DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1179
June 4, 2015

Statistical Justifications for Computationally Tractable Network Data Analysis

Tai Qin[1]
Department of Statistics
University of Wisconsin, Madison

[1]Research of TQ is supported by NSF Grant DMS1308877.

STATISTICAL JUSTIFICATIONS FOR COMPUTATIONALLY TRACTABLE NETWORK DATA ANALYSIS

by

Tai Qin

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 05/15/2015

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

Grace Wahba, IJ Schoenberg-Hilldale Professor, Statistics
Karl Rohe, Assistant Professor, Statistics
Ming Yuan, Professor, Statistics
Sushmita Roy, Assistant Professor, Biostatistics
Xiaojin Zhu, Associate Professor, Computer Science

To my wife Yi Fan, and parents Hanghui Li and Jianren Qin

Abstract

Our lives are embedded in networks. Many researchers wish to analyze these networks to gain a deeper understanding of the underlying mechanisms. Some types of underlying mechanisms generate communities (aka clusters or modularities) in the network. This work aims not merely to devise algorithms for community detection, but also to study the algorithms' estimation properties, to understand if and when we can make justifiable inferences from the estimated communities to the underlying mechanisms.

Spectral clustering is a popular algorithm for community detection. Yet it fails when dealing with networks with degree heterogeneity. Chapter 2 proposes regularized spectral clustering and extends the previous statistical estimation results (Chaudhuri et al. (2012) and Amini et al. (2012)) to the more canonical spectral clustering algorithm in a way that removes any assumption on the minimum degree and provides guidance on the choice of the tuning parameter. Moreover, our results show how the "star shape" in the eigenvectors – a common feature of empirical networks – can be explained by the Degree-Corrected Stochastic Blockmodel and the Extended Planted Partition model, two statistical models that allow for highly heterogeneous degrees.

Chapter 3 extends the regularized spectral clustering algorithm to directed networks with general degrees. Chapter 3 proposes and studies a spectral co-clustering algorithm called DI-SIM. Building on previous spectral co-clustering algorithms (e.g. Dhillon (2001)), DI-SIM incorporates regularization and projection steps. We show that these two steps are essential when there is a large amount of degree heterogeneity and several weakly connected nodes.

Chapter 4 studies a random network model where, ignoring log terms, number of clusters (or communities) K can grow proportionally to number of nodes N. Since the number of clusters must be smaller than the number of nodes, no reasonable model allows K to grow faster; thus, our asymptotic results are the "highest" dimensional. Furthermore, we develop a regularized maximum likelihood estimator that enjoys weak consistency under certain conditions. This is the first work to explicitly introduce

and demonstrate the advantages of statistical regularization in a parametric form for network analysis.

Chapter 5 proposes a fast and memory efficient community detection algorithm. It is built upon regularized spectral clustering and the Nyström extension with proper normalization and regularization. We provide a bound for its misclustering rate under the Degree-Corrected Stochastic Blockmodel.

Finally, chapter 6 studies the interaction between transitivity and sparsity. The two common features in empirical networks, implies that there are local regions of large sparse networks that are dense. We call this the blessing of transitivity and it has consequences for both modeling and inference. Extant research suggests that statistical inference for the Stochastic Blockmodel is more difficult when the edges are sparse. However, this conclusion is confounded by the fact that the asymptotic limit in all of the previous studies is not merely sparse, but also non-transitive. To retain transitivity, the blocks cannot grow faster than the expected degree. Thus, in sparse models, the blocks must remain asymptotically small. Previous algorithmic research demonstrates that small "local" clusters are more amenable to computation, visualization, and interpretation when compared to "global" graph partitions. This paper provides the first statistical results that demonstrate how these small transitive clusters are also more amenable to statistical estimation. Theorem 2 of chapter 6 shows that a "local" clustering algorithm can, with high probability, detect a transitive stochastic block of a fixed size (e.g. 30 nodes) embedded in a large graph. The only constraint on the ambient graph is that it is large and sparse–it could be generated at random or by an adversary–suggesting a theoretical explanation for the robust empirical performance of local clustering algorithms.

Acknowledgments

I am blessed to have Grace and Karl as my two advisors.

In my second year, I felt lost and depressed in the transition from taking classes to doing research. Grace introduced me to the topic of spectral methods which later turn out to be the main theme of this thesis. Research requires creativity and persistence. I am grateful to Grace for giving me infinity degree of freedom for choosing research topics and encouraging and supporting me to go ahead on whichever road I pick.

In my mailbox, there are more than five hundred emails if searching "Karl". Some of his emails displays sending time around 5 to 6am. Karl is always passionate and optimistic about work and life. Some of his energy and attitude toward life passed on me, which makes me a better researcher, and a better person. He is more like a friend than a mentor to me. May 31, 2013 was the submission deadline for a conference. We were working on it to the last minute at Karl's home. When we finished, it was already about 6:30pm. I went downstairs and found his whole family waiting for him to celebrate his little baby's second birthday. We worked on various projects, some failed, some turned out OK. I enjoyed all of them.

I am thankful to my thesis committee: Prof. Ming Yuan, Prof. Sushmita Roy and Prof. Xiaojin (Jerry) Zhu for the helpful advices and suggestions, and more importantly, for letting me pass.

Thank the faculty members for teaching me classes and conversations out side of class. Ming knows everything in statistical theory. I am grateful to him for pushing us aiming higher and thinking deeper. Prof. Wei-Yin Loh introduced me to Kaggle. The online data competition platform opens a window for me to learn practical statistical modeling. Thank Prof. Peter Qian, Prof. Sijian Wang, Prof. Garvesh Raskutt, Prof. Jun Shao, Prof. Douglas Bates, Prof. Chunming Zhang, Prof. Zhengjun Zhang, Prof. Krishnakumar Balasubramanian and Prof. Stephen Wright.

Thank the Thursday's group and the Friday's group for the discussions and inspiring ideas from different prospectives. Thank Zhigeng Geng (also my reliable badminton doubles team partner), Jing Kong, Bin Dai, Shilin Ding, Luwan Zhang, Han Chen, Shulei Wang, Hao Zhou, Yilin Zhang, Xiaowu Dai, and the Friday group:

Juhee Cho, Kim Donggyu, Mohammad Khabbazian, Thu Le, Norbert Binkiewicz, Zoe Russek, Nick Zaborek and Song Wang. Thank my friends, Zhishi, Haoyang, Lilun, Xu, Chandler, Yunfei, Chenliang, Jianan, Yan, Jiajie, Lu, Vincent, Lie, Yaodong and heng. Thank our small friday "practical probability theory training session". Life will be bored without my friends.

At last, thank my wife Yi, for making my life tasty, in every way it can be.

Contents

List of Figures

Chapter 1

Introduction

Recent advances in information technology have produced a deluge of data on complex systems with myriad interacting elements, easily represented by networks or graphs. Communities or clusters of highly connected actors are an essential feature in a multitude of empirical networks, and identifying these clusters helps answer vital questions in various fields. Depending on the area of interest, interacting elements may be metabolites, people, or computers. Their interactions can be represented in chemical reactions, friendships, or some type of communication. For example, on Facebook, groups of people sharing same interest or attending same college form various communities; a terrorist cell is a cluster in the communication network of terrorists; web pages that provide hyperlinks to each other form a community that may host discussions of a similar topic; a cluster in the network of biochemical reactions might contain metabolites with similar functions and activities. Networks (or graphs) appropriately describe these relationships. Therefore, the substantive questions in these various disciplines are, in essence, questions regarding the structure of networks. Given the demonstrated interest in making statistical inference from an observed network, it is essential to evaluate the ability of clustering algorithms to estimate the "true clusters" in a network model. Understanding when and why a clustering algorithm correctly estimates the "true communities" provides a rigorous understanding of the behavior of these algorithms and potentially leads to improved algorithms.

Networks can be complex. Community structures are confounded with many other features of various types of networks:

(a). In many networks, like the world wide web, their degrees are highly heterogeneous, some approximately follow power-law distribution (Adamic and Huberman (2000)). For a sparse network with strong degree heterogeneity,

standard spectral clustering often fails to function properly (Amini et al. (2012); Jin (2015)).

(b). Some networks consist of small communities. Dunbar (1992) suggests that humans do not have the social intellect to maintain stable communities larger than roughly 150 people (colloquially referred to as Dunbar's number). Leskovec et al. (2008) found a similar result in several other networks that were not composed of humans. The research of Leskovec et al. (2008) and Dunbar (1992) suggests that the community sizes should not grow asymptotically.

(c). Some networks are just too large to be even stored in computer memory for further analysis. In this case, it is important to find a balance point in the tradeoff between accuracy and computing resource.

(d). Transitivity and sparsity are two common features of many networks. Extant research suggests that statistical inference for the Stochastic Blockmodel is more difficult when the edges are sparse.

This work aims to devise and study regularized algorithms that detect communities under these confounding issues with theoretical performance guarantee. Chapter 2 introduces regularized spectral clustering(RSC) that improves the performance standard spectral clustering under scenario (a). Chapter 3 extends RSC to directed networks with degree heterogeneity. Chapter 4 studies a regularized maximum likelihood estimator that are proven to be useful under asymptotic settings that mimics scenario (b). Chapter 5 develops and studies a memory efficient spectral algorithm that deals with scenario (c). Lastly chapter 6 studies the interaction between transitivity and sparsity and shows that transitivity can be helpful in finding communities in sparse networks. Chapter 6 then introduces a local clustering algorithm that can, with high probability, detect transitive stochastic block of a fixed size (e.g. 30 nodes) embedded in a large graph.

Each chapter, although highly related to each other, is self-contained. Readers can start from any chapters of interests. Chapter 2 is adapted from Qin and Rohe (2013). Chapter 3 is largely adapted from Rohe, Qin, and Yu (2015). Chapter 4 is adapted from Rohe, Qin, and Fan (2014). Chapter 6 is adapted from Rohe and Qin (2013).

Chapter 2

Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel

2.1 Introduction

Spectral clustering is a fast and popular algorithm for finding clusters in networks. Recently, Chaudhuri et al. (2012) and Amini et al. (2012) proposed inspired variations on the algorithm that artificially inflate the node degrees for improved statistical performance. This chapter extends the previous statistical estimation results to the more canonical spectral clustering algorithm in a way that removes any assumption on the minimum degree and provides guidance on the choice of the tuning parameter. Moreover, our results show how the "star shape" in the eigenvectors – a common feature of empirical networks – can be explained by the Degree-Corrected Stochastic Blockmodel and the Extended Planted Partition model, two statistical models that allow for highly heterogeneous degrees. Throughout, this chapter characterizes and justifies several of the variations of the spectral clustering algorithm in terms of these models.

Several previous authors have studied the estimation properties of spectral clustering under various statistical network models (McSherry (2001); Dasgupta et al. (2004); Coja-Oghlan and Lanka (2009); Ames and Vavasis (2010); Rohe et al. (2011); Sussman et al. (2012b) and Chaudhuri et al. (2012)). Recently, Chaudhuri et al. (2012) and Amini et al. (2012) proposed two inspired ways of artificially inflating the node degrees in ways that provide statistical regularization to spectral clustering.

This chapter examines the statistical estimation performance of regularized spectral

clustering under the Degree-Corrected Stochastic Blockmodel (DC-SBM), an extension of the Stochastic Blockmodel (SBM) that allows for heterogeneous degrees (Holland and Leinhardt (1983); Karrer and Newman (2011)). The SBM and the DC-SBM are closely related to the planted partition model and the extended planted partition model, respectively. We extend the previous results in the following ways:

(a) In contrast to previous studies, this paper studies the regularization step with a canonical version of spectral clustering that uses k-means. The results do not require any assumptions on the minimum expected node degree; instead, there is a threshold demonstrating that higher degree nodes are easier to cluster. This threshold is a function of the leverage scores that have proven essential in other contexts, for both graph algorithms and network data analysis (see Mahoney (2012a) and references therein). These are the first results that relate leverage scores to the statistical performance of spectral clustering.

(b) This paper provides more guidance for data analytic issues than previous approaches. First, the results suggest an appropriate range for the regularization parameter. Second, our analysis gives a (statistical) model-based explanation for the "star-shaped" figure that often appears in empirical eigenvectors. This demonstrates how projecting the rows of the eigenvector matrix onto the unit sphere (an algorithmic step proposed by Ng et al. (2002)) removes the ancillary effects of heterogeneous degrees under the DC-SBM. Our results highlight when this step may be unwise.

Preliminaries: Throughout, we study undirected and unweighted graphs or networks. Define a graph as $G(E, V)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is the vertex or node set and E is the edge set. We will refer to node v_i as node i. E contains a pair (i, j) if there is an edge between node i and j. The edge set can be represented by the adjacency matrix $A \in \{0, 1\}^{n \times n}$. $A_{ij} = A_{ji} = 1$ if (i, j) is in the edge set and $A_{ij} = A_{ji} = 0$ otherwise. Define the diagonal matrix D and the normalized Graph Laplacian L, both elements of $\mathbb{R}^{N \times N}$, in the following way:

$$D_{ii} = \sum_j A_{ij}, \qquad L = D^{-1/2} A D^{-1/2}.$$

The following notations will be used throughout the paper: $||\cdot||$ denotes the spectral norm, and $||\cdot||_F$ denotes the Frobenius norm. For two sequence of variables $\{x_N\}$ and $\{y_N\}$, we say $x_N = \omega(y_N)$ if and only if $y_N/x_N = o(1)$. $\delta_{(.,.)}$ is the indicator function where $\delta_{x,y} = 1$ if $x = y$ and $\delta_{x,y} = 0$ if $x \neq y$.

2.2 The Algorithm: Regularized Spectral Clustering (RSC)

For a sparse network with strong degree heterogeneity, standard spectral clustering often fails to function properly (Amini et al. (2012); Jin (2015)). To account for this, Chaudhuri et al. (2012) proposed the regularized graph Laplacian that can be defined as

$$L_\tau = D_\tau^{-1/2} A D_\tau^{-1/2} \in \mathbb{R}^{N \times N}$$

where $D_\tau = D + \tau I$ for $\tau \geq 0$.

The spectral algorithm proposed and studied by Chaudhuri et al. (2012) divides the nodes into two random subsets and only uses the induced subgraph on one of those random subsets to compute the spectral decomposition. In this paper, we will study the more traditional version of spectral algorithm that uses the spectral decomposition on the entire matrix (Ng et al. (2002)). Define the regularized spectral clustering (RSC) algorithm as follows:

RSC

Input: Adjacency matrix $A \in \{0,1\}^{n \times n}$, regularizer $\tau \geq 0$ (Default: $\tau =$ average node degree), number of clusters K.

1. Given input adjacency matrix A, number of clusters K, and regularizer τ, calculate the regularized graph Laplacian L_τ. (As discussed later, a good default for τ is the average node degree.)

2. Find the eigenvectors $X_1, ..., X_K \in \mathbb{R}^N$ corresponding to the K largest eigenvalues of L_τ. Form $X = [X_1, ..., X_K] \in \mathbb{R}^{N \times K}$ by putting the eigenvectors into the columns.

3. Form the matrix $X^* \in \mathbb{R}^{N \times K}$ from X by normalizing each of X's rows to have unit length. That is, project each row of X onto the unit sphere of $\mathbb{R}^K$ ($X^*_{ij} = X_{ij}/(\sum_j X^2_{ij})^{1/2}$).

4. Treat each row of X^* as a point in $\mathbb{R}^K$, and run k-means with K clusters. This creates K non-overlapping sets $V_1, ..., V_K$ whose union is V.

5. Node i is assigned to cluster r if the i'th row of X^* is assigned to V_r.

Output: The clusters $V_1, ..., V_K$ from step (5).

This paper will refer to "standard spectral clustering" as the above algorithm with L replacing L_τ.

These spectral algorithms have two main steps: 1) find the principal eigenspace of the (regularized) graph Laplacian; 2) determine the clusters in the low dimensional eigenspace. Later, we will study RSC under the Degree-Corrected Stochastic Blockmodel and show rigorously how regularization helps to maintain cluster information in step (a) and why normalizing the rows of X helps in step (b). From now on, we use X_τ and X^*_τ instead of X and X^* to emphasize that they are related to L_τ. Let X^i_τ and $[X^*_\tau]^i$ denote the i'th row of X_τ and X^*_τ.

The next section introduces the Degree-Corrected Stochastic Blockmodel and its matrix formulation.

2.3 The Degree-Corrected Stochastic Blockmodel (DC-SBM)

In the Stochastic Blockmodel (SBM), each node belongs to one of K blocks. Each edge corresponds to an independent Bernoulli random variable where the probability of an edge between any two nodes depends only on the block memberships of the two nodes (Holland and Leinhardt (1983)). The formal definition is as follows.

Definition 2.1. *For a node set $\{1, 2, ..., N\}$, let $z : \{1, 2, ..., N\} \to \{1, 2, ..., K\}$ partition the N nodes into K blocks. So, z_i equals the block membership for node i. Let $\boldsymbol{B}$ be a $K \times K$ matrix where $\boldsymbol{B}_{ab} \in [0, 1]$ for all a, b. Then under the SBM, the probability of an edge between i and j is $P_{ij} = P_{ji} = \boldsymbol{B}_{z_i z_j}$ for any $i, j = 1, 2, ..., n$. Given z, all edges are independent.*

One limitation of the SBM is that it presumes all nodes within the same block have the same expected degree. The Degree-Corrected Stochastic Blockmodel (DC-SBM) (Karrer and Newman (2011)) is a generalization of the SBM that adds an additional set of parameters ($\theta_i > 0$ for each node i) that control the node degrees. Let $\mathbf{B}$ be a $K \times K$ matrix where $\mathbf{B}_{ab} \geq 0$ for all a, b. Then the probability of an edge between node i and node j is $\theta_i \theta_j \mathbf{B}_{z_i z_j}$, where $\theta_i \theta_j \mathbf{B}_{z_i z_j} \in [0, 1]$ for any $i, j = 1, 2, ..., n$. Parameters θ_i are arbitrary to within a multiplicative constant that is absorbed into $\mathbf{B}$. To make it identifiable, Karrer and Newman (2011) suggest imposing the constraint that, within each block, the summation of θ_i's is 1. That is, $\sum_i \theta_i \delta_{z_i, r} = 1$ for any block label r. Under this constraint, $\mathbf{B}$ has explicit meaning: If $s \neq t$, $\mathbf{B}_{st}$ represents the expected number of links between block s and block t and if $s = t$,

$\mathbf{B}_{st}$ is twice the expected number of links within block s. Throughout the paper, we assume that $\mathbf{B}$ is positive definite.

Under the DC-SBM, define $\mathscr{A} \triangleq \mathbb{E}A$. This matrix can be expressed as a product of the matrices,

$$\mathscr{A} = \Theta Z \mathbf{B} Z^T \Theta,$$

where (1) $\Theta \in \mathbb{R}^{N \times N}$ is a diagonal matrix whose ii'th element is θ_i and (2) $Z \in \{0,1\}^{N \times K}$ is the membership matrix with $Z_{it} = 1$ if and only if node i belongs to block t (i.e. $z_i = t$).

Population Analysis

Under the DC-SBM, if the partition is identifiable, then one should be able to determine the partition from $\mathscr{A}$. This section shows that with the population adjacency matrix $\mathscr{A}$ and a proper regularizer τ, RSC perfectly reconstructs the block partition.

Define the diagonal matrix $\mathscr{D}$ to contain the expected node degrees, $\mathscr{D}_{ii} = \sum_j \mathscr{A}_{ij}$ and define $\mathscr{D}_\tau = \mathscr{D} + \tau I$ where $\tau \geq 0$ is the regularizer. Then, define the population graph Laplacian $\mathscr{L}$ and the population version of regularized graph Laplacian $\mathscr{L}_\tau$, both elements of $\mathbb{R}^{N \times N}$, in the following way:

$$\mathscr{L} = \mathscr{D}^{-1/2} \mathscr{A} \mathscr{D}^{-1/2}, \qquad \mathscr{L}_\tau = \mathscr{D}_\tau^{-1/2} \mathscr{A} \mathscr{D}_\tau^{-1/2}.$$

Define $D_B \in \mathbb{R}^{K \times K}$ as a diagonal matrix whose (s,s)'th element is $[D_B]_{ss} = \sum_t B_{st}$. A couple lines of algebra shows that $[D_B]_{ss} = W_s$ is the total expected degrees of nodes from block s and that $\mathscr{D}_{ii} = \theta_i [D_B]_{z_i z_i}$. Using these quantities, the next Lemma gives an explicit form for $\mathscr{L}_\tau$ as a product of the parameter matrices.

Lemma 2.2. *(Explicit form for $\mathscr{L}_\tau$) Under the DC-SBM with K blocks with parameters $\{\boldsymbol{B}, Z, \Theta\}$, define θ_i^τ as:*

$$\theta_i^\tau = \frac{\theta_i^2}{\theta_i + \tau / W_{z_i}} = \theta_i \frac{\mathscr{D}_{ii}}{\mathscr{D}_{ii} + \tau}.$$

Let $\Theta_\tau \in \mathbb{R}^{n \times n}$ be a diagonal matrix whose ii'th entry is θ_i^τ. Define $B_L = D_B^{-1/2} B D_B^{-1/2}$, then $\mathscr{L}_\tau$ can be written

$$\mathscr{L}_\tau = \mathscr{D}_\tau^{-\frac{1}{2}} \mathscr{A} \mathscr{D}_\tau^{-\frac{1}{2}} = \Theta_\tau^{\frac{1}{2}} Z B_L Z^T \Theta_\tau^{\frac{1}{2}}.$$

Recall that $\mathscr{A} = \Theta Z \mathbf{B} Z^T \Theta$. Lemma 3.3 demonstrates that $\mathscr{L}_\tau$ has a similarly simple form that separates the block-related information (B_L) and node specific information (Θ_τ). Notice that if $\tau = 0$, then $\Theta_0 = \Theta$ and $\mathscr{L} = \mathscr{D}^{-\frac{1}{2}} \mathscr{A} \mathscr{D}^{-\frac{1}{2}} = \Theta^{\frac{1}{2}} Z B_L Z^T \Theta^{\frac{1}{2}}$. The next lemma shows that $\mathscr{L}_\tau$ has rank K and describes how its eigen-decomposition can be expressed in terms of Z and Θ.

