 2009. *Statistical analysis of network data*. New York: Springer.

Lei, Jing, and Alessandro Rinaldo. 2013. Consistency of spectral clustering in sparse stochastic block models. *arXiv preprint arXiv:1312.2050*.

Lei, Jing, Alessandro Rinaldo, et al. 2015. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* 43(1):215–237.

Lei, Jing, and Lingxue Zhu. 2014. A generic sample splitting approach for refined community recovery in stochastic block models. *arXiv preprint arXiv:1411.1469*.

Luce, R Duncan, and Albert D Perry. 1949. A method of matrix analysis of group structure. *Psychometrika* 14(2):95–116.

Lyzinski, Vince, Daniel Sussman, Minh Tang, Avanti Athreya, and Carey Priebe. 2013. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *arXiv preprint arXiv:1310.0532*.

Mao, X., P. Sarkar, and D. Chakrabarti. 2017. Estimating Mixed Memberships with Sharp Eigenvector Deviations. *ArXiv e-prints*. 1709.00407.

Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th acm/usenix internet measurement conference (imc'07)*. San Diego, CA.

Newman, Mark EJ. 2006. Modularity and community structure in networks. *Proceedis of the National Academy of Sciences* 103(23):8577–8582.

———. 2010. *Networks: an introduction*. Oxford University Press.

Newman, Mark EJ, and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69(2):026113.

Ng, Andrew. 2000. Cs229 lecture notes. *CS229 Lecture notes* 1(1):1–3.

Ng, Andrew Y, Michael I Jordan, and Yair Weiss. 2001. On spectral clustering1 analysis and an algorithm. *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press* 14:849–856.

Ng, Andrew Y, Michael I Jordan, Yair Weiss, et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2: 849–856.

Qin, Tai, and Karl Rohe. 2013. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in neural information processing systems*, 3120–3128.

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. 2007. An introduction to exponential random graph models for social networks. *Social networks* 29(2): 173–191.

Rohe, Karl, Sourav Chatterjee, Bin Yu, et al. 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* 39(4):1878–1915.

Rubin-Delanchy, P., C. E. Priebe, and M. Tang. 2017. Consistency of adjacency spectral embedding for the mixed membership stochastic blockmodel. *ArXiv e-prints*. 1705.04518.

Sarkar, Purnamrita, Peter J Bickel, et al. 2015. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics* 43(3):962–990.

Shi, Jianbo, and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22(8):888–905.

Singer-Cohen, K. B. 1995. Random intersection graphs. *PhD thesis, Department of Mathematical Sciences, The Johns Hopkins University.*

Tang, Minh, and Carey E. Priebe. 2016. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *https://arxiv.org/pdf/1607.08601v1.pdf.*

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4):395–416.

Wasserman, Stanley, and Katherine Faust. 1994. *Social network analysis: Methods and applications*, vol. 8. Cambridge university press.

Watts, Duncan J, and Steven H Strogatz. 1998. Collective dynamics of ?small-world?networks. *nature* 393(6684):440–442.

White, Halbert. 1982. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society* 1–25.

Yang, Jaewon, and Jure Leskovec. 2012. Defining and evaluating network communities based on ground-truth. *CoRR* abs/1205.6233.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu. 2014. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432.*

Zhao, Yunpeng, Elizaveta Levina, Ji Zhu, et al. 2012. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* 40(4):2266–2292.