Lemma 2.3. *(Eigen-decomposition for $\mathscr{L}_\tau$) Under the DC-SBM with K blocks and parameters $\{\boldsymbol{B}, Z, \Theta\}$, $\mathscr{L}_\lambda$ has K positive eigenvalues. The remaining $N - K$ eigenvalues are zero. Denote the K positive eigenvalues of $\mathscr{L}_\tau$ as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$ and let $\mathscr{X}_\tau \in \mathbb{R}^{N \times K}$ contain the eigenvector corresponding to λ_i in its i'th column. Define $\mathscr{X}_\tau^*$ to be the row-normalized version of $\mathscr{X}_\tau$, similar to X_τ^* as defined in the RSC algorithm in Section 2. Then, there exists an orthogonal matrix $U \in \mathbb{R}^{K \times K}$ depending on τ, such that*

1. $\mathscr{X}_\tau = \Theta_\tau^{\frac{1}{2}} Z (Z^T \Theta_\tau Z)^{-1/2} U$
2. $\mathscr{X}_\tau^* = ZU$, $Z_i \neq Z_j \Leftrightarrow Z_i U \neq Z_j U$, *where Z_i denote the i'th row of the membership matrix Z.*

This lemma provides four useful facts about the matrices $\mathscr{X}_\tau$ and $\mathscr{X}_\tau^*$. First, if two nodes i and j belong to the same block, then the corresponding rows of $\mathscr{X}_\tau$ (denoted as $\mathscr{X}_\tau^i$ and $\mathscr{X}_\tau^j$) both point in the same direction, but with different lengths: $||\mathscr{X}_\tau^i||_2 = (\frac{\theta_i^\tau}{\sum_j \theta_j^\tau \delta_{z_j, z_i}})^{1/2}$. Second, if two nodes i and j belong to different blocks, then $\mathscr{X}_\tau^i$ and $\mathscr{X}_\tau^j$ are orthogonal to each other. Third, if $z_i = z_j$ then after projecting these points onto the sphere as in $\mathscr{X}_\tau^*$, the rows are equal: $[\mathscr{X}_\tau^*]^i = [\mathscr{X}_\tau^*]^j = U_{z_i}$. Finally, if $z_i \neq z_j$, then the rows are perpendicular, $[\mathscr{X}_\tau^*]^i \perp [\mathscr{X}_\tau^*]^j$. Figure 1 illustrates the geometry of $\mathscr{X}_\tau$ and $\mathscr{X}_\tau^*$ when there are three underlying blocks. Notice that running k-means on the rows of $\mathscr{X}_\lambda^*$ (in right panel of Figure 1) will return perfect clusters.

Note that if Θ were the identity matrix, then the left panel in Figure 1 would look like the right panel in Figure 1; without degree heterogeneity, there would be no star shape and no need for a projection step. This suggests that the star shaped figure often observed in data analysis stems from the degree heterogeneity in the network.

2.4 Regularized Spectral Clustering with the Degree-corrected model

This section bounds the mis-clustering rate of Regularized Spectral Clustering under the DC-SBM. The section proceeds as follows: Theorem 2.4 shows that L_τ is close to

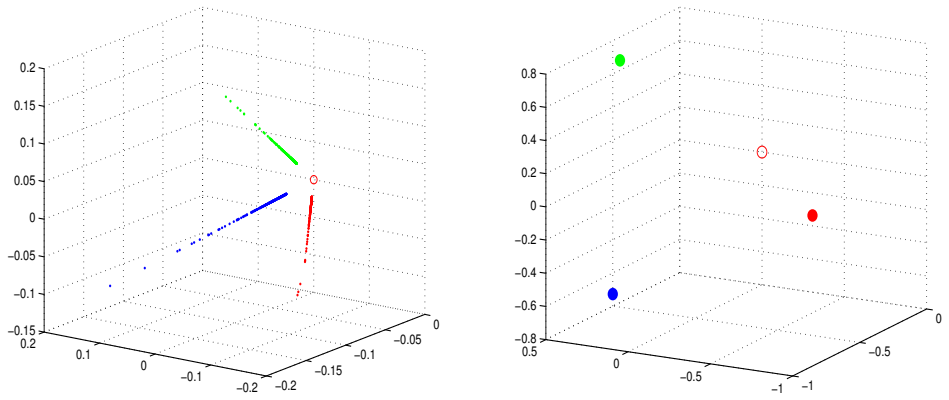

Figure 2.1: In this numerical example, $\mathscr{A}$ comes from the DC-SBM with three blocks. Each point corresponds to one row of the matrix $\mathscr{X}_\tau$ (in left panel) or $\mathscr{X}_\tau^*$ (in right panel). The different colors correspond to three different blocks. The hollow circle is the origin. Without normalization (left panel), the nodes with same block membership share the **s**ame direction in the projected space. After normalization (right panel), nodes with same block membership share the **s**ame position in the projected space.

$\mathscr{L}_\tau$. Theorem 2.5 shows that X_τ is close to $\mathscr{X}_\tau$ and that X_τ^* is close to $\mathscr{X}_\tau^*$. Finally, Theorem 2.7 shows that the output from RSC with L_τ is close to the true partition in the DC-SBM (using Lemma 2.3).

Theorem 2.4. *(Concentration of the regularized Graph Laplacian) Let G be a random graph, with independent edges and $pr(v_i \sim v_j) = p_{ij}$. Let δ be the minimum expected degree of G, that is $\delta = \min_i \mathscr{D}_{ii}$. For any $\epsilon > 0$, if $\delta + \tau > 3\ln N + 3\ln(4/\epsilon)$, then with probability at least $1 - \epsilon$,*

$$||L_\tau - \mathscr{L}_\tau|| \leq 4\sqrt{\frac{3\ln(4N/\epsilon)}{\delta + \tau}}. \tag{2.1}$$

Remark: This theorem builds on the results of Chung and Radcliffe (2011) and Chaudhuri et al. (2012) which give a seemingly similar bound on $||L - \mathscr{L}||$ and $||D_\tau^{-1}A - \mathscr{D}_\tau^{-1}\mathscr{A}||$. However, the previous papers require that $\delta \geq c\ln N$, where c is some constant. This assumption is not satisfied in a large proportion of sparse empirical networks with heterogeneous degrees. In fact, the regularized graph Laplacian is most interesting when this condition fails, i.e. when there are several

nodes with very low degrees. Theorem 2.4 only assumes that $\delta+\tau > 3\ln N + 3\ln(4/\epsilon)$. This is the fundamental reason that RSC works for networks containing some nodes with extremely small degrees. It shows that, by introducing a proper regularizer τ, $||L_\tau - \mathscr{L}_\tau||$ can be well bounded, even with δ very small. Later we will show that a suitable choice of τ is the average degree.

The next theorem bounds the difference between the empirical and population eigenvectors (and their row normalized versions) in terms of the Frobenius norm.

Theorem 2.5. *Let A be the adjacency matrix generated from the DC-SBM with K blocks and parameters $\{\boldsymbol{B}, Z, \Theta\}$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$ be the only K positive eigenvalues of $\mathscr{L}_\tau$. Let X_τ and $\mathscr{X}_\tau \in \mathbb{R}^{N\times K}$ contain the top K eigenvectors of L_τ and $\mathscr{L}_\tau$ respectively. Define $m = \min_i\{||\mathscr{X}_\tau^i||_2\}$ as the length of the shortest row in $\mathscr{X}_\tau$. Let X_τ^* and $\mathscr{X}_\tau^* \in \mathbb{R}^{N\times K}$ be the row normalized versions of X_τ and $\mathscr{X}_\tau$, as defined in step 3 of the RSC algorithm.*

For any $\epsilon > 0$ and sufficiently large N, assume that $\delta + \tau > 3\ln N + 3\ln(4/\epsilon)$, then with probability at least $1-\epsilon$, the following holds,

$$||X_\tau - \mathscr{X}_\tau\mathscr{O}||_F \leq c_0\frac{1}{\lambda_K}\sqrt{\frac{K\ln(4N/\epsilon)}{\delta+\tau}}, \quad \textit{and} \quad ||X_\tau^* - \mathscr{X}_\tau^*\mathscr{O}||_F \leq c_0\frac{1}{m\lambda_K}\sqrt{\frac{K\ln(4N/\epsilon)}{\delta+\tau}}. \tag{2.2}$$

The proof of Theorem 2.5 can be found in the appendix.

Next we use Theorem 2.5 to derive a bound on the mis-clustering rate of RSC. To define "mis-clustered", recall that RSC applies the k-means algorithm to the rows of X_τ^*, where each row is a point in $\mathbb{R}^K$. Each row is assigned to one cluster, and each of these clusters has a centroid from k-means. Define $C_1, \ldots, C_n \in \mathbb{R}^K$ such that C_i is the centroid corresponding to the i'th row of X_τ^*. Similarly, run k-means on the rows of the population eigenvector matrix $\mathscr{X}_\tau^*$ and define the population centroids $\mathcal{C}_1, \ldots, \mathcal{C}_n \in \mathbb{R}^K$. In essence, we consider node i correctly clustered if C_i is closer to $\mathcal{C}_i$ than it is to any other $\mathcal{C}_j$ for all j with $Z_j \neq Z_i$.

The definition is complicated by the fact that, if any of the $\lambda_1, \ldots, \lambda_K$ are equal, then only the subspace spanned by their eigenvectors is identifiable. Similarly, if any of those eigenvalues are close together, then the estimation results for the individual eigenvectors are much worse that for the estimation results for the subspace that they span. Because clustering only requires estimation of the correct subspace, our definition of correctly clustered is amended with the rotation $\mathscr{O}^T \in R^{K\times K}$, the matrix which minimizes $\|X_\tau^*\mathscr{O}^T - \mathscr{X}_\tau^*\|_F$. This is referred to as the orthogonal Procrustes problem and Schönemann (1966) shows how the singular value decomposition gives the solution.

Definition 2.6. *If $C_i\mathscr{O}^T$ is closer to $\mathcal{C}_i$ than it is to any other $\mathcal{C}_j$ for j with $Z_j \neq Z_i$, then we say that node i is correctly clustered. Define the set of mis-clustered nodes:*

$$\mathscr{M} = \{i : Existsj \neq i, s.t. ||C_i\mathscr{O}^T - \mathcal{C}_i||_2 > ||C_i\mathscr{O}^T - \mathcal{C}_j||_2\}. \tag{2.3}$$

The next theorem bounds the mis-clustering rate $|\mathscr{M}|/N$.

Theorem 2.7. ***(Main Theorem)*** *Suppose $A \in \mathbb{R}^{N\times N}$ is an adjacency matrix of a graph G generated from the DC-SBM with K blocks and parameters $\{\boldsymbol{B}, Z, \Theta\}$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$ be the K positive eigenvalues of $\mathscr{L}_\tau$. Define $\mathscr{M}$, the set of mis-clustered nodes, as in Definition 2.6. Let δ be the minimum expected degree of G. For any $\epsilon > 0$ and sufficiently large N, assume (a) and (b) as in Theorem 2.5. Then with probability at least $1-\epsilon$, the mis-clustering rate of RSC with regularization constant τ is bounded,*

$$|\mathscr{M}|/N \leq c_1 \frac{K \ln(N/\epsilon)}{Nm^2(\delta+\tau)\lambda_K^2}. \tag{2.4}$$

Remark 1 (Choice of τ): The quality of the bound in Theorem 2.7 depends on τ through three terms: $(\delta+\tau), \lambda_K$, and m. Setting τ equal to the average node degree balances these terms. In essence, if τ is too small, there is insufficient regularization. Specifically, if the minimum expected degree $\delta = O(\ln N)$, then we need $\tau \geq c(\epsilon)\ln N$ to have enough regularization to satisfy condition (b) on $\delta+\tau$. Alternatively, if τ is too large, it washes out significant eigenvalues.

To see that τ should not be too large, note that

$$C = (Z^T\Theta_\tau Z)^{1/2} B_L (Z^T\Theta_\tau Z)^{1/2} \in \mathbb{R}^{K\times K} \tag{2.5}$$

has the same eigenvalues as the largest K eigenvalues of $\mathscr{L}_\tau$ (see supplementary materials for details). The matrix $Z^T\Theta_\tau Z$ is diagonal and the (s,s)'th element is the summation of θ_i^τ within block s. If $\mathrm{E}M = \omega(N\ln N)$ where $M = \sum_i D_{ii}$ is the sum of the node degrees, then $\tau = \omega(M/N)$ sends the smallest diagonal entry of $Z^T\Theta_\tau Z$ to 0, sending λ_K, the smallest eigenvalue of C, to zero.

The trade-off between these two suggests that a proper range of τ is $(\alpha\frac{\mathbb{E}M}{N}, \beta\frac{\mathbb{E}M}{N})$, where $0 < \alpha < \beta$ are two constants. Keeping τ within this range guarantees that λ_K is lower bounded by some constant depending only on K. In simulations, we find that $\tau = M/N$ (i.e. the average node degree) provides good results. The theoretical results only suggest that this is the correct rate. So, one could adjust this by a multiplicative constant. Our simulations suggest that the results are not sensitive to such adjustments.

Remark 2 (Thresholding m): Mahoney (2012a) (and references therein) shows how the leverage scores of A and L are informative for both data analysis and algorithmic stability. For L, the leverage score of node i is $||X^i||_2^2$, the length of the ith row of the matrix containing the top K eigenvectors. Theorem 2.7 is the first result that explicitly relates the leverage scores to the statistical performance of spectral clustering. Recall that m^2 is the minimum of the squared row lengths in $\mathscr{X}_\tau$, that is the minimum leverage score in both $\mathscr{L}_\tau$. This appears in the denominator of (5.7). The leverage scores in $\mathscr{L}_\tau$ have an explicit form

$$||\mathscr{X}_\tau^i||_2^2 = \frac{\theta_i^\tau}{\sum_j \theta_j^\tau \delta_{z_j,z_i}}.$$

So, if node i has small expected degree, then θ_i^τ is small, rendering $||\mathscr{X}_\tau^i||_2$ small. This can deteriorate the bound in Theorem 2.7. The problem arises from projecting X_τ^i onto the unit sphere for a node i with small leverage; it amplifies a noisy measurement. Motivated by this intuition, the next corollary focuses on the high leverage nodes. More specifically, let m^* denote the threshold. Define S to be a subset of nodes whose leverage scores in $\mathscr{L}_\tau$, $||\mathscr{X}_\tau^i||$ exceed the threshold m^*:

$$S = \{i : ||\mathscr{X}_\tau^i|| \geq m^*\}.$$

Then by applying k-means on the set of vectors $\{[X_\tau^*]^i, i \in S\}$, we cluster these nodes. The following corollary bounds the mis-clustering rate on S.

Corollary 2.8. *Let $N_1 = |S|$ denote the number of nodes in S and define $\mathscr{M}_1 = \mathscr{M} \cap S$ as the set of mis-clustered nodes restricted in S. With the same settings and assumptions as in Theorem 2.7, let $\gamma > 0$ be a constant and set $m^* = \gamma/\sqrt{N}$. If $N/N_1 = O(1)$, then by applying k-means on the set of vectors $\{[X_\tau^*]^i, i \in S\}$, we have with probability at least $1 - \epsilon$, there exist constant c_2 independent of ϵ, such that*

$$|\mathscr{M}_1|/N_1 \leq c_2 \frac{K \ln(N_1/\epsilon)}{\gamma^2(\delta + \tau)\lambda_K^2}. \tag{2.6}$$

In the main theorem (Theorem 2.7), the denominator of the upper bound contains m^2. Since we do not make a minimum degree assumption, this value potentially approaches zero, making the bound useless. Corollary 2.8 replaces Nm^2 with the constant γ^2, providing a superior bound when there are several small leverage scores.

If λ_K (the Kth largest eigenvalue of $\mathscr{L}_\tau$) is bounded below by some constant and $\tau = \omega(\ln N)$, then Corollary 2.8 implies that $|\mathscr{M}_1|/N_1 = o_p(1)$. The above

thresholding procedure only clusters the nodes in S. To cluster all of the nodes, define the thresholded RSC (t-RSC) as follows:

(a) Follow step (1), (2), and (3) of RSC as in section 2.1.

(b) Apply k-means with K clusters on the set $\mathrm{S} = \{i, ||X_\tau^i||_2 \geq \gamma/\sqrt{N}\}$ and assign each of them to one of $V_1, ..., V_K$. Let $C_1, ..., C_K$ denote the K centroids given by k-means.

(c) For each node $i \notin S$, find the centroid C_s such that $|||[X_\tau^*]^i - C_s||_2 = min_{1 \leq t \leq K} |||[X_\tau^*]^i - C_t||_2$. Assign node i to V_s.

(d) Output $V_1, ...V_K$.

Remark 3 (Applying to SC): Theorem 2.7 can be easily applied to the standard SC algorithm under both the SBM and the DC-SBM by setting $\tau = 0$. In this setting, Theorem 2.7 improves upon the previous results for spectral clustering.

Define the four parameter Stochastic Blockmodel $SBM(p, r, s, K)$ as follows: p is the probability of an edge occurring between two nodes from the same block, r is the probability of an out-block linkage, s is the number of nodes within each block, and K is the number of blocks.

Because the SBM lacks degree heterogeneity within blocks, the rows of $\mathscr{X}$ within the same block already share the same length. So, it is not necessary to project X^i's to the unit sphere. Under the four parameter model, $\lambda_K = (K[r/(p-r)] + 1)^{-1}$ (Rohe et al. (2011)). Using Theorem 2.7, with p and r fixed and $p > r$, and applying k-means to the rows of X, we have

$$|\mathscr{M}|/N = O_p\left(\frac{K^2 \ln N}{N}\right). \tag{2.7}$$

If $K = o(\sqrt{\frac{N}{\ln N}})$, then $|\mathscr{M}|/N \to 0$ in probability. This improves the previous results that required $K = o(N^{1/3})$ (Rohe et al. (2011)). Moreover, it makes the results for spectral clustering comparable to the results for the MLE in Choi et al. (2012).

2.5 Simulation and Analysis of Political Blogs

This section compares five different methods of spectral clustering. Experiment 1 generates networks from the DC-SBM with a power-law degree distribution. Experiment 2 generates networks from the standard SBM. Finally, the benefits of regularization

are illustrated on an empirical network from the political blogosphere during the 2004 presidential election (Adamic and Glance (2005)).

The simulations compare (1) standard spectral clustering (SC), (2) RSC as defined in section 2, (3) RSC without projecting X_τ onto unit sphere (RSC_wp), (4) regularized SC with thresholding (t-RSC), and (5) spectral clustering with perturbation (SCP) (Amini et al. (2012)) which applies SC to the perturbed adjacency matrix $A_{per} = A + a11^T$. In addition, experiment 2 compares the performance of RSC on the subset of nodes with high leverage scores (RSC on S) with the other 5 methods. We set $\tau = M/N$, threshold parameter $\gamma = 1$, and $a = M/N^2$ except otherwise specified.

Experiment 1

This experiment examines how degree heterogeneity affects the performance of the spectral clustering algorithms. The Θ parameters (from the DC-SBM) are drawn from the power law distribution with lower bound $x_{min} = 1$ and shape parameter $\beta \in \{2, 2.25, 2.5, 2.75, 3, 3.25, 3.5\}$. A smaller β indicates to greater degree heterogeneity. For each fixed β, thirty networks are sampled. In each sample, $K = 3$ and each block contains 300 nodes ($N = 900$). Define the signal to noise ratio to be the expected number of in-block edges divided by the expected number of out-block edges. Throughout the simulations, the SNR is set to four and the expected average degree is set to eight.

The left panel of Figure 2 plots β against the misclustering rate for SC, RSC, RSC_wp, t-RSC, SCP and RSC on S. Each point is the average of 30 sampled networks. Each line represents one method. If a method assigns more than 95% of the nodes into one block, then we consider all nodes to be misclustered. The experiment shows that (1) if the degrees are more heterogeneous ($\beta \leq 3.5$), then regularization improves the performance of the algorithms; (2) if $\beta < 3$, then RSC and t-RSC outperform RSC_wp and SCP, verifying that the normalization step helps when the degrees are highly heterogeneous; and, finally, (3) uniformly across the setting of β, it is easier to cluster nodes with high leverage scores.

Experiment 2

This experiment compares SC, RSC, RSC_wp, t-RSC and SCP under the SBM with no degree heterogeneity. Each simulation has $K = 3$ blocks and $N = 1500$ nodes. As in the previous experiment, SNR is set to four. In this experiment, the average degree has three different settings: $10, 21, 30$. For each setting, the results are averaged over 50 samples of the network.

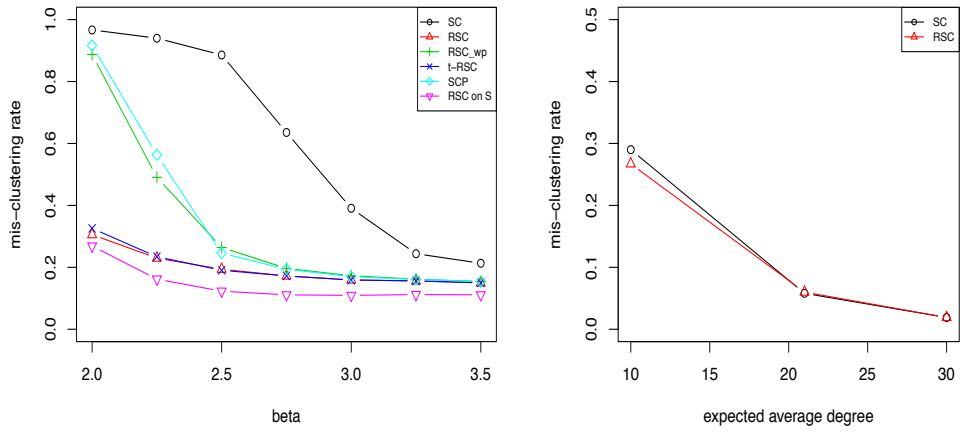

Figure 2.2: Left Panel: Comparison of Performance for SC, RSC, RSC_wp, t-RSC, SCP and (RSC on S) under different degree heterogeneity. Smaller β corresponds to greater degree heterogeneity. Right Panel: Comparison of Performance for SC and RSC under SBM with different sparsity.

The right panel of Figure 2 shows the misclustering rate of SC and RSC for the three different values of the average degree. SCP, RSC_wp, t-RSC perform similarly to RSC, demonstrating that under the standard SBM (i.e. without degree heterogeneity) all spectral clustering methods perform comparably. The one exception is that under the sparsest model, SC is less stable than the other methods.

Analysis of Blog Network

This empirical network is comprised of political blogs during the 2004 US presidential election (Adamic and Glance (2005)). Each blog has a known label as liberal or conservative. As in Karrer and Newman (2011), we symmetrize the network and consider only the largest connected component of 1222 nodes. The average degree of the network is roughly 15. We apply RSC to the data set with τ ranging from 0 to 30. In the case where $\tau = 0$, it is standard Spectral Clustering. SC assigns 1144 out of 1222 nodes to the same block, failing to detect the ideological partition. RSC detects the partition, and its performance is insensitive to the τ. With $\tau \in [1, 30]$, RSC misclusters (80 ± 2) nodes out of 1222.

If RSC is applied to the 90% of nodes with the largest leverage scores (i.e. excluding the nodes with the smallest leverage scores), then the misclustering rate among these high leverage nodes is 44/1100, which is almost 50% lower. This illustrates how the leverage score corresponding to a node can gauge the strength of the clustering evidence for that node relative to the other nodes.

We tried to compare these results t the regularized algorithm in Chaudhuri et al. (2012). However, because there are several very small degree nodes in this data, the values computed in step 4 of the algorithm in Chaudhuri et al. (2012) sometimes take negative values. Then, step 5 (b) cannot be performed.

2.6 Discussion

In this chapter, we give theoretical, simulation, and empirical results that demonstrate how a simple adjustment to the standard spectral clustering algorithm can give dramatically better results for networks with heterogeneous degrees. Our theoretical results add to the current results by studying the regularization step in a more canonical version of the spectral clustering algorithm. Moreover, our main results require no assumptions on the minimum node degree. This is crucial because it allows us to study situations where several nodes have small leverage scores; in these situations, regularization is most beneficial. Finally, our results demonstrate that choosing a tuning parameter close to the average degree provides a balance between several competing objectives.

Chapter 3

Regularized Co-clustering for Directed Graphs

3.1 Introduction

Co-clustering (a.k.a. bi-clustering) was first proposed in Hartigan (1972) for data arranged in a matrix $M \in \mathbb{R}^{n \times d}$. In addition to clustering the rows of M into k_r clusters, co-clustering simultaneously clusters the columns of M into k_c clusters. In the past decade, co-clustering has become an important data analytic technique in biological applications (e.g. Madeira and Oliveira (2004), Tanay et al. (2004), Tanay et al. (2005), Madeira et al. (2010)), text processing (e.g. Dhillon (2001), Bisson and Hussain (2008)), and natural language processing (e.g. Freitag (2004), Rohwer and Freitag (2004)). In these settings, Banerjee et al. (2004) describes how co-clustering dramatically reduces the number of parameters that one needs to estimate. This leads to three advantages over traditional clustering: (1) more interpretable results, (2) faster computation, and (3) implicit statistical regularization.

Previous applications of co-clustering have involved matrices where the rows and columns index different sets of objects. For example, in text processing, the rows correspond to documents, and the columns correspond to words. Element i, j of this matrix denotes how many times word j appears in document i. The row clusters correspond to clusters of similar documents and the column clusters correspond to clusters of similar words. In contrast, this paper applies co-clustering to a matrix where the rows and columns index the same set of nodes. The ith row of the matrix identifies the *outgoing* edges for node i; two nodes are in the same row cluster if they send edges to several of the same nodes. The ith column of this matrix identifies the *incoming* edges for node i; two nodes are in the same row co-cluster if they send edges

to several of the same nodes. As such, each node i is in two types of clusters (one for the ith column and one for the ith row). Comparing these two distinct partitions of the nodes can lead to novel insights when compared to the standard co-clustering applications where the rows and columns index different sets.

This paper proposes and studies a spectral co-clustering algorithm called DI-SIM. Building on previous spectral co-clustering algorithms (e.g. Dhillon (2001)), DI-SIM incorporates regularization and projection steps. These two steps are essential when there is a large amounts of degree heterogeneity and several weakly connected nodes. The name DI-SIM has three meanings. First, because DI-SIM co-clusters the nodes, it uses two distinct (but related) similarity measures between nodes: "the number of common parents" and "the number of common offspring" to create two different partitions of the nodes. In this sense, DI-SIM means two similarities and two partitions. Second, DI- denotes that this algorithm is specifically for *di*rected graphs. Finally, DI-SIM, pronounced "dice 'em", dices data into clusters.

3.2 The di-sim Algorithm

Let $G = (V, E)$ denote a graph, where V is a vertex set and E is an edge set. The vertex set $V = \{1, \ldots, n\}$ contains vertices or nodes. These are the actors in the graph. This paper considers unweighted, directed edges. So, the edge set E contains a pair (i, j) if there is an edge, or relationship, from node i to node j: $i \to j$. The graph can be represented as an adjacency matrix $A \in \{0, 1\}^{n \times n}$:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \text{ is in the edge set} \\ 0 & \text{otherwise.} \end{cases}$$

If the adjacency matrix is symmetric, then the graph is undirected. We are interested in exploring the asymmetries in A.

The graph Laplacian is a function of the adjacency matrix. It is fundamental to spectral graph theory and the spectral clustering algorithm (Chung (1997); von Luxburg (2007)). Several previous papers have proposed and or studied various ways of regularizing the graph Laplacian; these regularization steps improve the statistical performance of various spectral algorithms (Page et al. (1999); Andersen et al. (2006); Chaudhuri et al. (2012); Amini et al. (2013); Qin and Rohe (2013); Joseph and Yu (2014)). This paper generalizes the regularization proposed in Chaudhuri et al. (2012) to directed graphs. Define the regularized graph Laplacian $L \in \mathbb{R}^{n \times n}$ for directed graphs with the diagonal matrices $P \in \mathbb{R}^{n \times n}$ and $O \in \mathbb{R}^{n \times n}$, regularization parameter

$\tau \geq 0$, and identity matrix $I \in R^{n \times n}$,

$$\begin{array}{rcl} P_{jj} & = & \sum_k A_{kj} = \sum_k \mathbf{1}\{k \to j\} \text{ and } P^\tau = P + \tau I; \\ O_{ii} & = & \sum_k A_{ik} = \sum_k \mathbf{1}\{i \to k\} \text{ and } O^\tau = O + \tau I; \text{ and} \\ L_{ij} & = & \frac{A_{ij}}{\sqrt{O_{ii}^\tau P_{jj}^\tau}} = \frac{\mathbf{1}\{i \to j\}}{\sqrt{O_{ii}^\tau P_{jj}^\tau}} = [(O^\tau)^{-1/2} A (P^\tau)^{-1/2}]_{ij}. \end{array} \tag{3.1}$$

P_{jj} is the number of nodes that send an edge to node j, or the number of parents to node j. Similarly, O_{ii} is the number of nodes to which i sends an edge, or the number of offspring to node i. A more standard definition of the graph Laplacian is $I - O^{-1/2} A O^{-1/2}$. Our definition also uses P in the normalization and it does not contain $I-$. These changes are essential to our theoretical results and many of the interpretations of DI-SIM would not hold otherwise. The regularized degree matrices, P^τ and O^τ, artificially inflate every degree by a constant τ. In the setting of undirected graphs, Qin and Rohe (2013) showed that in order to make the asymptotic bounds informative, τ should grow proportionally to the average node degree, $\sum_i O_{ii}/n$. Note that $\sum_i O_{ii}/n = \sum_j P_{jj}/n$ since the out degree equals to the in degree. We use the average node degree as the default value for τ.

To apply DI-SIM to a bipartite graph on disjoint sets of vertices U and V (e.g. U contains words and V contains documents), let U index the rows of A and V index the columns of A. As such, A is rectangular and $A_{ij} = 1$ if and only if $i \in U$ shares an edge with $j \in V$ (e.g. word i is contained in document j). While the dimensions of O, P, and L must change to reflect that A is rectangular, the definitions in Equations (6.9) remain the same.

Throughout, for $x \in \mathbb{R}^d$, $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$, for $M \in \mathbb{R}^{d \times p}$, $\|M\|$ denotes the spectral norm and $\|M\|_F$ denotes the Frobenius norm. With the above notation, DI-SIM is defined as follows.

DI-SIM

Input: Adjacency matrix $A \in \{0,1\}^{n \times n}$, regularizer $\tau \geq 0$ (Default: $\tau =$ average node degree), number of row-clusters k_y, number of column-clusters k_z.

(1) Compute the regularized graph Laplacian $L = (O^\tau)^{-1/2} A (P^\tau)^{-1/2}$.

(2) Compute the top K left and right singular vectors $X_L \in \mathbb{R}^{n \times K}, X_R \in \mathbb{R}^{n \times K}$, where $K = \min\{k_y, k_z\}$.

(3) Normalize each row of X_L and X_R to have unit length. That is, define $X_L^* \in \mathbb{R}^{n \times K}, X_R^* \in \mathbb{R}^{n \times K}$, such that

$$[X_L^*]_i = \frac{[X_L]_i}{\|[X_L]_i\|_2}, \; [X_R^*]_j = \frac{[X_R]_j}{\|[X_R]_j\|_2},$$

where $[X_L]_i$ is the ith row of X_L and similarly for $[X_L^*]_i, [X_R]_j, [X_R^*]_j$.

(4) Cluster the rows of X_L^* into k_r clusters with $(1+\alpha)$-approximate k-means (Kumar et al. (2004)). Because each row of X_L^* corresponds to a node's sending pattern in the graph, the results cluster the nodes' sending patterns.

(5) Cluster the receiving patterns by performing step (4) on the matrix X_R^* with k_z clusters.

Output: The clusters from step (4) and (5).

When A is undirected, then the left and right singular vectors of L are equal to each other and equal to the eigenvectors of L. In this special case, DI-SIM is equivalent to previous versions of undirected spectral clustering (e.g. see von Luxburg (2007), Qin and Rohe (2013)).

3.3 Stochastic co-Blockmodel

This section proposes a statistical model for a directed graph with dual notions of stochastic equivalence. Despite the fact that DI-SIM is not a model based algorithm, when the graph is sampled from this model, DI-SIM will estimate these dual partitions.

Stochastic equivalence, a model based similarity

Stochastic equivalence is a fundamental concept in classical social network analysis. In the Stochastic Blockmodel, two nodes are in the same block if and only if they are stochastically equivalent (Holland et al. (1983)). In a directed network, two nodes a and b are stochastically equivalent if and only if both of the following hold:

$$P(a \to x) = P(b \to x) \ \ \forall x \ \text{ and} \tag{3.2}$$
$$P(x \to a) = P(x \to b) \ \ \forall x \tag{3.3}$$

where $a \to x$ denotes the event that a sends an edge to x. Separating these two notions allows for co-clustering structure. Two nodes a and b are *stochastically equivalent senders* if and only if Equation 3.2 holds. Two nodes a and b are *stochastically equivalent receivers* if and only if Equation 3.3 holds. These two concepts correspond to a model based notion of co-clusters and they are simultaneously represented in the new Stochastic co-Blockmodel.

A statistical model of co-clustering in directed graphs

The Stochastic Blockmodel provides a model for a random network with K well defined blocks, or communities (Holland et al. (1983)). The Stochastic co-Blockmodel is an extension of the Stochastic Blockmodel.

This model naturally generalizes to bi-partite graphs, where the rows and the columns of A index different sets of actors (e.g. words and documents). As such, the rest of the paper allows for a different number of rows (N_r) and columns (N_c) in the adjacency matrix A. Using the notation from the previous sections, a directed graph would satisfy $N_r = N_c = n$.

Definition 3.1. *Define three nonrandom matrices, $Y \in \{0,1\}^{N_r \times k_y}$, $Z \in \{0,1\}^{N_c \times k_z}$ and $B \in [0,1]^{k_y \times k_z}$. Each row of Y and each row of Z has exactly one 1 and each column has at least one 1. Under the* ***Stochastic co-Blockmodel*** *(ScBM), the adjacency matrix $A \in \{0,1\}^{N_r \times N_c}$ is random such that $\mathrm{E}(A) = YBZ^T$. Further, each edge is independent, so the probability distribution factors*

$$P(A) = \prod_{i,j} P(A_{ij}).$$

Without loss of generality, we will always presume that $k_y \leq k_z$.

In the Stochastic Blockmodel, $\mathrm{E}(A) = ZBZ^T$. In the ScBM, $\mathrm{E}(A) = YBZ^T$. In this definition, Y and Z record two types of block membership which correspond to

the two types of stochastic equivalence (Equations 3.2 and 3.3). Denote y_i as the ith row of Y and z_i to be the ith row of Z.

Proposition 3.2. *Under the ScBM for a directed graph, if $y_i = y_j$, then nodes i and j are stochastically equivalent senders, Equation 3.2. Similarly, if $z_i = z_j$, then nodes i and j are stochastically equivalent receivers, Equation 3.3.*

Wang and Wong (1987) previously proposed and studied a directed Stochastic Blockmodel. However, our aims are different. Where Wang and Wong (1987) sought to understand the dependence between A_{ij} and A_{ji}, the current paper seeks to understand the co-clustering structure of the blocks. Importantly, where we use two types of stochastic equivalence (sending and receiving), Wang and Wong (1987) uses only one type of stochastic equivalence which implies that if two nodes are stochastically equivalent senders, then the nodes are also stochastically equivalent receivers and vice versa. By encoding co-clustering structure, the ScBM more closely aligns with the concept of separately exchangeable arrays (e.g. see Diaconis and Janson (2007) and Wolfe and Choi (2014)).

Degree Correction for co-Blockmodel

The degree-corrected Stochastic Blockmodel generalizes the Stochastic Blockmodel to allow for nodes in the same block to have highly heterogeneous degrees (Karrer and Newman (2011)). Theorem 3.7 below studies a similar generalization of the ScBM. The Degree-Corrected Stochastic co-Blockmodel (DC-ScBM) adds two sets of parameters ($\theta_i^y > 0, i = 1, ..., N_r$ and $\theta_j^z > 0, j = 1, ..., N_c$) that control the in- and out-degrees for each node. Let $\mathbf{B}$ be a $k_y \times k_z$ matrix where $\mathbf{B}_{ab} \geq 0$ for all a, b. Then, under the DC-ScBM

$$P(A_{ij} = 1) = \theta_i^y \theta_j^z \mathbf{B}_{y_i z_j}$$

where $\theta_i^y \theta_j^z \mathbf{B}_{y_i z_j} \in [0, 1]$. Note that parameters θ_i^y and θ_j^z are arbitrary to within a multiplicative constant that is absorbed into $\mathbf{B}$. To make it identifiable, we impose the constraint that within each row block, the summation of θ_i^ys is 1. That is, for each row-block s,

$$\sum_i \theta_i^y \mathbf{1}(Y_{is} = 1) = 1.$$

Similarly, for any column-block t, we impose

$$\sum_j \theta_j^z \mathbf{1}(Z_{jt} = 1) = 1.$$

Under this constraint, $\mathbf{B}$ has explicit meaning: $\mathbf{B}_{st}$ represents the expected number of links from row-block s to column-block t. Under the DC-ScBM, define $\mathscr{A} \triangleq \mathbb{E}A$. This matrix can be expressed as a product of the matrices,

$$\mathscr{A} = \Theta_y Y \mathbf{B} Z^T \Theta_z,$$

where Θ_y is a diagonal matrix whose ii'th element is θ_i^y and Θ_z is defined similarly with θ_j^z.

3.4 Estimating the Degree Corrected Stochastic co-Blockmodel with di-sim

Theorem 3.7 bounds the number of nodes that DI-SIM "misclusters". This demonstrates that the co-clusters from DI-SIM estimate both the row- and column-block memberships, one in matrix Y and the other in matrix Z, corresponding to the two types of stochastic equivalence. This implies that the two notions of stochastic equivalence relate to the two sets of singular vectors of L.

In a diverse set of large empirical networks, the optimal clusters, as judged by a wide variety of graph cut objective functions, are not very large (Leskovec et al. (2008)). To account for this, the results below limit the growth of community sizes by allowing the number of communities to grow with the number of nodes. Previously, Rohe et al. (2011); Choi et al. (2012); Rohe et al. (2014) have also studied this high dimensional setting for the undirected Stochastic Blockmodel.

Several previous papers have explored the use of spectral tools to aid the estimation of the Stochastic Blockmodel, including McSherry (2001); Dasgupta et al. (2004); Coja-Oghlan and Lanka (2009); Ames and Vavasis (2010); Rohe et al. (2011); Sussman et al. (2012a); Chaudhuri et al. (2012); Joseph and Yu (2014); Qin and Rohe (2013); Sarkar and Bickel (2013); Krzakala et al. (2013); Jin (2015); and Lei and Rinaldo (2015). The results below build on this previous literature in several ways. Theorem 3.7 gives the first statistical estimation results for directed graphs or bipartite graphs with general degree distributions. Because we study a graph that is directed, DI-SIM uses the leading singular vectors of a sparse and asymmetric matrix. As such, the proof required novel extensions of previous proof techniques. These techniques allow the results to also hold for bipartite graphs; previous results for bipartite graphs have only studied computationally intractable techniques, e.g. Flynn and Perry (2012); Wolfe and Choi (2014). For directed graphs and particularly for bipartite graphs, it is not necessarily true that the number of sending clusters should equal the number of

receiving clusters. Theorem 3.7 below does not presume that the number of sending clusters equals the number of receiving clusters; the theoretical results highlight the statistical price that is paid when they are not equal. Finally, we study a sparse degree corrected model and the theoretical results highlight the importance of the regularization and projection steps in DI-SIM.

Previous theoretical papers that use the non-regularized graph Laplacian all require that the minimum degree grows with the number of nodes (e.g. Rohe et al. (2011); Sarkar and Bickel (2013); Lei and Rinaldo (2015)). However, in many empirical networks, most nodes have 1, 2, or 3 edges. In these settings, the non-regularized graph Laplacian often has highly localized eigenvectors that are uninformative for estimating large partitions in the graph. Because DI-SIM uses a *regularized* graph Laplacian, the concentration of the singular vectors does not require a growing minimum node degree. Several previous papers have realized the benefits of regularizing the graph Laplacian (e.g. Page et al. (1999); Andersen et al. (2006); Amini et al. (2013); Chaudhuri et al. (2012); Qin and Rohe (2013); Joseph and Yu (2014)). While the regularized singular vectors concentrate without a growing minimum degree, the weakly connected nodes effect the conclusions through their statistical leverage scores. From the perspective of numerical linear algebra, the leverage scores and the localization of the singular vectors are essential to controlling the algorithmic difficulty of computing the singular vectors (Mahoney, 2012a).

Population notation

Recall that $\mathscr{A} = \mathbb{E}(A)$ is the population version of the adjacency matrix A. Under the Degree-Corrected Stochastic co-Blockmodel,

$$\mathscr{A} = \Theta_y Y \mathbf{B} Z^T \Theta_z,$$

Similar to Equation (6.9), define regularized population versions of O, P, and L as

$$\begin{aligned} &\mathscr{O}_{jj} = \textstyle\sum_k \mathscr{A}_{kj} \\ &\mathscr{P}_{ii} = \textstyle\sum_k \mathscr{A}_{ik} \\ &\mathscr{O}_\tau = \mathscr{O} + \tau I, \qquad \mathscr{P}_\tau = \mathscr{P} + \tau I \\ &\mathscr{L} = \mathscr{O}_\tau^{-\frac{1}{2}} \mathscr{A} \mathscr{P}_\tau^{-\frac{1}{2}} \end{aligned} \tag{3.4}$$

where $\mathscr{O}$ and $\mathscr{P}$ are diagonal matrices.

Define $O_B \in \mathbb{R}^{k_y \times k_y}$ as a diagonal matrix whose (s,s)'th element is $[O_B]_{ss} = \sum_t \mathbf{B}_{st}$. Similarly define $P_B \in \mathbb{R}^{k_z \times k_z}$ as a diagonal matrix whose (t,t)'th element is

$[P_B]_{tt} = \sum_s \mathbf{B}_{st}$. A couple lines of algebra shows that $[O_B]_{ss}$ is the total expected out-degrees of row nodes from block s and that $\mathscr{O}_{ii} = \theta_i^Y [O_B]_{y_i y_i}$. Similarly $[P_B]_{tt}$ is the total expected in-degrees of column nodes from block t and that $\mathscr{P}_{jj} = \theta_j^Z [P_B]_{z_j z_j}$. Define $B_L = O_B^{-1/2} \mathbf{B} P_B^{-1/2}$.

The population graph Laplacian $\mathscr{L}$ has an alternative expression in terms of Y and Z.

Lemma 3.3. *(Explicit form for $\mathscr{L}_\tau$) Under the DC-ScBM with parameters $\{\boldsymbol{B}, Y, Z, \Theta_Y, \Theta_Z\}$, define $\Theta_{Y,\tau} \in \mathbb{R}^{N_r \times N_r}$ ($\Theta_{Z,\tau} \in \mathbb{R}^{N_c \times N_c}$) to be diagonal matrix where*

$$[\Theta_{Y,\tau}]_{ii} = \theta_i^Y \frac{\mathscr{O}_{ii}}{\mathscr{O}_{ii} + \tau} \qquad [\Theta_{Z,\tau}]_{jj} = \theta_j^Z \frac{\mathscr{P}_{jj}}{\mathscr{P}_{jj} + \tau}.$$

Then $\mathscr{L}$ has the following form,

$$\mathscr{L} = \mathscr{O}_\tau^{-\frac{1}{2}} \mathscr{A} \mathscr{P}_\tau^{-\frac{1}{2}} = \Theta_{Y,\tau}^{\frac{1}{2}} Y B_L Z^T \Theta_{Z,\tau}^{\frac{1}{2}}.$$

The proof of Lemma 3.3 is in Section B.2, in the appendix.

Definition of misclustered

Rigorous discussions of clustering require careful attention to identifiability. In the ScBM, the *order* of the columns of Y and Z are unidentifiable. This leads to difficulty in defining "misclustered". Theorem 3.7 uses the following definition of misclustered that is extended from Rohe et al. (2011).

By the singular value decomposition, there exist orthonormal matrices $\mathscr{X}_L \in \mathbb{R}^{N_r \times k_y}$ and $\mathscr{X}_R \in \mathbb{R}^{N_c \times k_y}$ and diagonal matrix $\Lambda \in \mathbb{R}^{k_y \times k_y}$ such that

$$\mathscr{L} = \mathscr{X}_L \Lambda \mathscr{X}_R^T.$$

Define $\mathscr{X}_L^*$ and $\mathscr{X}_R^*$ as the row normalized population singular vectors,

$$[\mathscr{X}_L^*]_i = \frac{[\mathscr{X}_L]_i}{||[\mathscr{X}_L]_i||_2}, \quad [\mathscr{X}_R^*]_j = \frac{[\mathscr{X}_R]_j}{||[\mathscr{X}_R]_j||_2}.$$

Unless stated otherwise, we will presume without loss of generality that $k_y \leq k_z$. If $\text{rank}(B) = k_y$, then there exist matrices $\mu^y \in \mathbb{R}^{k_y \times k_y}$ and $\mu^z \in \mathbb{R}^{k_z \times k_y}$ such that $Y\mu^y = \mathscr{X}_L^*$ and $Z\mu^z = \mathscr{X}_R^*$ (implied by Lemma B.4 in the appendix). Moreover, the rows of μ^y are distinct; with a slightly stronger assumption, the rows of μ^z are also distinct. As such, k-means applied to the rows of $\mathscr{X}_L^*$ will reveal the partition in Y.

Similarly for μ^z, $\mathscr{X}_R^*$, and Z. As such, DI-SIM applied to the population Laplacian, $\mathscr{L}$, can discover the block structure in the matrices Y and Z.

Let $X_L \in \mathbb{R}^{N_r \times k_y}$ be a matrix whose orthonormal columns are the right singular vectors corresponding to the largest k_y singular values of L. DI-SIM applies k-means (with k_y clusters) to the rows of X_L^*, denoted as $u_1, \ldots, u_{N_r}$. Each row is assigned to one cluster and each cluster has a centroid.

Definition 3.4. *For $i = 1, \ldots, N_r$, define $c_i^L \in \mathbb{R}^{k_y}$ to be the centroid corresponding to u_i after running $(1+\alpha)$-approximate k-means on $u_1, \ldots, u_{N_r}$ with k_y clusters.*

If c_i^L is closer to some population centroid other than its own, i.e. $y_j \mu^y$ for some $y_j \neq y_i$, then we call node i Y-misclustered. This definition must be slightly complicated by the fact that the coordinates in X_L must first align with the coordinates in $\mathscr{X}_L$. So, the definitions below include an additional rotation matrix $\mathscr{R}_L$.

Definition 3.5. *The set of nodes Y-misclustered is*

$$\mathscr{M}_y = \left\{ i : \|c_i^L - y_i \mu^y \mathscr{R}_L\|_2 > \|c_i^L - y_j \mu^y \mathscr{R}_L\|_2 \ \text{ for any } \ y_j \neq y_i \right\}, \tag{3.5}$$

where $\mathscr{R}_L$ is the orthonormal matrix that solves Wahba's problem $\min \|X_L - \mathscr{X}_L \mathscr{R}_L\|_F$, i.e. it is the procrustean transformation.

Defining Z-misclustered, requires defining c_i^R and μ^z analogous to the previous definitions.

Definition 3.6. *The set of nodes Z-misclustered is*

$$\mathscr{M}_z = \left\{ i : \|c_i^R - z_i \mu^z \mathscr{R}_R\|_2 > \|c_i^R - z_j \mu^z \mathscr{R}_R\|_2 \ \text{ for any } \ z_j \neq z_i \right\}, \tag{3.6}$$

where $\mathscr{R}_R$ is the orthonormal matrix that solves Wahba's problem $\min \|X_R - \mathscr{X}_R \mathscr{R}_R\|_F$, i.e. it is the procrustean transformation.

Asymptotic performance

Define

$$H = (Y^T \Theta_{Y,\tau} Y)^{1/2} B_L (Z^T \Theta_{Z,\tau} Z)^{1/2}.$$

$H \in \mathbb{R}^{k_y \times k_z}$ shares same top K singular values with the population graph Laplacian $\mathscr{L}$. Define $H_{\cdot j}$ as the jth column of H, and define

$$\gamma_z = \min_{i \neq j} \|H_{\cdot i} - H_{\cdot j}\|_2. \tag{3.7}$$

When $k_z > k_y$, γ_z controls the additional difficulty in estimating Z.

Define m_y as the minimum row length of $\mathscr{X}_L$. Similarly define m_z as the minimum row length of $\mathscr{X}_R$. That is,

$$m_y = \min_{i=1,..,N_r} ||[\mathscr{X}_L]_i||_2, \qquad m_z = \min_{j=1,..,N_c} ||[\mathscr{X}_R]_j||_2. \tag{3.8}$$

These are the minimum leverage scores for the matrices $\mathscr{L}\mathscr{L}^T$ and $\mathscr{L}^T\mathscr{L}$.

The next theorem bounds the sizes of the sets of misclustered nodes, $|\mathscr{M}_y|$ and $|\mathscr{M}_z|$.

Theorem 3.7. *Suppose $A \in \mathbb{R}^{N_r \times N_c}$ is an adjacency matrix sampled from the Degree-Corrected Stochastic co-Blockmodel with k_y left blocks and k_x right blocks. Let $K = \min\{k_y, k_z\} = k_y$. Define $\mathscr{L}$ as in Equation 3.4. Define $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ as the K nonzero singular values of $\mathscr{L}$. Let $\mathscr{M}_y$ and $\mathscr{M}_z$ be the sets of Y- and Z-misclustered nodes (Equations 3.5 and 3.6) by DI-SIM. Let δ be the minimum expected row and column degree of A, that is $\delta = \min(\min_i \mathscr{O}_{ii}, \min_j \mathscr{P}_{jj})$. Define γ_z, m_y and m_z as in Equations 3.7 and 3.8. For any $\epsilon > 0$, if $\delta + \tau > 3\ln(N_r + N_c) + 3\ln(4/\epsilon)$, then with probability at least $1 - \epsilon$,*

$$\frac{\mathscr{M}_y}{N_r} \leq c_0(\alpha)\frac{K\ln(4(N_r + N_c)/\epsilon)}{N_r\lambda_K^2 m_y^2(\delta + \tau)}, \tag{3.9}$$

$$\frac{\mathscr{M}_z}{N_c} \leq c_1(\alpha)\frac{K\ln(4(N_r + N_c)/\epsilon)}{N_c\lambda_K^2 m_z^2\gamma_z^2(\delta + \tau)}. \tag{3.10}$$

A proof of Theorem 3.7 is contained in the appendix.

Because $\|\mathscr{X}_L\|_F^2 = K$, the average leverage score $||[\mathscr{X}_L]_i||_2$ is $\sqrt{K/N_r}$. If the m_y is of the same order, with λ_K and K fixed, then $\frac{\mathscr{M}_y}{N_r}$ goes to zero when $\delta + \tau$ grows faster than $\ln(N_r + N_c)$. In sparse graphs, δ is fixed and so τ must grow with n. To ensure that λ_K remains fixed while τ is growing, it is necessary for the *average* degree to also grow.

In many empirical networks, the vast majority of nodes have very small degrees; this is a regime in which δ is not growing. In such networks, the bounds in Equations (3.9) and (3.10) are vacuous unless $\tau > 0$. While these equations are upper bounds, the simulations in the appendix show that for sparse networks (i.e. δ small), these bounds align with the performance of DI-SIM. Moreover, the performance of DI-SIM is drastically improves with statistical regularization.

These results highlight the sensitivity to the smallest leverage scores m_y and m_z. When there are excessively small leverage scores, then the bound above can become

meaningless. However, a slight modification of DI-SIM that excludes the low leveraged points from the k-means step and the clustering results, obtains a vastly improved bound. If one computes the leading singular vectors and only runs k-means on the with the observations i that satisfy $||[\mathscr{X}_L]_i||_2 > \eta\sqrt{K/N}$, then the theoretical results are much improved. Denote the nodes misclustered by this procedure as $\mathscr{M}_y^*$. Let there be N^* nodes with $||[\mathscr{X}_L]_i||_2 > \eta\sqrt{K/N}$. If $N/N^* = O(1)$ and the population eigengap λ_K is not asymptotically diminishing, then

$$\frac{\mathscr{M}_y^*}{N^*} \leq c_2(\alpha)\frac{\ln((N_r + N_c)/\epsilon)}{\eta^2(\delta + \tau)}.$$

The proof mimics the proof of Theorem 3.7.

In Theorem 3.7, the bound for $\mathscr{M}_z$ exceeds the bound for $\mathscr{M}_y$ because the bound for $\mathscr{M}_z$ contains an additional term γ_z. This asymmetry stems from allowing $k_z \geq k_y$. In fact, if $k_y = k_z$, then γ_z can be removed, making the bounds identical. However, if $k_z > k_y$, then Rank($\mathscr{L}$) is at most k_y. So, the singular value decomposition represents the data in k_y dimensions and the k-means steps for both the left and the right clusters are done in k_y dimensions. In estimating Y, there is one dimension in the singular vector representation for each of the k_y blocks. At the same time, the singular value representation shoehorns the k_z blocks in Z into less than k_z dimensions. So, there is less space to separate each of the k_z clusters, obscuring the estimation of Z.

To further understand the bound in Theorem 3.7, define the following toy model.

Definition 3.8. *The **four parameter ScBM** is an ScBM parameterized by $K \in \mathbb{N}, s \in \mathbb{N}, r \in (0,1)$, and $p \in (0,1)$ such that $p + r \leq 1$. The matrices $Y, Z \in \{0,1\}^{n\times K}$ each contain s ones in each column and $B = pI_K + r\mathbf{1}_K\mathbf{1}_K^T$.*

In the four parameter ScBM, there are K left- and right-blocks each with s nodes and the node partitions in Y and Z are not necessarily related. If $y_i = z_j$, then $P(i \to j) = p + r$. Otherwise, $P(i \to j) = r$.

Corollary 3.9. *Assume the four parameter ScBM, with same number of rows and columns, and r, p fixed and K growing with $N = Ks$. Since δ is growing with n, set $\tau = 0$. Then,*

$$\lambda_K = \frac{1}{K(r/p) + 1},$$

where λ_K is the Kth largest singular value of $\mathscr{L}$. Moreover,

$$N^{-1}(|\mathscr{M}_y| + |\mathscr{M}_z|) = O_p\left(\frac{K^2 \log N}{N}\right).$$

The proportion of nodes that are misclustered converges to zero, as long as number of clusters $K = o(\sqrt{N/\log N})$.

The proof of Corollary 3.9 is contained in the Appendix.

3.5 Simulation

The theoretical results of Theorem 3.7 identify (1) the expected node degree and (2) the spectral gap as essential parameters that control the clustering performance of DI-SIM. The simulations investigate DI-SIM's non-asymptotic sensitivity to these quantities under the four parameter Stochastic Co-Blockmodel (Definition 3.8). Moreover, the simulations investigate the performance under the model without degree correction and with degree correction.

Both simulations use $k = 5$ blocks for both Y and Z. Each of the five blocks contains 400 nodes. So, $n = 2000$. When the model is degree corrected, $\theta_1, \ldots, \theta_n$ are iid with $\theta_i \stackrel{d}{=} \sqrt{Z + .169}$ where $Z \sim$ exponential(1). The addition of .169 ensures that $\mathrm{E}(\theta_i) \approx 1$ and thus the expected degrees are unchanged between the degree corrected model and the model without degree correction.

In the first simulation, the expected node degree is represented on the horizontal axis; the out of block probability r and the in block probability $p + r$ change in a way that keeps the spectral gap of $\mathscr{L}$ fixed across the horizontal axis. In the second simulation, the spectral gap is represented on the horizontal axis; the probabilities p and r change so that the expected degree $pk + rn$ remains fixed at twenty. In both simulations, the partition matrices Y and Z are sampled independently and uniformly over the set of matrices with $s = 400$ and $k = 5$.

To design the parameter settings of p and r, note that the population graph Laplacian $\mathscr{L}$ is a rank k matrix. So, its $k + 1$ eigenvalue is $\lambda_{k+1} = 0$ and the spectral gap is $\lambda_k - \lambda_{k+1} = \lambda_k$. Corollary 3.9 says that the kth eigenvalue of $\mathscr{L}$ for $\tau = 0$ is

$$\lambda_k = \frac{1}{k(r/p) + 1}.$$

To keep the spectral gap λ_k fixed, it is equivalent to keeping r/p fixed.

We use the k-means++ algorithm (Kumar et al. (2004), Borchers (2012)) with ten initializations. Only the results for Y-misclustered (Definition 3.5) are reported.

Simulation 1

This simulation investigates the sensitivity of DI-SIM to a diminishing number of edges. Figure 3.1 displays the simulation results for a sequence of nine equally spaced values of the expected degree between 5 and 16. To decrease the variability of the plot, each simulation was run twenty times; only the average is displayed. The solid line corresponds to setting the regularization parameter equal to zero ($\tau = 0$). The line with longer dashes represents $\tau = 1$. The line with small dashes represents the average degree, $\tau = \frac{1}{n}\sum_i P_{ii}$.

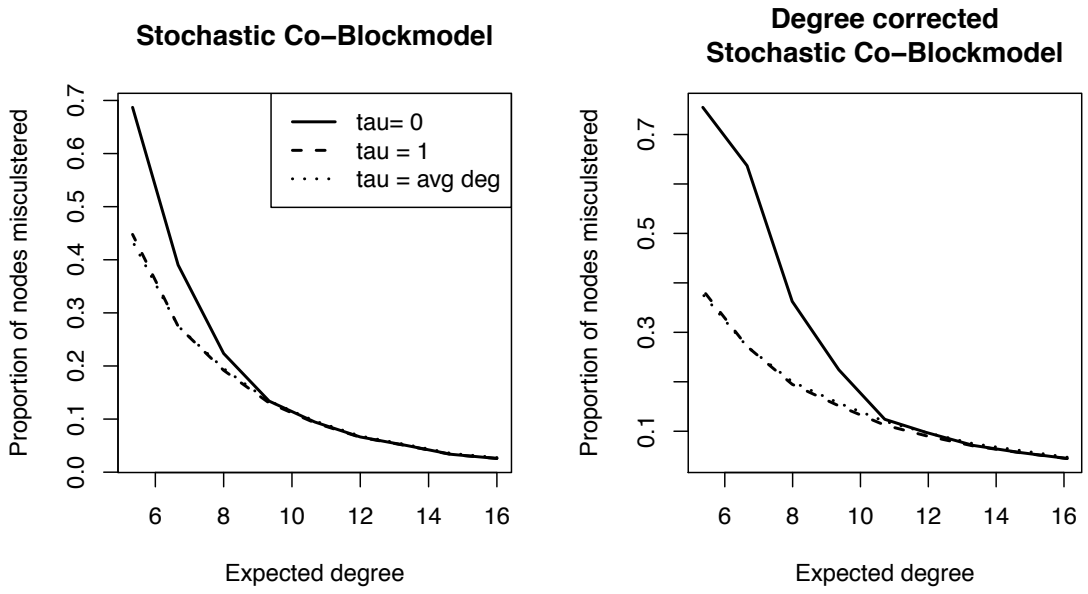

Figure 3.1: In the simulation on the left, the data comes from the four parameter Stochastic Co-Blockmodel. On the right, the data comes from the same model, but with degree correction. The θ_i parameters have expectation one. In both models, $k = 5$ and $s = 400$. The probabilities p and r vary such that $p = 5r$, keeping the spectral gap fixed at $\lambda_k = 1/2$. This simulation shows that for small expected degree, regularization decreases the proportion of nodes that are misclustered. Moreover, the benefits of regularization are more pronounced under the degree corrected model.

Figure 3.1 demonstrates two things. First, the number of misclustered nodes increases as the expected degree goes to zero. Second, regularization decreases the

number of misclustered nodes for small values of the expected degree.

Simulation 2

This simulation investigates the sensitivity of DI-SIM to a diminishing spectral gap λ_k. Figure 3.1 displays the simulation results for a sequence of nine equally spaced values of the spectral gap, between .3 and .6. In each simulation, the expected degree is held constant at twenty. To decrease the variability, each simulation was run twenty times; only the average is displayed. The solid line corresponds to setting the regularization parameter equal to zero ($\tau = 0$). The line with longer dashes represents $\tau = 1$. The line with small dashes represents the average degree, $\tau = \frac{1}{n}\sum_i P_{ii}$.

Figure 3.2 demonstrates two things. First, the number of misclustered nodes increases as the spectral gap goes to zero. Second, regularization yields slight benefits when the spectral gap is small and the model is degree corrected.

3.6 Discussion

Related SVD methods

Several other researchers have used SVD to explore and understand different network features.

Kleinberg (1999) proposed the concept of "hubs and authorities" for hyperlink-induced topic search (HITS). This algorithm that was a precursor to Google's PageRank algorithm (Page et al. (1999)). The SVD plays a key role in this algorithm. The SVD also played a key role in Hoff (2009), where the left and right singular vectors estimate "sender-specific and receiver-specific latent nodal attributes". Like DI-SIM, the algorithms in Kleinberg (1999) and Hoff (2009) use the SVD to investigate asymmetric features of directed graphs.

Dhillon (2001) suggested an algorithm similar to DI-SIM that was to be applied to bipartite graphs in which the rows and columns of L correspond to different entities (e.g. documents and words). There are three key differences between DI-SIM and the algorithm in Dhillon (2001). First, Dhillon (2001) does not use regularization. So, the definition of L remains the same, but $\tau = 0$. The regularization step helps DI-SIM when L has highly localized singular vectors; this often happens when several nodes have very small degrees. Second, Dhillon (2001) does not project the rows of the singular vectors onto the sphere. The project step helps DI-SIM when the node degrees are highly heterogeneous. Finally, to estimate K clusters, Dhillon (2001) only uses $\lceil \log_2 K \rceil$ singular vectors ($\lceil x \rceil$ is the smallest integer greater than x).

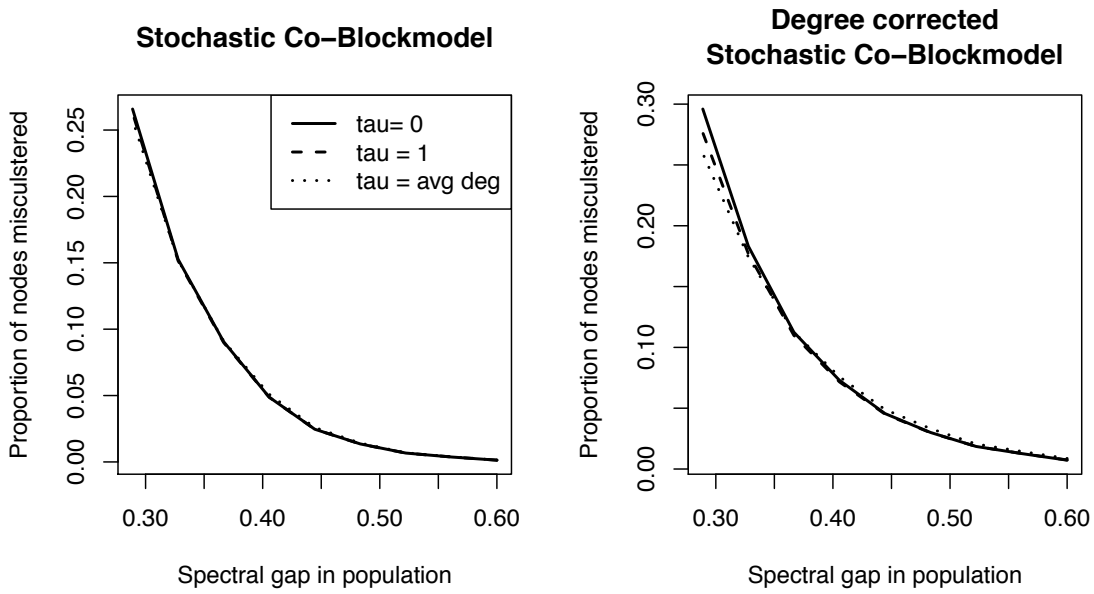

Figure 3.2: In the simulation on the left, the data comes from the four parameter Stochastic Co-Blockmodel. On the right, the data comes from the same model, but with degree correction. The θ_i parameters have expectation one. In both models, $k = 5$ and $s = 400$. The spectral gap, displayed on the horizontal axis, changes because the probabilities p and r change. The values of p and r vary in a way that keeps the expected degree fixed at twenty for all simulations. Without degree correction, the three separate lines are difficult to distinguish because they are nearly identical. Under the degree corrected model, regularization improves performance when the spectral gap is small.

While it is much faster to only compute $\log_2 K$ singular vectors, there is additional information contained in the remaining top K singular vectors. For example, under the four parameter ScBM, $\lambda_2 = \cdots = \lambda_K$. As such, there is not an eigengap after the $\lceil \log_2 K \rceil$th singular value.

SVD has been used in other forms of discrete data, most notably in correspondence analysis (CA). In fact, DI-SIM normalizes the rows and columns in an identical fashion to CA. CA has similarities to principal components analysis, but it is applicable to categorical data in contingency tables and is built on a beautiful set of algebraic ideas (Holmes (2006)). The methodology was first published in Hirschfeld (1935) and

(like spectral clustering) it has been rediscovered and reapplied several times over (Guttman (1959)). While there exists a deep algorithmic, algebraic, and heuristic understanding of CA, it is rarely conceived through a statistical model; Goodman (1986) is one exception. Wasserman et al. (1990) study how one could use CA to study relational data, but was particularly interested in two-way or bipartite networks. Anderson et al. (1992) mentions CA and visual inspection as one possible way to construct blocks in a Stochastic Blockmodel. The previous CA literature has not explored the parameter estimation performance of CA under any of these models, nor has the literature explored the dual partitions under a directed graph. Algorithmically, the CA literature does not employ the regularization step (using τ) for sparse data. Nor does it employ the projection step, where the rows of the singular vector matrices are normalized to have unit length. This is a potentially fruitful area for further research in CA.

In research that was contemporaneous to this paper's tech report (Rohe and Yu (2012)), both Wolfe and Choi (2014) and Flynn and Perry (2012) studied likelihood formulations of co-clustering in the network setting. Wolfe and Choi (2014) studied a "non-parametric" model that assumes the nodes are separately exchangeable. This is a generalization of the Stochastic co-Blockmodel. Flynn and Perry (2012) uses a profile likelihood formulation to develop a consistent estimator of the Stochastic co-Blockmodel.

Conclusion

By extending both spectral clustering and the Stochastic Blockmodel to a co-clustering framework, this paper aims to better conceptualize clustering in directed graphs; co-clustering is a meaningful procedure for directed networks and helps to guide the development of reasonable questions for network researchers.

Given that empirical graphs can be sparse, with highly heterogeneous node degrees, we propose a novel spectral algorithm DI-SIM that incorporates both the regularization and projection steps. Investigating the statistical properties of DI-SIM required several theoretical novelties that build on the extensive literature for spectral algorithms. The results highlight the importance of regularization and the statistical leverage scores. Importantly, because of the regularization, the convergence of the singular vectors does not require a growing minimum degree. Moreover, because the theory accommodates a "degree corrected" model, it was necessary to project the rows of X_L and X_R onto the sphere. Finally, these results extend to bipartite graphs, where the rows and columns of the adjacency matrix index different sets of objects.

Chapter 4

The Highest Dimensional Stochastic Blockmodel with a Regularized Estimator

4.1 Introduction

In the high dimensional Stochastic Blockmodel for a random network, the number of clusters (or blocks) K grows with the number of nodes N. Two previous studies have examined the statistical estimation performance of spectral clustering and the maximum likelihood estimator under the high dimensional model; neither of these results allow K to grow faster than $N^{1/2}$. We study a model where, ignoring log terms, K can grow proportionally to N. Since the number of clusters must be smaller than the number of nodes, no reasonable model allows K to grow faster; thus, our asymptotic results are the "highest" dimensional. To push the asymptotic setting to this extreme, we make additional assumptions that are motivated by empirical observations in physical anthropology (Dunbar, 1992), and an in depth study of massive empirical networks (Leskovec, Lang, Dasgupta, and Mahoney, 2008). Furthermore, we develop a regularized maximum likelihood estimator that leverages these insights and we prove that, under certain conditions, the proportion of nodes that the regularized estimator misclusters converges to zero. This is the first paper to explicitly introduce and demonstrate the advantages of statistical regularization in a parametric form for network analysis.

The Stochastic Blockmodel is a model for a random network. The "blocks" in the model correspond to the concept of "true communities" that we want to study. In the Stochastic Blockmodel, N actors (or nodes) each belong to one of K blocks and

the probability of a connection between two nodes depends only on the memberships of the two nodes (Holland and Leinhardt, 1983). This paper adds to the rigorous understanding of the maximum likelihood estimator (MLE) under the Stochastic Blockmodel.

There has been significant interest in how various clustering algorithms perform under the Stochastic Blockmodel (for example, Bickel and Chen (2009); Rohe, Chatterjee, and Yu (2011); Choi, Wolfe, and Airoldi (2012); Bickel, Chen, and Levina (2011); Zhao, Levina, and Zhu (2011a); Celisse, Daudin, and Pierre (2011); Channarond, Daudin, and Robin (2011); Flynn and Perry (2012); Bickel, Choi, Chang, and Zhang (2012); Sussman, Tang, Fishkind, and Priebe (2012b)). In a parallel line of research, several authors have studied clustering algorithms on the Planted Partition Model, a model nearly identical to the Stochastic Blockmodel. For example, McSherry (2001) studies a spectral algorithm to recover the planted partition and analyzes the estimation performance of this algorithm. Chaudhuri, Chung, and Tsiatas (2012) improves upon this algorithm by introducing a type of regularization and proving consistency results under the planted partition model.

In the previous literature, two papers have studied the high dimensional Stochastic Blockmodel, where the number of blocks K grows with the number of nodes N (Rohe, Chatterjee, and Yu, 2011; Choi, Wolfe, and Airoldi, 2012). The impetus for a high dimensional model comes from two empirical observations. First, Leskovec, Lang, Dasgupta, and Mahoney (2008) found that in a large corpus of empirical networks, the tightest clusters (as judged by several popular clustering criteria) were no larger than 100 nodes, even though some of the networks had several million nodes. This result echoes similar findings in Physical Anthropology. Dunbar (1992) took various measurements of brain size in 38 different primates and found that the size of the neocortex divided by the size of the rest of the brain had a log-linear relationship with the size of the primate's natural communities. In humans, the neocortex is roughly four times larger than the rest of the brain. Extrapolating the log-linear relationship estimated from the 38 other primates, Dunbar (1992) suggests that humans do not have the social intellect to maintain stable communities larger than roughly 150 people (colloquially referred to as Dunbar's number). Leskovec et al. (2008) found a similar result in several other networks that were not composed of humans. The research of Leskovec et al. (2008) and Dunbar (1992) suggests that the block sizes in the Stochastic Blockmodel should not grow asymptotically. Rather, block sizes should remain fixed (or grow very slowly).

In the previous research of Rohe, Chatterjee, and Yu (2011) and Choi, Wolfe, and Airoldi (2012), the average block size grows at least as fast as $N^{3/4}$ and $N^{1/2}$ respectively. Even though these asymptotic results allow for K to grow with N,

K does not grow fast enough. The average block size quickly surpasses Dunbar's number. In this paper, we introduce the highest dimensional asymptotic setting that allows $K = N \log^{-5} N$ and $N/K = \log^5 N$. Thus, under this asymptotic setting, the size of the clusters grows much more slowly. We call it the "highest" dimensional because, ignoring the log term, K cannot grow any faster. If it did, then eventually $K > N$ and there would necessarily be blocks containing zero nodes. To create a sparse graph, the out-of-block probabilities decay roughly as $\log^{\gamma} N/N$ in the highest dimensional setting, where $\gamma > 0$ is some constant. To ensure that a block's induced subgraph remains connected, the in-block probabilities are only allowed to decay slowly like $\log^{-1} N$. We show that under this asymptotic setting, a regularized maximum likelihood estimator (RMLE) can estimate the block partition for most nodes.

This paper departs from the previous high dimensional estimators of Rohe, Chatterjee, and Yu (2011) and Choi, Wolfe, and Airoldi (2012) by introducing a restricted parameter space for the Stochastic Blockmodel. In several high dimensional settings, regularization restricts the full parameter space providing a path to consistent estimators (Negahban, Ravikumar, Wainwright, and Yu, 2010). If the true parameter setting is close to the restricted parameter space, then regularization trades a small amount of bias for a potentially large reduction in variance. For example, in the high dimensional regression literature, sparse regression techniques such as the LASSO restrict the parameter space to produce sparse regression estimators (Tibshirani, 1996). Several authors have also suggested parameter space restrictions for high dimensional covariance estimation, e.g. Fan, Fan, and Lv (2008); Friedman, Hastie, and Tibshirani (2008); Ravikumar, Wainwright, Raskutti, and Yu (2011). Parameter space restrictions have also been applied in Linear Discriminant Analysis (Tibshirani, Hastie, Narasimhan, and Chu, 2002). In graph inference, previous authors have explored various ways of incorporating statistical regularization into eigenvector computations (Chaudhuri, Chung, and Tsiatas, 2012; Amini, Chen, Bickel, and Levina, 2012; Mahoney and Orecchia, 2010; Perry and Mahoney, 2011; Mahoney, 2012b).

In this paper, we propose restricting the parameter space for the Stochastic Blockmodel. These restrictions are supported by empirical observations (Dunbar, 1992; Leskovec, Lang, Dasgupta, and Mahoney, 2008), and they result in a statistically regularized estimator. We will show that the RMLE is suitable in the highest dimensional asymptotic setting. This work is distinct from previous approaches to regularization in graph inference because we study a parametric method, the MLE.

4.2 Preliminaries

Highest Dimensional Asymptotic Setting

In the Stochastic Blockmodel (SBM), each node belongs to one of K blocks. Each edge corresponds to an independent Bernoulli random variable where the probability of an edge between any two nodes depends only on the two nodes' block memberships (Holland and Leinhardt, 1983). The formal definition is as follows.

Definition 4.1. *For a node set $\{1, 2, ..., N\}$, let P_{ij} denote the probability of including an edge linking node i and j. Let $\tilde{z} : \{1, 2, ..., N\} \to \{1, 2, ..., K\}$ partition the N nodes into K blocks. So, $\tilde{z}_i$ equals the block membership for node i. $\tilde{z}$ specifies all true clusters in the model. Let $\boldsymbol{\theta}$ be a $K \times K$ matrix where $\theta_{ab} \in [0, 1]$ for all a, b. Then $P_{ij} = \theta_{\tilde{z}_i \tilde{z}_j}$ for any $i, j = 1, 2, ..., n$. So under the SBM, the the probability of observing adjacency matrix A is*

$$P(A) = \prod_{i<j} \theta_{\tilde{z}_i \tilde{z}_j}{}^{A_{ij}} (1 - \theta_{\tilde{z}_i \tilde{z}_j})^{(1-A_{ij})}.$$

The distribution factors over $i < j$ because we only consider undirected graphs without self-loops.

The highest dimensional asymptotic setting, defined in Definition 4.2, restricts the parameters of the SBM in two ways. First, because empirical evidence suggests that community sizes do not grow with the size of the network, this setting allows s, defined to be the population of the smallest block, to grow very slowly. The second restriction ensures that the sampled networks will have sparse edges. At a high level, there are two types of edges, "in-block edges" that connect nodes in the same block and "out-of-block edges" that connect nodes in different blocks. In order to ensure sparse edges in the high dimensional setting, it is necessary that both the number of out-of-block edges and the number of in-block edges do not grow too fast. To control the number of out-of-block edges, the off-diagonal elements of θ must be (roughly) on the order of $1/N$, otherwise the graph will be dense. The definition allows a set Q to prevent this restriction from becoming too stringent; if $(a, b) \in Q$, then θ_{ab} is not required to shrink as the network grows, allowing blocks a and b to have a tight connection. As for the in-block edges, the slowly growing communities prevent these from creating a dense network; the number of in-block edges connected to each node is bounded by the size of the block population. As such, the highest dimensional asymptotic setting allows the probability of an in-block connection to remain fixed or

decay slowly. It is necessary to prevent these probabilities from converging to zero too quickly because in such small blocks, it would quickly erase any community structure.

Definition 4.2. *The highest dimensional asymptotic setting is an SBM with the following asymptotic restrictions.*

(R1) For s equal to the population of the smallest block and $x_n = \omega(y_n) \Leftrightarrow y_n/x_n = o(1)$,

$$s = \omega(\log^{\beta} N),\ \beta > 4.$$

(R2) Let (c, d) be the interval between c and d and let Q contain a subset of the indices for $\boldsymbol{\theta}$. For constants C and $f(N) = o(s/\log N)$,

$$\theta_{ab} = \theta_{ba} \in \begin{cases} (\log^{-1} N, 1 - \log^{-1} N) & a = b \\ (1/N^2, Cf(N)/N) & a < b, \{a, b\} \notin Q \\ (\log^{-1} N, 1 - \log^{-1} N) & a < b, \{a, b\} \in Q. \end{cases}$$

Assumption (R1) requires that the population of the smallest block $s = \omega(\log^{\beta} N), \beta > 4$. This includes the scenario where each block size is very small (e.g. $o(\log^5 N)$). In this case, the expected degree for each node is $o(\log^5 N)$. In the next sections we will introduce the RMLE and then show that it can identify the blocks under the highest dimensional asymptotic setting.

Regularized Maximum Likelihood Estimator

Under the highest dimensional asymptotic setting, the number of parameters in $\boldsymbol{\theta}$ is quadratic in K and the sample size available for estimating each parameter in $\boldsymbol{\theta}$ is as small as s^2. For tractable estimation in the "large K small s" setting, we propose an RMLE.

Recall that $\tilde{z}$ denotes the true partition. Let z denote any arbitrary partition. The log-likelihood for an observed adjacency matrix A under the SBM w.r.t node partition z is

$$L(A; z, \boldsymbol{\theta}) = \log P(A; z, \boldsymbol{\theta}) = \sum_{i<j} \{A_{ij} \log \theta_{z_i z_j} + (1 - A_{ij}) \log(1 - \theta_{z_i z_j})\}.$$

For fixed class assignment z, let N_a denote the number of nodes assigned to class a, and let n_{ab} denote the maximum number of possible edges between class a and b; i.e.,

$n_{ab} = N_a N_b$ if $a \neq b$ and $n_{aa} = \binom{N_a}{2}$. For an arbitrary partition z, the MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{(\boldsymbol{z})} = \arg \max_{\boldsymbol{\theta} \in [0,1]^{K \times K}} L(A; z, \boldsymbol{\theta}).$$

This is a symmetric matrix in the parameter space $\Theta = [0,1]^{K \times K}$. It is straightforward to show

$$\hat{\theta}_{ab}^{(z)} = \frac{1}{n_{ab}} \sum_{i<j} A_{ij} 1\{z_i = a, z_j = b\}, \quad \forall a, b = 1, 2, ..., K$$

By substituting $\hat{\boldsymbol{\theta}}^{(\boldsymbol{z})}$ into $L(A; z, \boldsymbol{\theta})$, we get the profiled log-likelihood (Bickel and Chen (2009)). Define

$$L(A; z) = L(A; z, \hat{\boldsymbol{\theta}}^{(\boldsymbol{z})}).$$

Define $\hat{z} = \arg\max_z L(A; z)$ as the MLE of $\tilde{z}$. To define the RMLE, define the restricted parameter space, $\Theta^R \subset \Theta$, by the following regularization:

$$\Theta^R = \left\{ \boldsymbol{\theta} \in [0,1]^{K \times K} : \theta_{ab} = c, \ \forall \ a \neq b \ \text{ and for } c \in [0,1] \right\}.$$

If $\theta \in \Theta^R$, then all off-diagonal elements of θ are equal. We call the new estimator "regularized" because, where Θ has $K(K+1)/2$ free parameters, Θ^R has only $K+1$ free parameters.

Given class assignment z, The RMLE $\boldsymbol{\theta}^{R,(z)}$ is the maximizer of $L(A; z, \boldsymbol{\theta})$ within Θ^R.

$$\boldsymbol{\theta}^{R,(z)} = \arg \max_{\boldsymbol{\theta} \in \Theta^R} L(A; z, \boldsymbol{\theta}).$$

The optimization problem within Θ^R can be treated as an unconstrained optimization problem within $[0,1]^{K+1}$ since we force the off-diagonal elements of $\boldsymbol{\theta}$ to be equal to some number r. It has a closed form solution:

$$\hat{\theta}_{ab}^{R,(z)} = \begin{cases} \hat{\theta}_{aa}^{(z)} = \frac{1}{n_{aa}} \sum_{i<j} A_{ij} 1\{z_i = a, z_j = b\} & a = b, \\ \hat{r}^{(z)} = \frac{1}{n_{out}} \sum_{i<j} A_{ij} 1\{z_i \neq z_j\} & a \neq b. \end{cases}$$

Here $n_{out} = \sum_{a<b} n_{ab}$ is the maximum number of possible edges between all different blocks. The Regularized MLE for θ_{aa} is exactly the same as ordinary MLE, while the Regularized MLE for $\theta_{ab}, a \neq b$ is set to be equal to the total off-diagonal average. Finally, by substituting $\hat{\boldsymbol{\theta}}^{R,(z)}$ into $L(A; z, \boldsymbol{\theta})$, define the regularized profile

log-likelihood to be

$$L^R(A; z) = L(A; z, \hat{\boldsymbol{\theta}}^{R,(z)}) = \sup_{\boldsymbol{\theta} \in \Theta^R} L(A; z, \boldsymbol{\theta}),$$

and denote the RMLE of the true partition $\tilde{z}$ to be

$$\hat{z}^R = \arg\max_z L^R(A; z). \tag{4.1}$$

4.3 Performance of the RMLE in the highest dimensional asymptotic setting

Our main result shows that most nodes are correctly clustered by the RMLE under the highest dimensional asymptotic setting. This result requires the definition of "correctly clustered" from Choi, Wolfe, and Airoldi (2012).

Definition 4.3. *For any estimated class assignment z, define $N_e(z)$ as the number of incorrect class assignments under z, counted for every node whose true class under $\tilde{z}$ is not in the majority within its estimated class under z.*

The main result, Theorem 4.4, uses the KL divergence between two Bernoulli distributions. This is defined as

$$D(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

Recall that under the highest dimensional asymptotic setting, Q denotes the off diagonal indices of $\boldsymbol{\theta}$ that do not asymptotically decay. Additionally, n_{ab} denotes the total number of possible edges between nodes in block a and nodes in block b. Define $|Q|$ as the number of possible tight edges across different blocks,

$$|Q| = \sum_{\{a,b\} \in Q} n_{ab}. \tag{4.2}$$

The following theorem is our main result. It shows that under the highest dimensional asymptotic setting, the proportion of nodes that the RMLE misclusters converges to zero.

Theorem 4.4. *Under the highest dimensional asymptotic setting in Definition 4.2, N is the total number of nodes, and s is the population of the smallest block. Assume*

that the set of friendly block pairs Q (defined in R2 of Definition 4.2) is small enough that $|Q| = o(Ns)$, where $|Q|$ is defined in Equation 4.2. Furthermore, for the matrix of probabilities $\boldsymbol{\theta}$, assume that for any distinct class pairs (a, b), there exists a class c such that the following condition holds:

$$D\left(\theta_{ac} \| \frac{\theta_{ac} + \theta_{bc}}{2}\right) + D\left(\theta_{bc} \| \frac{\theta_{ac} + \theta_{bc}}{2}\right) \geq C\frac{MK}{N^2} \tag{4.3}$$

Under these assumptions, RMLE $\hat{z}^R$ defined in Equation 4.1 satisfies

$$\frac{N_e(\hat{z}^R)}{N} = o_p(1),$$

where $N_e(z)$ is the number of misclustered nodes defined in Definition 4.3.

This theorem requires two main assumptions. The first main assumption is $|Q| = o(Ns)$. Define the number of expected edges $M = \sum_{i<j} EA_{ij}$. Under the highest dimensional asymptotic setting, this first assumption implies that M grows slowly, specifically $M = \omega(N(\log N)^{3+\delta})$, where $\delta > 0$. The second main assumption says that every distinct class pair (a, b) has at least one class c that satisfies Equation 4.3. This assumption relates to the identifiability of $\tilde{z}$ under the highest dimensional asymptotic setting. For example, if $(a, b) \notin Q$, then choosing $c = a$ satisfies the assumption in Equation 4.3, because θ_{aa} is large and θ_{ba} is small. However, if $(a, b) \in Q$, then there should exist at least one class c to make θ_{ac},θ_{bc} identifiable. Otherwise, blocks a and b should be merged into the same block. Interestingly, this assumption is not strong enough to ensure that $\tilde{z}$ maximizes $E(L^R(A, \cdot))$, but this is not relevant for our asymptotic results. If one is concerned about this abnormality, it would be enough to assume in R2 (in the definition of the highest dimensional asymptotic setting) that if $\{a, b\} \in Q$, then $\theta_{ab} < \Delta$. This ensures that the probabilities in the set Q are smaller than the in-block probabilities. Such an assumption does not change the asymptotic result.

While theorem 4.4 does not make an explicit assumption about the size of the largest block, Equation 4.3 makes an implicit assumption because the size of the largest blocks affects the number of edges M. Equation 4.3 is satisfied when $MK/N^2 \to 0$ and the set Q does not interfere. For example, if $|Q| = 0$ and the largest block is $O(N^{1/2-\epsilon})$ for some $\epsilon > 0$, then Equation 4.3 is satisfied.

4.4 Simulations

This section compares the RMLE's and the MLE's ability to estimate the block memberships in the Stochastic Blockmodel. In our simulations, the RMLE outperforms the MLE in a wide range of scenarios, particularly when there are several blocks and when the out-of-block probabilities are not too heterogeneous.

Implementation

Computing the exact RMLE and MLE is potentially computationally intractable owing to the combinatorial nature of the parameter space. In this simulation, we fit the MLE with the pseudo-likelihood algorithm proposed in Amini, Chen, Bickel, and Levina (2012). A slight change to the pseudo-likelihood algorithm can fit the RMLE as well; immediately after the pseudo-likelihood algorithm updates $\theta^{(z)}$, we replace the off-diagonal elements with the average of the off-diagonal elements.

This pseudo-likelihood implementation of the RMLE often returns an estimated partition that contained empty sets; for example, if the model was simulated with $K = 30$ blocks and the algorithm was told to estimate 30 blocks, it often estimates fewer than 30 blocks. When the pseudo-likelihood implementation of the RMLE discards a block, the simulations below "reseeds" a new block. This reseeding is done by the following algorithm that was motivated by follow-up work to the current paper (see Rohe and Qin (2013)):

1. Find the block (as defined by the current iteration of the partition) with the smallest empirical in-block probability.

2. For each node in this block, take its neighborhood and remove any nodes that do not connect to any other nodes in the neighborhood. Call this the transitive neighborhood.

3. Combine into a new block both (1) the node with the most nodes in its transitive neighborhood with (2) this node's transitive neighborhood.

We found it beneficial to do this reseeding not only when blocks disappear, but also whenever they are smaller than two nodes.

Section C demonstrates how this reseeding provides consistently better values of the restricted likelihood, that is $L^R(A, \hat{z}_{reseed}) \geq L^R(A, \hat{z}_{no.reseed})$ where $\hat{z}_{reseed}$ is the partition estimated with reseeding and $\hat{z}_{no.reseed}$ is the partition estimated without reseeding. In essence, the reseeding technique is helping the pseudo-likelihood implementation of RMLE attain larger likelihood scores.

Similarly to the suggestion in Amini, Chen, Bickel, and Levina (2012), we initialize the pseudo-likelihood algorithm with spectral clustering using the regularized graph Laplacian (Chaudhuri, Chung, and Tsiatas, 2012). Specifically, it runs k-means on the top K eigenvectors of the matrix $D_\tau^{-1/2} A D_\tau^{-1/2}$, where $D_\tau^{-1/2}$ is a diagonal matrix whose i,ith element is $1/\sqrt{D_{ii}+\tau}$. $D_{ii} = \sum_j A_{ij}$ is the degree of node i and tuning parameter τ is set to be the average degree of all nodes, as was proposed in Qin and Rohe (2013).

Numerical results

This section contains two sets of simulations. In the first set of simulations, K is growing while everything else remains fixed. The second set of simulations investigate the sensitivity of the algorithms to heterogeneous values in the off-diagonal elements of θ.

The results in Figure 4.1 compare the RMLE and MLE under an asymptotic regime that keeps the population of each block fixed at twenty nodes and simply adds blocks. The horizontal axes corresponds to K growing from ten to one hundred. In both the left and the right panel, the probability of a connection between two nodes in the same block is $8/20$. In the left panel, the probability of a connection between two nodes in separate blocks is $5/N$. In the right panel, it is $10/N$. Under these two asymptotics, the expected number of "signal" edges connected to each node is eight, while the expected number of "noisy" edges is either five or ten. The vertical axis in both figures is $N_e(\hat{z})/N$, the proportion of misclustered nodes.

The results in Figure 4.2 examine the sensitivity of the algorithms to deviations from the model in Figure 4.1 that makes the off-diagonal elements of θ equal to one another. In all simulations, the expected number of "signal edges" per node is eight, the expected number of "noisy edges" per node is 5, $s = 20$, and $K = 40$. On the left side of Figure 4.2, the off-diagonal elements of θ come from the Gamma distribution. In the top left figure, the shape parameter in the Gamma distribution (α) varies along the horizontal axis. While the shape parameter varies, the rate parameter changes to ensure that each node has an expected out-of-block degree equal to five.[1] Under our scaling of the rate parameter, the variance of the Gamma distribution is proportional to $1/\alpha$. As such, the small values of α make the out-of-block probabilities more heterogeneous, deviating further from the implicit model. For a point of reference, recall that $\alpha = 1$ gives the exponential distribution. Our simulations present $\alpha \in (.1, .55)$, more variable than the exponential distribution. For values of α greater than .18, the RMLE outperforms the MLE. The bottom left

[1]Since θ is now random, this expectation is taken over both A and θ.

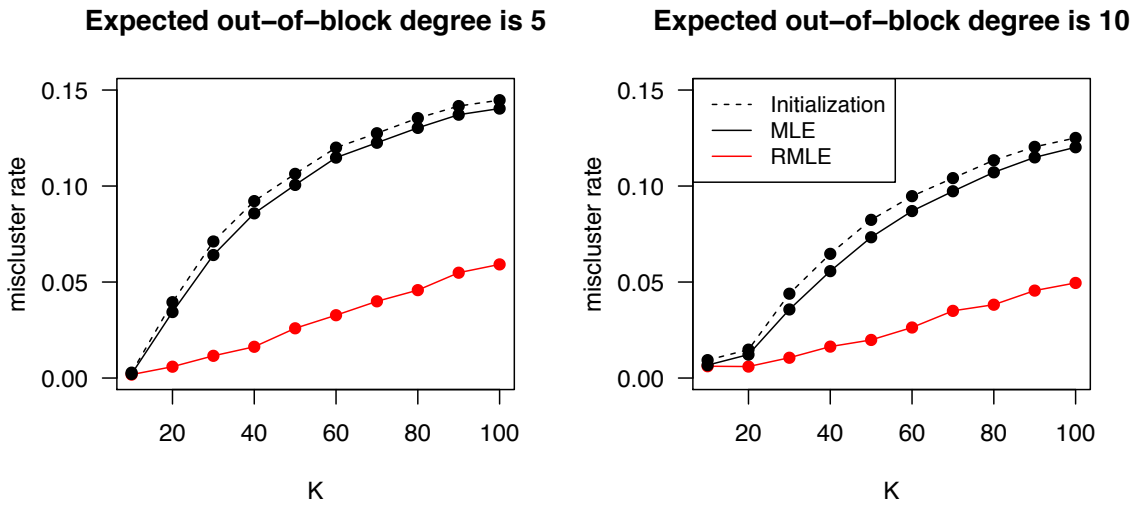

Figure 4.1: In this simulation, across a wide range of K, the RMLE misclusters fewer nodes than the MLE. In each simulation, every block contains 20 nodes and K grows from 10 to 100 along the horizontal axis. The vertical axis displays the proportion of nodes misclustered. Both algorithms are initialized with regularized spectral clustering and the results for this initialization are displayed by the dashed line. The MLE makes minor improvements to the initialization, while the RMLE makes more significant improvements. Each point in this figure represents the average of 300 simulations. All methods were run on the same simulated adjacency matrices.

plot shows the top left 400×400 submatrix of the adjacency matrix for a simulated example when $\alpha = .18$; the block pattern is clearly recognizable at this level of α, suggesting that the RMLE is surprisingly robust to deviations from the implicit model.

The plots on the right side of Figure 4.2 are similar, except the off-diagonal elements of θ are scaled Bernoulli(p) random variables. Note that when $p = 1$, this simulation would be identical to a setting in Figure 4.1. The scaling ensures that the expected out-of-block degree is always five. Here, the break-even point is around $p = .14$ and the bottom right figure shows the top left 400×400 submatrix of the adjacency matrix for a sample when $p = .14$; the block pattern is clearly recognizable for this level of p. In both of these cases, the RMLE appears robust to deviations from the implied model. At the same time, for small levels of p and α, the MLE misclusters fewer nodes than the RMLE.

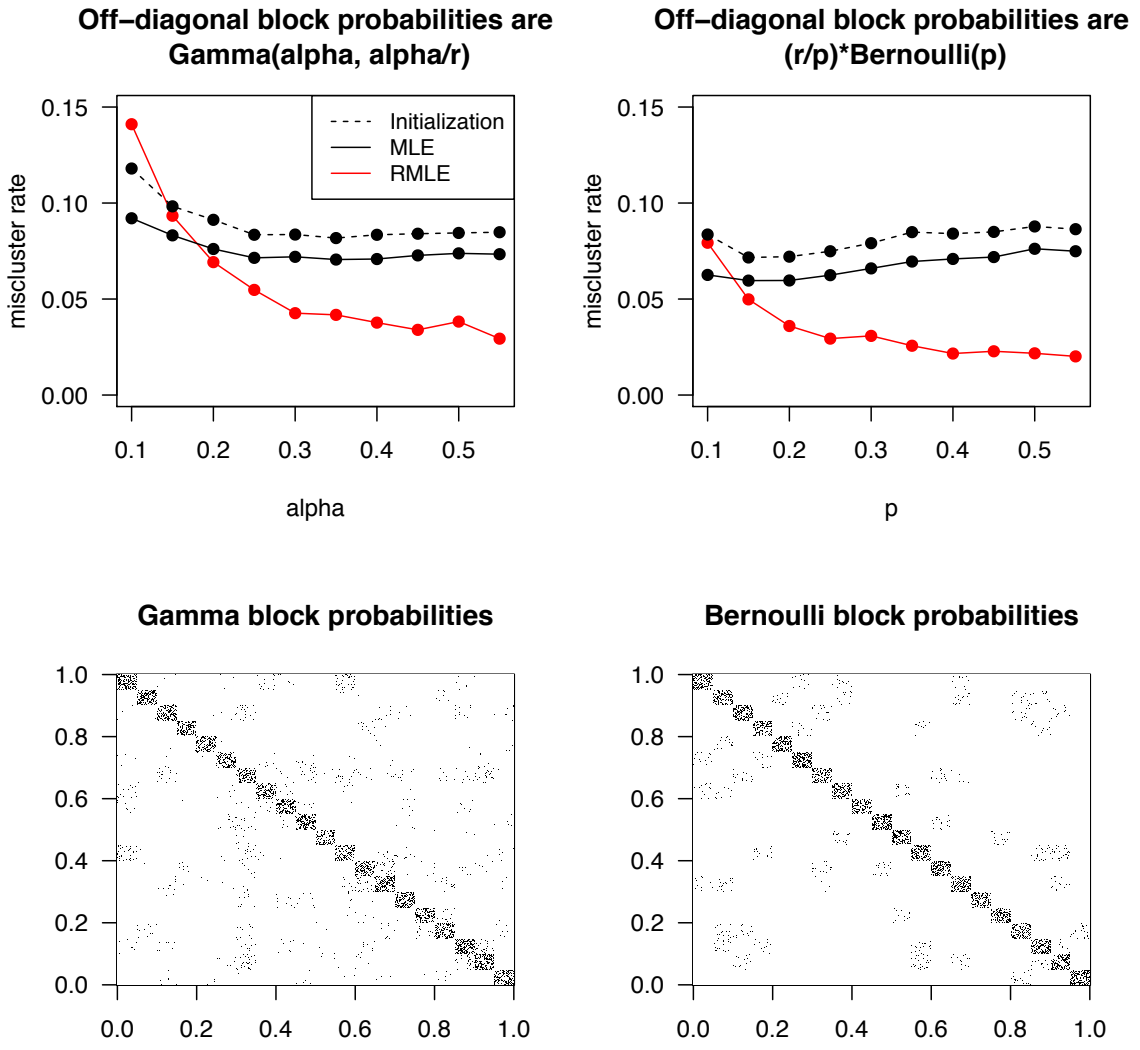

Figure 4.2: These figures investigate the sensitivity of the algorithms to deviations from the RMLE's "implied model" that has homogeneous off-diagonal elements in θ. The top left figure displays results when these elements of θ come from the Gamma distribution with varying shape parameter. The top right figure displays results when these elements of θ come from the Bernoulli distribution with varying probability p. In both cases, adjustments are made so that each node has five expected out-of-block neighbors. The bottom plots illustrate the how these heterogenous probabilities manifest in the adjacency matrix; in both cases, A is sampled with the parameterization that corresponds to the break-even point between the MLE and the RMLE. Each point represents an average over 200 simulations

4.5 Discussion

This paper examines the theoretical properties of the regularized maximum likelihood estimator (RMLE) under the highest dimensional asymptotic setting, showing that under a novel and relevant asymptotic regime, regularization allows for weakly consistent estimation of the block memberships.

Under the highest dimensional asymptotic setting, the size of the communities grows at a poly-logarithmic rate, not at a polynomial rate, aligning with several empirical observations (Dunbar, 1992; Leskovec, Lang, Dasgupta, and Mahoney, 2008). There are two natural implications of the block populations growing this slowly. Under any Stochastic Blockmodel, to ensure the sampled graph has sparse edges, the probability of an out-of-block connection must decay. In previous "low-dimensional" analyses, it was also necessary for the probability of an in-block connection to decay. The first clear implication of small blocks is that the probability of an in-block connection must stay bounded away from zero. Otherwise, a block's induced subgraph will become disconnected. The second implication of small block sizes is that the number of off diagonal elements in Θ grows nearly quadratically with N, while the number of in-block parameters (diagonal elements of Θ) grows linearly with N.

The proposed estimator, restricts the parameter space of the SBM in a way that leverages both of these implications. Since the out-of-block edge probabilities decay to zero, we maximize the likelihood over a parameter space that estimates the probabilities as equal. Theorem 4.4 shows that under the highest dimensional asymptotic setting and certain conditions that are similar to identifiability conditions, the RMLE can estimate the correct block for most nodes. Correspondingly, the simulation section demonstrates the advantages of the RMLE over the MLE. Overall, this paper represents a first step in applying statistically regularized estimators to high dimensional network analysis in a parametric setting. Because of the computational issues involved in computing both the MLE and the RMLE, future work will propose a "local estimator" that (1) incorporates the insights gained from the current analysis and (2) is computationally straight-forward.

Chapter 5

A Normalized and Regularized Nyström Extension for Clustering Network with General Degrees

5.1 Introduction

With the help of fast developing technology, we find ourselves being connected to each other more than ever before. Our lives are embedded in networks: social, biological, communication, etc. Many researchers are trying to analyze these networks to gain a deeper understanding of the underlying mechanisms. Some types of underlying mechanisms generate communities (aka clusters or modularities) in the network. As some networks tend to grow so overwhelmingly massive that it is difficult even to store and analyze them in the first place. This chapter has two aims:

(a) To devise algorithms for community detection that are highly practical – scalable, memory efficient and fast.

(b) To understand if and when we can make justifiable inferences from the estimated communities to the underlying mechanisms.

This chapter proposes a fast and memory efficient algorithm based on standard spectral clustering and a variation of the Nystsröm extension. We then examines its statistical estimation performance under the Degree-Corrected Stochastic Blockmodel (DC-SBM), an extension of the Stochastic Blockmodel (SBM) that allows for heterogeneous degrees (Holland and Leinhardt (1983),Karrer and Newman (2011)).

Spectral Clustering

Spectral clustering is a popular technique for finding communities in networks. Several previous authors have studied the estimation properties of spectral clustering under various statistical network models (McSherry (2001); Dasgupta et al. (2004); Qin and Rohe (2013); Coja-Oghlan and Lanka (2009); Ames and Vavasis (2010); Rohe et al. (2011); Lei and Rinaldo (2015) and Chaudhuri et al. (2012)).

The algorithm is defined in terms of a graph G, represented by a vertex set V and an edge set E. The vertex set $V = \{v_1, v_2, \ldots, v_N\}$ contains vertices or nodes. We will refer to node v_i as node i. E contains a pair (i, j) if there is an edge between node i and j. The edge set can be represented by the adjacency matrix $G \in \{0, 1\}^{n \times n}$. $G_{ij} = G_{ji} = 1$ if (i, j) is in the edge set and $G_{ij} = G_{ji} = 0$ otherwise. Define Graph Laplacian L and diagonal matrix D both elements of $\mathbb{R}^{N \times N}$ in the following way:

$$D_{ii} = \sum_j G_{ij}, \qquad L = D^{-1/2} G D^{-1/2}. \tag{5.1}$$

The spectral clustering algorithm is defined as follows:

Spectral clustering

Input: Adjacency matrix $G \in \{0, 1\}^{N \times N}$, number of clusters K.

1. Get top K eigenvectors of L and form N by K matrix E by putting eigenvectors into its columns.
2. Treat each row of E as a point in $\mathbb{R}^K$, and run k-means with K clusters. This creates K non-overlapping sets $V_1, ..., V_K$ whose union is V. Node i is assigned to cluster r if the i'th row of E is assigned to V_r.

Output: The clusters $V_1, ..., V_K$.

Compared to maximum likelihood based methods and modularity based methods, spectral clustering is computationally more tractable and reasonably fast.

However, when applied to massive networks with degree heterogeneity, spectral clustering falls short in three aspects: (a). Degree heterogeneity significantly influence the stability of spectral clustering, which makes spectral clustering return highly unbalanced or meaningless clusters. (b). For dense networks, the time complexity of spectral clustering is $O(N^2K)$ where N is the number of nodes and K is the number of desired clusters. (c). To compute top eigenvectors, it requires the whole network to be stored into memory first.

To handle networks with heterogeneous degrees, Chaudhuri et al. (2012) and Qin and Rohe (2013) studied regularized versions of spectral clustering that stabilize its performance. Yet both extensions suffer the same time and space bottlenecks as spectral clustering To overcome all three practical disadvantages while maintaining decent clustering performance of spectral clustering, next section will introduce a memory efficient version of regularized spectral clustering.

The Nystsröm Extension

The Nystsröm extension was originally introduced to compute the numerical solution of an integral equation by replacing the integral with a representative weighted sum. Let $K : [a, b] \times [a, b] \to \mathbb{R}$ be an SPSD kernel and $(u_k, \lambda_k), k \in \mathbb{N}$ denote its eigenfunction and eigenvalue pairs:

$$\int_a^b K(x, y)u_k(y)dy = \lambda_k u_k(x), \quad k \in \mathbb{N}.$$

Let $\{(x_i, x_j)\}_{i,j=1}^n \in [a, b] \times [a, b]$ be n^2 distinct points. Define $K_n \in \mathbb{R}^{n\times n}$ where $K_n(i, j) = K(x_i, x_j)$. Let $\{(v_k, \hat{\lambda}_k)\}_{1:n}$ represent the n eigenvector and eigenvalue pairs of K_n/n:

$$\frac{1}{n}\sum_{j=1}^n K_n(i, j)v_k(j) = \hat{\lambda}_k v_k(i), k = 1, 2, ..., n.$$

Then u_k can be approximated by $\hat{u}_k$ as follows:

$$\hat{u}_k(x) = \frac{1}{n\hat{\lambda}_k}\sum_{i=1}^n K(x, x_i)v_k(i).$$

The key idea of the Nystsröm extension is to use partial information about the kernel (n^2 evaluations) to get eigen decomposition of a simpler system K_n and then extend it to approximate the whole set of eigenfunctions of the kernel.

Williams and Seeger (2001) applied similar idea to approximate eigenvectors of SPSD matrices. Let $G \in \mathbb{R}^{N\times N}$ and be partitioned as

$$G = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix},$$

where $A \in \mathbb{R}^{N_1 \times N_1}$. Spectral decomposition of A gives $A \approx X_A \Lambda_A X_A^T$, where X_A contains top K eigenvectors of A. The Nystsröm extension then provides an approximation for K eigenvectors of G:

$$\tilde{X} := \begin{bmatrix} X_A \\ B X_A \Lambda_A^{-1} \end{bmatrix}.$$

Consequently, one can approximate the original matrix G as follows,

$$\tilde{W} := \tilde{X} \Lambda_A \tilde{X}^T = \begin{bmatrix} A & B^T \\ B & B A^{-1} B^T \end{bmatrix}.$$

The Nystsröm extension has various applications in matrix approximation(Belabbas and Wolfe (2009); Drineas and Mahoney (2005)) and image segmentation (Fowlkes et al. (2004)). This paper proposes a memory efficient spectral clustering algorithm for the task of community detection. The proposed algorithm first compute properly normalized top eigenvalue and eigenvector pairs of a sub-adjacency matrix A. Then a regularized version of Nystsröm extension is applied to get extended eigenvectors that contain community information of the whole graph. Lastly k-means is applied to the extended eigenvectors to get clusters of all the nodes. Its estimation performance is then study under the DC-SBM. Different from previous works, this work is the first to study the statistical estimation performance under a parametric framework.

The Degree-Corrected Stochastic Blockmodel

In the Stochastic Blockmodel (SBM), each node belongs to one of K blocks. Each edge corresponds to an independent Bernoulli random variable where the probability of an edge between any two nodes depends only on the block memberships of the two nodes (Holland and Leinhardt (1983)). The formal definition is as follows.

Definition 5.1. *For a node set $\{1, 2, ..., N\}$, let $z : \{1, 2, ..., N\} \to \{1, 2, ..., K\}$ partition the N nodes into K blocks. So, z_i equals the block membership for node i. Let $\boldsymbol{P}$ be a $K \times K$ matrix where $\boldsymbol{P}_{ab} \in [0, 1]$ for all a, b. Then under the SBM, the probability of an edge between i and j is $P_{ij} = P_{ji} = \boldsymbol{P}_{z_i z_j}$ for any $i, j = 1, 2, ..., n$. Given z, all edges are independent.*

One limitation of the SBM is that it presumes all nodes within the same block have the same expected degree. The Degree-Corrected Stochastic Blockmodel (DC-SBM) (Karrer and Newman (2011)) is a generalization of the SBM that adds an additional

set of parameters ($\theta_i > 0$ for each node i) that control the node degrees. Let $\mathbf{P}$ be a $K \times K$ matrix where $\mathbf{P}_{ab} \geq 0$ for all a, b. Then the probability of an edge between node i and node j is $\theta_i\theta_j\mathbf{P}_{z_iz_j}$, where $\theta_i\theta_j\mathbf{P}_{z_iz_j} \in [0, 1]$ for any $i, j = 1, 2, ..., n$. The scale of parameter θ_i is arbitrary to within a multiplicative constant that is absorbed into $\mathbf{P}$. To make it identifiable, Karrer and Newman (2011) suggest imposing the constraint that, within each block, the summation of θ_i's is 1. That is, $\sum_i \theta_i\delta_{z_i,r} = 1$ for any block label r. Under this constraint, $\mathbf{P}$ has explicit meaning: If $s \neq t$, $\mathbf{P}_{st}$ represents the expected number of links between block s and block t and if $s = t$, $\mathbf{P}_{st}$ is twice the expected number of links within block s. Throughout the paper, we assume that $\mathbf{P}$ is positive definite.

Under the DC-SBM, define $\mathscr{G} \triangleq EG$. This matrix can be expressed as a product of the matrices,

$$\mathscr{G} = \Theta Z \mathbf{P} Z^T \Theta,$$

where (1) $\Theta \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose ii'th element is θ_i and (2) $Z \in \{0, 1\}^{N\times K}$ is the membership matrix with $Z_{it} = 1$ if and only if node i belongs to block t (i.e. $z_i = t$).

Preliminaries

For any two subsets $V_1, V_2 \subseteq V$. Let $G(V_1, V_2)$ denote the submatrix of G with row indices restricted to V_1 and column indices restricted to V_2.

Let $V = V_A \cup V_B, V_A \cap V_B = \emptyset$ be a partition of the vertex set. where $|V_A| = N_A$ and $|V_B| = N_B, N_A + N_B = N$. Let $A = G(V_A, V_A) \in \{0, 1\}^{N_A\times N_A}$ and $B = G(V_B, V_A) \in \{0, 1\}^{N_B\times N_A}$. Define the diagonal matrix $D_A \in \mathbb{R}^{N_A\times N_A}, D_B \in \mathbb{R}^{N_B\times N_B}$ and the normalized graph laplacian L_A of A, in the following way:

$$[D_A]_{ii} = \sum_j A_{ij}, \qquad [D_B]_{ii} = \sum_j B_{ij}, \qquad L_A = D_A^{-1/2} A D_A^{-1/2}.$$

Similarly, define the population adjacency submatrices, $\mathscr{A} = \mathscr{G}(V_A, V_A)$ and $\mathscr{B} = \mathscr{G}(V_B, V_A)$. Further define $Z_A \in \{0, 1\}^{N_A\times K}$ and $Z_B \in \{0, 1\}^{N_B\times K}$ to be the community labels for nodes in V_A and V_B respectively.

The following notations will be used throughout the paper: For vector, $\|\cdot\|$ denotes l_2 norm. For matrix, $\|\cdot\|$ denotes the spectral norm, and $\|\cdot\|_F$ denotes the Frobenius norm. For two sequence of variables $\{x_N\}$ and $\{y_N\}$, we say $x_N = \omega(y_N)$ if and only if $y_N/x_N = o(1)$. $\delta_{(.,.)}$ is the indicator function where $\delta_{x,y} = 1$ if $x = y$ and $\delta_{x,y} = 0$ if $x \neq y$.

5.2 Algorithm: Memory efficient regularized spectral clustering(mRSC)

The algorithm includes three steps, The first stage is to compute spectral decomposition of a submatrix A of adjacency matrix G, and the second stage involves "projecting" the rest of the nodes onto the properly stretched eigenspace and get the extended eigenvectors. Lastly, k-means is applied to the extended eigenvectors to get clusters of all nodes. The formal algorithm is described below.

mRSC

Input: Two sub-adjacency matrices $A \in \{0,1\}^{N_A \times N_A}, B \in \{0,1\}^{N_B \times N_A}$, regularizer $\tau \geq 0$ (Default: τ = average row degree in B), number of clusters K.

1. Get top K eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K$ of L_A and its corresponding eigenvectors $\tilde{X}_1, ..., \tilde{X}_K \in \mathbb{R}^{N_A}$. Form $\Lambda = diag(\lambda_1, ..., \lambda_K)$ and $\tilde{X} = [\tilde{X}_1, ..., \tilde{X}_K] \in \mathbb{R}^{N_A \times K}$ by putting the eigenvectors into the columns.

2. Compute the normalized eigenvectors $X \in \mathbb{R}^{N_A \times K}$ where $X = D_A^{-1/2}\tilde{X}$.

3. Compute the extended eigenvectors $Y \in \mathbb{R}^{N_B \times K}$:

$$Y = (D_B + \tau I)^{-1} B X \Lambda^{-1}.$$

4. Form matrix X^* and Y^* by normalizing each row of X and Y to have unit length. Form $F \in \mathbb{R}^{N \times K}$ by combining X^* and Y^*, $F^T = [X^{*T}, Y^{*T}]$.

5. Treat each row of F as a point in $\mathbb{R}^K$, and run k-means with K clusters. This creates K non-overlapping sets $V_1, ..., V_K$ whose union is V. Node i is assigned to cluster r if the i'th row of F is assigned to V_r.

Output: The clusters $V_1, ..., V_K$ from step (6).

Step 2 normalizes the eigenvectors of L_A by a factor of $D_A^{-1/2}$. The normalization will be shown crucial for the algorithm to work under the DC-SBM in the next section. Step 3 is different from the standard Nyström extension, because each row is normalized by its degree plus a regularizing factor τ. This extra step helps the concentration of degrees to their expectations for those low degree nodes and in turn makes our result more general by weakening the assumption on the minimum expected

degree. Step 4 further projects each row onto unit sphere. This was suggested by Ng et al. (2002) and Qin and Rohe (2013).

Related works

Chaudhuri et al. (2012) studies a related algorithm first divides the nodes into two random subsets and then requires two runs of spectral decomposition and projection. In addition, their algorithm need to combine the clustering results of the two subsets of node. Inspired by their work, the proposed mRSC requires only one run of spectral decomposition and k-means on the extended eigenvectors clusters the whole set of nodes.

Fowlkes et al. (2004) proposes a similar spectral algorithm that applies the Nyström extension. Their algorithm computes the orthogonalized eigenvectors and then approximate normalized graph laplacian for the whole graph. Their algorithm is shown to be effective by several data applications.

5.3 Population analysis

This section shows that under the DC-SBM, with the two submatrices $\mathscr{A}$ and $\mathscr{B}$ of population adjacency matrix and a proper regularizer τ, mRSC perfectly reconstructs the block partition.

Define the diagonal matrix $\mathscr{D}_{\mathscr{A}}$ to contain the expected node degrees of A, $[\mathscr{D}_{\mathscr{A}}]_{ii} = \sum_j \mathscr{A}_{ij}$ and define $\mathscr{D}_{\mathscr{B}}$ where $[\mathscr{D}_{\mathscr{B}}]_{ii} = \sum_j \mathscr{B}_{ij}$. Define the population graph Laplacian of $\mathscr{A}$,as $\mathscr{L}_{\mathscr{A}} \in \mathbb{R}^{N_A \times N_A}$, in the following way:

$$\mathscr{L}_{\mathscr{A}} = \mathscr{D}_{\mathscr{A}}^{-1/2} \mathscr{A} \mathscr{D}_{\mathscr{A}}^{-1/2}.$$

Define $\Theta_A = \Theta(V_A, V_A)$ and $\Theta_B = \Theta(V_B, V_B)$.

The next two lemmas give explicit forms of the population version of X and Y.

Lemma 5.2. *(Eigen-decomposition for $\mathscr{L}_{\mathscr{A}}$) Under the DC-SBM with K blocks and parameters $\{P, Z, \Theta\}$, Assume P is positive definite, then $\mathscr{L}_{\mathscr{A}}$ has K positive eigenvalues. The remaining $N_A - K$ eigenvalues are zero. Denote the K positive eigenvalues of $\mathscr{L}_{\mathscr{A}}$ as $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq ... \geq \bar{\lambda}_K > 0$ and let $\tilde{\mathscr{X}} \in \mathbb{R}^{N_A \times K}$ contain the eigenvector corresponding to $\bar{\lambda}_i$ in its i'th column. Define $\mathscr{X} = \mathscr{D}_{\mathscr{A}}^{-1/2} \tilde{\mathscr{X}}$ and define $\mathscr{X}^*$ to be the row-normalized $\mathscr{X}$, similar to X^* as defined in the mRSC algorithm in Section 2. Let $V = diag[(Z_A^T \Theta_A Z_A) P (Z_A^T \Theta_A Z_A) \mathbf{1}]$. Then, there exists an orthogonal matrix $U \in \mathbb{R}^{K \times K}$, such that*

1. $\mathscr{X} = Z_A V^{-1/2} U$,

2. $\mathscr{X}^* = Z_A U$.

Lemma 5.3. *(Structure of $\mathscr{Y}$ and $\mathscr{Y}^*$) Define the population version of Y as*

$$\mathscr{Y} = (\mathscr{D}_{\mathscr{B}} + \tau I)^{-1} \mathscr{B} \mathscr{X} \bar{\Lambda}^{-1},$$

and define $\mathscr{Y}^*$ by normalizing each row of $\mathscr{Y}$ to have unit length. Then

1. $\mathscr{Y} = (\mathscr{D}_{\mathscr{B}} + \tau I)^{-1} \mathscr{D}_{\mathscr{B}} Z_B V^{-1/2} U$,

2. $\mathscr{Y}^* = Z_B U$,

where U is the same orthogonal matrix as in Lemma 5.2.

Define $\mathscr{F} = [\mathscr{X}^{*T}, \mathscr{Y}^{*T}]^T$. Lemma 5.2 and lemma 5.3 reveal two important facts:

First, after projection, population eigenvector $\mathscr{X}^*$ and extended population eigenvector $\mathscr{Y}^*$ share similar simple form. Notice that $U \in \mathbb{R}^{K \times K}$ is orthogonal matrix, rows of U are K distinct points that are orthogonal to each other. Hence, if two nodes i and j belong to the same block ($z_i = z_j$), then no matter which partition they belong to, the corresponding rows of $\mathscr{X}^*$ or $\mathscr{Y}^*$ are equal. If two nodes i and j belong to two different blocks ($z_i \neq z_j$), then their corresponding rows in $\mathscr{X}^*$ or $\mathscr{Y}^*$ are perpendicular. Notice that running k-means on the rows of $\mathscr{F}$ will return perfect clusters.

Second, lemma 5.2 and lemma 5.3 also indicates that unbalanced clusters will make clustering more difficult. For now we set $\tau = 0$. Before projection, population eigenvector $\mathscr{X}^*$ and extended population eigenvector $\mathscr{Y}^*$ both share the same form: $\mathscr{X} = Z_A V^{-1/2} U$ and $\mathscr{Y} = Z_B V^{-1/2} U$. They share K points in the K dimensional space, each representing one underlying cluster. for node i, its corresponding row is $Z_i V^{-1/2} U$. Its distance to origin is $1/\sqrt{V_{z_i, z_i}}$. Notice that by definition of V, V_{ss} is the expected volume of cluster s within A. Hence, if the clusters are highly unbalanced, the cluster will be close the the others.

This section shows that mRSC applied to the population adjacency matrix $\mathscr{A}$ and $\mathscr{B}$ results in perfect community recovery. The next section will show that the perturbation between empirical and population eigenvectors is small. Consequently, we may expect good clustering performance on networks generated from the DC-SBM.

5.4 Perturbation analysis and a bound on mis-clustering rate

The next lemma bounds the distance between L_A and $\mathscr{L}_{\mathscr{A}}$.

Lemma 5.4. *(Concentration of Graph Laplacian) Let δ_A be the minimum expected degree of A, that is $\delta_A = \min_i[\mathscr{D}_{\mathscr{A}}]_{ii}$. For any $\epsilon > 0$, if $\delta_A > 3\ln(4N_A/\epsilon)$, then with probability at least $1-\epsilon$,*

$$\|L_A - \mathscr{L}_{\mathscr{A}}\| \leq 2\sqrt{\frac{3\ln(4N_A/\epsilon)}{\delta_A}}, \tag{5.2}$$

and consequently,

$$|\lambda_i - \bar{\lambda}_i| \leq 2\sqrt{\frac{3\ln(4N_A/\epsilon)}{\delta_A}}, \tag{5.3}$$

for all $1 \leq i \leq N_A$.

The next theorem bounds the difference between the empirical and population (extended) eigenvectors (and their row normalized versions) in terms of the Frobenius norm.

Theorem 5.5. *Let δ_B be the minimum expected degree of B, that is $\delta_B = \min_i[\mathscr{D}_{\mathscr{B}}]_{ii}$. Define $V_{max} = max_i V_{ii}$, which is the maximum expected volume of clusters within A. For any $\epsilon > 0$, assume the following:*

(a). There exists constant Δ, such that $\bar{\lambda}_i - \bar{\lambda}_{i+1} \geq \Delta, i = 1, ..., K-1$ and $\bar{\lambda}_K \geq \Delta$,

(b). $\delta_A > 3\ln(4N_A/\epsilon)$,

(c). $\delta_B + \tau > 4\ln(4N_A/\epsilon)$,

then,

$$\|X^* - \mathscr{X}^*\mathscr{O}\|_F \leq c_0 \frac{\sqrt{KV_{max}\ln(4N_A/\epsilon)}}{\delta_A\bar{\lambda}_K}, \tag{5.4}$$

$$\|Y^* - \mathscr{Y}^*\mathscr{O}\|_F \leq c_1 \max\Big\{\frac{\sqrt{KV_{max}N_B(\delta_B+\tau)\ln(N_A/\epsilon)}}{\delta_A\delta_B\bar{\lambda}_K^2}, \frac{\sqrt{KV_{max}N_B\ln(N_B/\epsilon)}}{\sqrt{\delta_A}\delta_B\bar{\lambda}_K}\Big\}, \tag{5.5}$$

where $\mathscr{O} \in \mathbb{R}^{K\times K}$ *is diagonal matrix with either* 1 *or* -1 *as its diagonal element.* c_0, c_1 *are two constants.*

Assumption (a) is not essential for the algorithm to work. It is only for technical convenience in further analysis. With Assumption (a), each eigenvector is identifiable up to a sign difference. This is the reason of introducing sign matrix $\mathscr{O}$. Schönemann (1966) shows how the singular value decomposition gives the proper $\mathscr{O}$ that aligns the empirical and population eigenvectors.

Next we use Theorem 5.5 to derive a bound on the mis-clustering rate of mRSC. Definition 5.6 defines "mis-clustered", it follows Qin and Rohe (2013). Recall that the algorithm applies the k-means algorithm to the rows of $F = [X^{*T}; Y^{*T}]^T$, where each row is a point in $\mathbb{R}^K$. Each row is assigned to one cluster, and each of these clusters has a centroid from k-means. Define $C_1, \ldots, C_n \in \mathbb{R}^K$ such that C_i is the centroid corresponding to the i'th row of F. Similarly, run k-means on the rows of the population version $\mathscr{F} = [\mathscr{X}^*; \mathscr{Y}^*]$ and define the population centroids $\mathcal{C}_1, \ldots, \mathcal{C}_n \in \mathbb{R}^K$. In essence, we consider node i correctly clustered if C_i is closer to $\mathcal{C}_i$ than it is to any other $\mathcal{C}_j$ for all j with $Z_j \neq Z_i$.

Definition 5.6. *If* $C_i\mathscr{O}$ *is closer to* $\mathcal{C}_i$ *than it is to any other* $\mathcal{C}_j\mathscr{O}$ *for* j *with* $Z_j \neq Z_i$*, then we say that node* i *is correctly clustered. Define the set of mis-clustered nodes:*

$$\mathscr{M} = \{i : \exists j \neq i, s.t. \|C_i\mathscr{O} - \mathcal{C}_i\|_2 > \|C_i\mathscr{O} - \mathcal{C}_j\|_2\}. \tag{5.6}$$

Here $\mathscr{O}$ is to adjust the sign difference between each columns of F and $\mathscr{F}$.

The next theorem bounds the mis-clustering rate $|\mathscr{M}|/N$.

Theorem 5.7. ***(Main Theorem)*** *Suppose* $G \in \{0,1\}^{N\times N}$ *is an adjacency matrix of a graph* $G(V, E)$ *generated from the DC-SBM with* K *blocks and parameters* $\{\boldsymbol{P}, Z, \Theta\}$*. Let* $V = V_A \cup V_B, V_A \cap V_B = \emptyset$ *be a partition of vertex set and* $|V_A| = N_A, |V_B| = N_B$ *chosen independent of* G*. Define* $\mathscr{M}$*, the set of mis-clustered nodes, as in Definition 5.6. Let* δ_A, δ_B *be the minimum expected degree of* A *and* B*. As* $N_A \to \infty$ *and* $N_B \to \infty$*, assume the following:*

(a). There exists constant Δ*, such that* $\bar{\lambda}_i - \bar{\lambda}_{i+1} \geq \Delta, i = 1, ..., K-1$ *and* $\bar{\lambda}_K \geq \Delta$*,*

(b). $\delta_A = \omega(\ln N_A)$*,*

(c). $\delta_B + \tau = \omega(\ln N_B)$*,*

then the mis-clustering rate of mRSC with regularization constant τ *is bounded,*

$$\frac{|\mathscr{M}|}{N} = O_p\bigg(\frac{N_A}{N} \bullet \frac{KV_{max}\ln N_A}{N_A\delta_A^2} + \frac{N_B}{N} \bullet \frac{KV_{max}}{\delta_A\delta_B}\max\big\{\frac{(\delta_B+\tau)\ln N_A}{\delta_A\delta_B}, \frac{\ln N_B}{\delta_B}\big\}\bigg) \quad (5.7)$$

Remark 1 (Interaction between A and B): Theorem 5.7 gives a general bound on the misclustering rate of mRSC under the DC-SBM. The bound consists of two parts, the first part bounds the misclustering rate within V_A. This bound is identical to the standard results for spectral clustering as stated in chapter 2. The second part bounds the misclustering rate within V_B. For the second part, notice the denominator includes both the minimum expected degree of A and B. The quality of the bound depends heavily on the interaction between expected degree density of both A and B. When A is dense, the assumption on δ_B is weakened. If A is sparse ($\delta_A \asymp \ln N_A$), then even of B is dense($\delta_A \asymp N_A$), weak consistency can not be achieved.

Remark 2(Applying to standard SBM): We can further interpret the result under the standard Stochastic Blockmodel by the following corollary.

Corollary 5.8. *Under the SBM, if the condition number $\kappa(V)$ is bounded, where $V = diag[(Z_A^T\Theta_A Z_A)P(Z_A^T\Theta_A Z_A)\mathbf{1}]$, then with the same assumption as in Theorem 5.7, the mis-clustering rate of mRSC with regularization constant $\tau = 0$ is bounded,*

$$\frac{|\mathscr{M}|}{N} = O_p\bigg(\frac{N_A}{N} \bullet \frac{\ln N_A}{\delta_A} + \frac{N_B}{N} \bullet \max\big\{\frac{N_A\ln N_A}{\delta_A\delta_B}, \frac{N_A\ln N_B}{\delta_B^2}\big\}\bigg) \quad (5.8)$$

If $N_A \asymp N^\alpha, \alpha \in (0,1]$ and $\delta_A/\ln N \to \infty$, then the bound is simplifies as

$$\frac{|\mathscr{M}|}{N} = O_p\bigg(\max\big\{\frac{N^\alpha\ln N}{\delta_A\delta_B}, \frac{N^\alpha\ln N}{\delta_B^2}\big\}\bigg).$$

In order for weak consistency, it requires $\delta_A\delta_B = \omega(N^\alpha\ln N)$ and $\delta_B = \omega(\sqrt{N^\alpha\ln N})$.

Remark 3(regularization constant τ): It appears that τ is on the numerator and hence makes looser bound. However, it also appears in assumption (c). In fact, regularization weakens the assumption on the minimum expected degree of B and makes the bounded more general to those nodes with low degrees.

5.5 Simulation Study

Experiment 1

This experiment compares different sampling schemes for the mRSC under the DC-SBM: random sampling, weighted sampling, sampling with hard threshold. More specifically, the five algorithms are:

1. mRSC_random_sample: random sample 10% nodes as V_A.
2. mRSC_weighted_sample: sample 10% nodes as V_A with weight proportional to node degree.
3. mRSC_degree_threshold: $V_A = \{i, D_{ii} \geq quantile(D, 90\%)\}$.
4. mRSC_oracle: $V_A = \{i, \mathscr{D}_{ii} \geq quantile(\mathscr{D}, 90\%)\}$.
5. Regularized spectral clustering(RSC). RSC replaces Graph Laplacian L with L_τ in spectral clustering. L_tau is defined as $D_\tau^{-1/2} G D_\tau^{-1/2}$ where $D_\tau = D + \tau I$.

Throughout the simulation section, tuning parameter τ in RSC and mRSC are both fixed to me 1. It is suggested by Qin and Rohe (2013), which states that the RSC is not very sensitive to tuning parameter.

Networks with sizes ranging from 900 to 10800 are generated from the DC-SBM with cluster number $K = 3$ and parameter Θ drawn from the power law distribution with lower bound $x_{min} = 1$ and shape parameter $\beta = 3$. For each fixed size, 50 networks are generated. Define the signal to noise ratio to be the expected number of in-block edges divided by the expected number of out-block edges. Throughout the simulations, the SNR is set to four and the expected average degree is set to $8/900 * N$, where N is the network size.

The upper panel of Figure 1 plots network size against the mis-clustering rate for the mRSC with four sampling scheme and the RSC. Each point is the average of 50 sampled networks. Each line represents one method. If a method assigns more than 95% of the nodes into one block, then we consider all nodes to be mis-clustered. The experiment shows that (1) sampling nodes with high degrees performs better than random sampling and weighted sampling. Its performance is close to the oracle method where we use information of expected degrees which is inaccessible in practice. (2) RSC works best since it utilized full network while the mRSC uses 10% of the information. (3). The lower panel plots on log scale of the mis-clustering rate. Under log scale, mis-clustering rates of all five methods converges to zero at the same rate but with different constants.

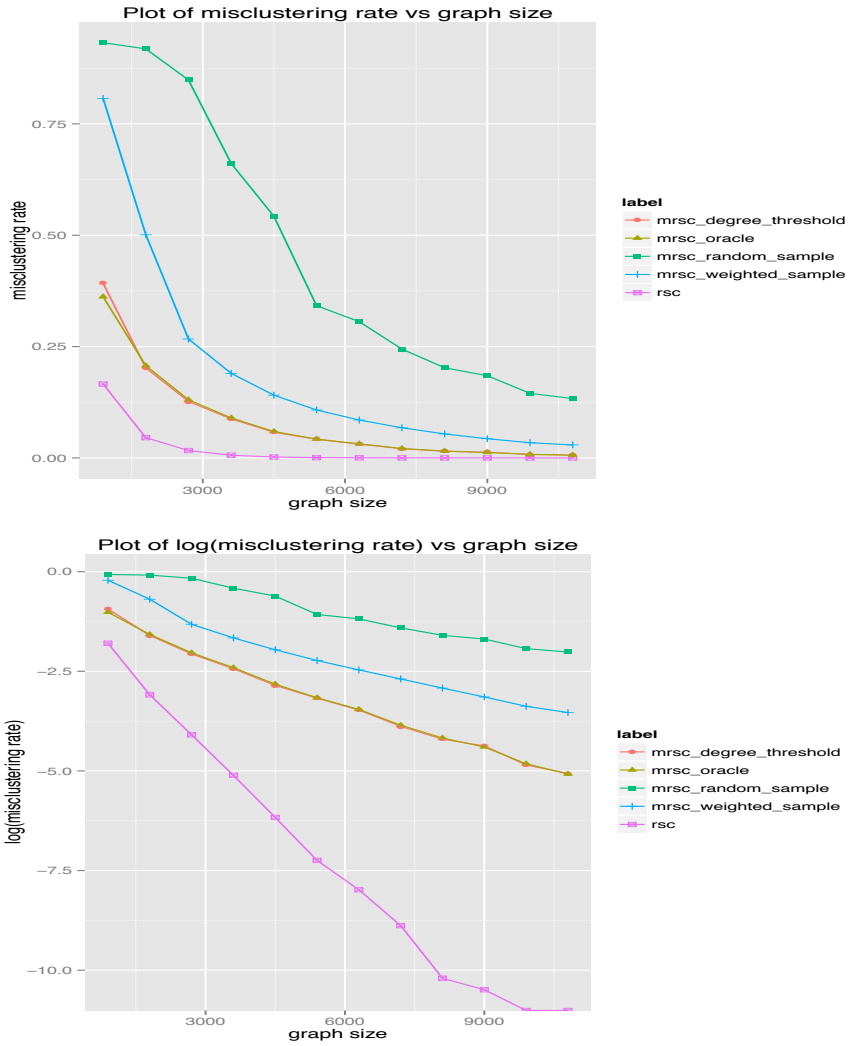

Figure 5.1: Upper Panel: Comparison of Performance for RSC, mRSC_random_sample, mRSC_weighted_sample, mRSC_degree_threshold, mRSC_oracle for networks with different size under the DC-SBM. Lower Panel: Same as left panel but with misclustering rate in log scale.

Experiment 2

The second simulation study the influence of degree heterogeneity on the performance of mRSC. The Θ parameters (from the DC-SBM) are drawn from the power law distribution with lower bound $x_{min} = 1$ and shape parameter $\beta \in \{2, 2.25, 2.5, 2.75, 3, 3.25, 3.5\}$. A smaller β indicates to greater degree heterogeneity. For each fixed β, fifty networks are sampled. In each sample, $K = 3$ and each block contains 300 nodes ($N = 900$). The SNR is set to four and the expected average degree is set to eight.

Three algorithms are compared: (1). mRSC_0.5: $V_A = \{i, D_{ii} \geq quantle(D, 50\%)\}$, (2). mRSC_0.25: $V_A = \{i, D_{ii} \geq quantle(D, 75\%)\}$, (3). RSC. Figure 2 plots β against the mis-clustering rate for mRSC_0.5, mRSC_0.25 and RSC . Each point is the average of 50 sampled networks with standard error.

It shows that (1) if the degrees are highly heterogeneous ($\beta \leq 2.5$), mRSC is more stable than RSC. (2) the performance of mRSC is not monotone in the level of heterogeneity.

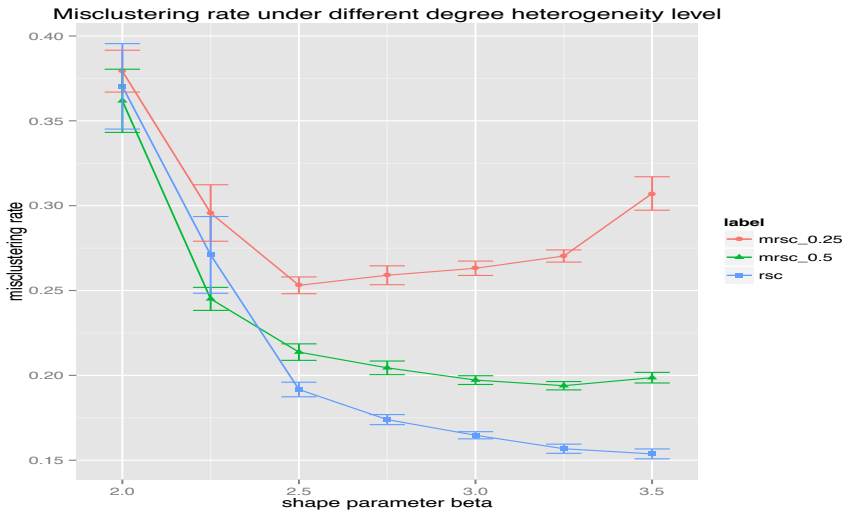

Figure 5.2: Comparison of mRSC and RSC under different degree heterogeneity levels.

Experiment 3

This experiment compares mRSC with the naive Nyström algorithm and RSC under the SBM with no degree heterogeneity. Networks with size from 3000 to 33000 are generated from the SBM with in-block linkage probability $p = 0.01$ and between-block linkage probability $q = p/3$ In this experiment, the average degree is $16/3000 * N$. For each setting, the results are averaged over 10 samples of the network.

Figure 3 shows (a) Even in networks with no degree heterogeneity, sampling high degree nodes is still better than random sapling. (b) Proper regularization and normalization makes mRSC perform better than the naive Nyström algorithm. (c). RSC gives lowest mis-clustering rate, yet for large networks ($N > 20000$), RSC suffers from memory limitation.

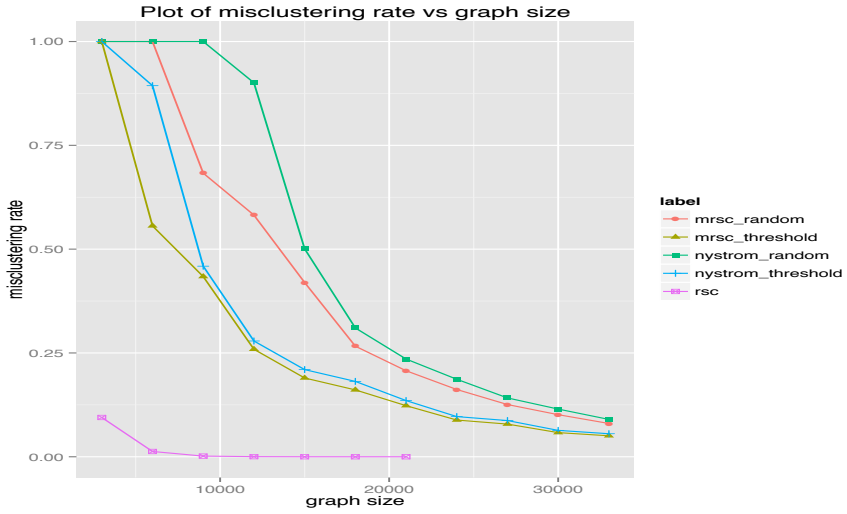

Figure 5.3: Comparison of mRSC, RSC and Nyström methods under the SBM. mRSC_random: random sample 10% nodes as V_A. mRSC_threshold: $V_A = \{i, D_{ii} \geq quantle(D, 90\%)\}$. nystrom_random: random sample 10% nodes as V_A. nystrom_threshold: $V_A = \{i, D_{ii} \geq quantle(D, 90\%)\}$.

5.6 Discussion

This chapter proposes a memory efficient spectral algorithm for community detection – mRSC. We further studies the algorithm under a parametric model DC-SBM

that allows for degree heterogeneity. To account for the degree heterogeneity, our algorithm requires several key steps of normalizing and regularizing, which distinguish mRSC from previous related methods. Theorem 5.7 gives a general bound on the misclustering rate of mRSC under DC-SBM and justifies the importance of normalizing and regularizing.

The algorithm can be applied to online learning settings. Consider a fixed network of size N, whose spectral decomposition and K cluster centers in the eigenspace have been computed and stored. When a new node joins the network. We can apply step 3 in mRSC to compute Y^*_{N+1} and assign it to cluster t iff Y^*_{N+1} is closest to cluster center C_t. Clustering the additional node requires time and space complexity $O(N)$.

Finding a good partition is crucial for mRSC to perform well. Empirically, for graphs with general degrees, their are two ways, (1) Set a degree threshold T, and assign nodes with degree greater than T to V_A, the rest nodes forms V_B. (2) with probability proportional to $\frac{D_i}{\sum_j^N D_j}$, assign node i to V_A. Developing theoretical guarantees of these sampling schemes under the DC-SBM for community detection is a future research direction.

Chapter 6

The Blessing of Transitivity in Sparse and Stochastic Networks

6.1 Introduction

Advances in information technology have generated a barrage of data on highly complex systems with interacting elements. Depending on the substantive area, these interacting elements could be metabolites, people, or computers. Their interactions could be represented in chemical reactions, friendship, or some type of communication. Networks (or graphs) describe these relationships. Therefore, the questions about the relationships in these data are questions regarding the structure of networks. Several of these questions are more naturally phrased as questions of inference; they are questions not just about the realized network, but about the mechanism that generated the network. To study questions in graph inference, it is essential to study algorithms under model parameterizations that reflect the fundamental features of the network of interest.

Sparsity and transitivity are two fundamental and recurring features. In sparse graphs, the number of edges in a network is orders of magnitude smaller than the number of possible edges; the average element has 10s or 100s of relationships, even in networks with millions of other elements. Transitivity describes the fact that friends of friends are likely to be friends. The interaction of these simple and localized features has profound implications for determining the set of realistic statistical models. They imply that in large sparse graphs, there are local and dense regions. This is the blessing of transitivity.

One essential inferential goal is the discovery of communities or clusters of highly connected actors. These form essential feature in a multitude of empirical networks,

and identifying these clusters helps answer vital scientific questions in many fields. A terrorist cell is a cluster in the communication network of terrorists; web pages that provide hyperlinks to each other form a community that might host discussions of a similar topic; a cluster in the network of biochemical reactions might contain metabolites with similar functions and activities. Several papers, that are briefly reviewed below, have proved theoretical results for various graph clustering algorithms under the Stochastic Blockmodel, a parametric model, where the model parameters correspond to a true partition of the nodes. Often, these estimators are also studied under the exchangeable random graph model, a non-parametric generalization of the Stochastic Blockmodel. The overarching goal of this paper is to show (1) how sparse and transitive models require a novel asymptotic regime and (2) how the blessing of transitivity makes edges become more informative in the asymptote, allowing for statistical inference even when cluster size and expected degrees do not grow with the number of nodes.

The first part of this paper studies how sparsity and transitivity interact in the Stochastic Blockmodel, and more generally, in the exchangeable random graph model. Interestingly, if a Stochastic Blockmodel is both sparse and transitive, then it has small blocks. The second part of this paper (1) introduces an intuitive and fast local clustering technique to find small clusters; (2) proposes the local Stochastic Blockmodel, which presumes a single stochastic block is embedded in a sparse and potentially adversarially chosen network; and (3) proves that if the proposed local clustering technique is initialized with any point in the stochastic block, then it returns the block with high probability. Figure 6.1 illustrates the types of clusters found by the proposed algorithm; in this case from a social network on epinions.com containing over 76,000 people.

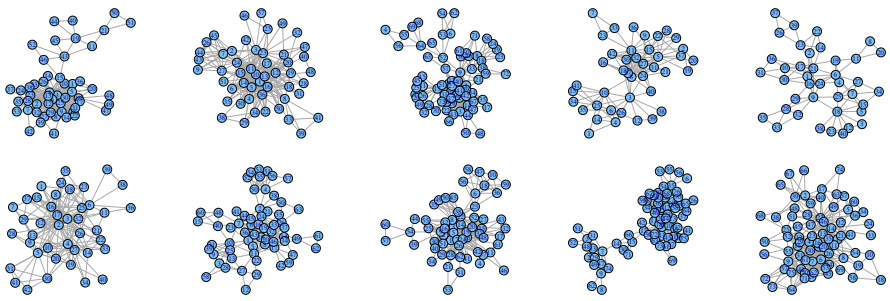

Figure 6.1: Local clusters from a sparse 76k node social network from epinions.com. Created with the igraph library in R (Csardi and Nepusz, 2006).

Preliminaries

Networks, or graphs, are represented by a vertex set and an edge set, $G = (V, E)$, where $V = \{1, \ldots, n\}$ contains the actors and

$$E = \{(i, j) : \text{ there is an edge from } i \text{ to } j\}.$$

The edge set can be represented by the adjacency matrix $A \in \{0, 1\}^{n \times n}$:

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

This paper only considers undirected graphs. That is, $(i, j) \in E \Rightarrow (j, i) \in E$. The adjacency matrix of such a graph is symmetric. Many of the results in the paper have a simple extension to weighted graphs, where $A_{ij} \geq 0$. For simplicity, we only discuss unweighted graphs. For $i \in V$, let $d_i = \sum_\ell A_{i\ell}$ denote the degree of node i. Define N_i as the neighborhood of node i,

$$N_i = \{j : (i, j) \in E\}.$$

Define the transitivity ratio of A as

$$trans(A) = \frac{\text{number of closed triplets in } A}{\text{number of connected triples of vertices in } A}.$$

Watts and Strogatz (1998) introduced an alternative measure of transitivity, the clustering coefficient. The local clustering coefficient, $C(i)$, is defined as the density of the subgraph induced by N_i, that is the number of edges between nodes in N_i divided by the total number of possible edges. The clustering coefficient for the entire network is the average of these values.

$$C = \frac{1}{n} \sum_i C(i)$$

This is related to the triangles in the graph because an edge (j, k) between two nodes in N_i makes a triangle with node i.

Statistical models for random networks

Suppose $A = \{A_{ij} : i, j \geq 1\}$ is an infinite array that is binary, symmetric, and random. If A is (jointly) exchangeable, that is

$$A \stackrel{d}{=} A_{\sigma,\sigma} = \{A_{\sigma(i),\sigma(j)} : i, j \geq 1\},$$

for any arbitrary permutation σ, then the Aldous-Hoover representation says there exists *i.i.d.* random variables $\xi_1, \xi_2, \ldots$ and an additional independent random variable α such that conditional on these variables, the elements of A are statistically independent (Hoover, 1979; Aldous, 1981; Kallenberg, 2005). The global parameter α controls the edge density and for ease of notation, it is often dropped (Bickel et al., 2011). One should think of this result as an extension of DeFinetti's Theorem to infinite exchangeable arrays.

While this representation has only been proven to be equivalent to exchangeability for infinite arrays, it is convenient to adopt this representation for finite graphs.

Definition 6.1. *Symmetric adjacency matrix $A \in \{0,1\}^{n\times n}$ follows the **exchangeable random graph model** if there exists i.i.d. random variables $\xi_1, \ldots, \xi_n$ such that probability distribution of A satisfies*

$$P(A|\xi_1, \ldots, \xi_n) = \prod_{i<j} P(A_{ij}|\xi_i, \xi_j).$$

For brevity, we will sometimes refer to this as the exchangeable model.

Independently of the research on infinite exchangeable arrays, Hoff et al. (2002) proposed the latent space model which assumes that (1) each person has a set of latent characteristics (e.g. past schools, current employer, hobbies, etc.) and (2) it is only these characteristics that produce the dependencies between edges. Specifically, conditional on the latent space characteristics, the relationships (or lack thereof) are independent. The Latent Space Model is equivalent to the exchangeable model in Definition 6.1.

The Stochastic Blockmodel is an exchangeable model that was first defined in Holland and Leinhardt (1983).

Definition 6.2. *The **Stochastic Blockmodel** is an exchangeable random graph model with $\xi_1, \ldots, \xi_n \in \{1, \ldots, K\}$ and*

$$P(A_{ij} = 1|\xi_i, \xi_j) = \Theta_{\xi_i,\xi_j}$$

for some $\Theta \in [0,1]^{K\times K}$.

In this model, the ξ_i correspond to group labels. The diagonal elements of Θ correspond to the probability of within-block connections. The off-diagonal elements correspond to the probabilities of between-block connections. When the diagonal elements are sufficiently larger than the off-diagonal elements, then the sampled network will have clusters that correspond to the blocks in the model.

The next subsection briefly reviews the existing literature that examines the consistency of various estimators for the partition created by the latent variables $\xi_1, \ldots, \xi_n$ in the Stochastic Blockmodel.

Previous research

This paper builds on an extensive body of literature examining various types of statistical estimators for the latent partition $\xi_1, \ldots, \xi_n$ in the Stochastic Blockmodel. These estimators fall into four different categories.[1]

1. Several have studied estimators that are solutions to discrete optimization problems (e.g. Bickel and Chen (2009); Choi et al. (2012); Zhao et al. (2011b,a); Flynn and Perry (2012)). These objective functions are the likelihood function for the Stochastic Blockmodel or the Newman-Girvan modularity (Newman and Girvan, 2004), a measure that corresponds to cluster quality.

2. Others have have studied various approximations to the likelihood that lead to more computationally tractable estimators. For example, Celisse et al. (2011) and Bickel et al. (2012) studied the variational approximation to the likelihood function and Chen et al. (2012a) studied the maximum pseudo-likelihood estimator.

3. Building on on spectral graph theoretic results (Donath and Hoffman, 1973; Fiedler, 1973), several researchers have studied the statistical performance of spectral algorithms for estimating the partition in the Stochastic Blockmodel (McSherry, 2001; Dasgupta et al., 2004; Giesen and Mitsche, 2005; Coja-Oghlan and Lanka, 2009; Rohe et al., 2011; Rohe and Yu, 2012; Chaudhuri et al., 2012; Jin, 2015; Sussman et al., 2012b; Fishkind et al., 2013). Others have studied estimators that are solutions to semi-definite programs (Ames and Vavasis, 2010; Oymak and Hassibi, 2011; Chen et al., 2012b).

[1]The works cited in this section give a sample of the previous literature on statistical inference for the Stochastic Blockmodel; it is not meant to be an exhaustive list. In particular, there are several highly relevant papers in the Computer Science literatures on (i) the planted partition model and (ii) the planted clique problem. The curious reader should consult the references in Ames and Vavasis (2010) and Chen et al. (2012b).

4. More recently, Bickel et al. (2011); Channarond et al. (2011); Rohe and Yu (2012) have developed methods to stitch together network motifs, or simple "local" measurements on the network, in a way that estimates the partition in the Stochastic Blockmodel. Bickel et al. (2011) draws a parallel between this motif-type of estimator and method of moments estimation.

All of the previous results described above are sensitive to both (a) the population of the smallest block and (b) the expected number of edges in the graph; larger blocks and higher expected degrees lead to stronger conclusions. This limitation arrises because the proofs rely on some form of concentration of measure for a function of sufficiently many variables. Bigger blocks and more edges yield more variables, and thus, more concentration. This paper shows how transitivity leads to a different type of concentration of measure, where each edge becomes asymptotically more informative. As such, the results in this paper extend to blocks of fixed sizes and bounded expected degrees.

Section 6.3 proposes the `LocalTrans` algorithm that exploits the triangles built by network transitivity. As such, it is most similar to motif-type estimators in bullet (4). Our analysis of `LocalTrans` is the first to study whether a local algorithm (i.e. initialized from a single node) can estimate a block in the Stochastic Blockmodel. The emphasis on local structure aligns with the aims of network scan statistics; these compute a "local" statistic on the subgraph induced by N_i for all i and then return the maximum over all i. In the literature on network scan statistics, Rukhin and Priebe (2012) and Wang et al. (2013) have previously studied the anomaly detection properties under random graph models, including a version of the Stochastic Blockmodel.

6.2 Transitivity in sparse exchangeable random graph models

In this section, Proposition 6.3 and Theorem 6.4 show that previous parameterizations of the sparse exchangeable models and sparse Stochastic Blockmodels lack transitivity in the asymptote. That is, the sampled networks are asymptotically sparse, but they are not asymptotically transitive. Theorem 6.8 concludes the section by describing a parameterizations that produce sparse *and* transitive networks.

Define

$$p_{\max} = \max_{\xi_i, \xi_j} P(A_{ij} = 1 | \xi_i, \xi_j) \tag{6.2}$$

as the largest possible probability of an edge under the exchangeable model. In the statistics literature, previous parameterizations of sparse Stochastic Blockmodels, and

sparse exchangeable models, have all ensured sparsity by sending $p_{\max} \to 0$.

Define

$$p_\Delta = P(A_{uv} = 1 | A_{iu} = A_{iv} = 1) \tag{6.3}$$

a population measure of the transitivity in the model. It is the probability of completing a triangle, conditionally on already having two edges.

By sending $p_{\max}$ to zero, a model removes transitivity.

Proposition 6.3. *Under the exchangeable random graph model* (6.1)

$$p_\Delta \leq p_{\max},$$

where these probabilities are defined in equations (6.2) *and* (6.3)*.*

The next theorem gives conditions that imply the transitivity ratio of the sampled network converges to zero.

Theorem 6.4. *Under the exchangeable random graph model (Definition 6.1), define* $\lambda_n = E(d_i)$ *as the expected node degree. If* $\lambda_n \to \infty$, $\lambda_n = o(n)$, *and*

$$p_{\max} = o\left(\frac{P(A_{ij} = 1)}{P(A_{ij} = 1 | A_{i\ell} = 1)}\right), \tag{6.4}$$

where $p_{\max}$ *is defined in* (6.2)*, then* $trans(A) \xrightarrow{P} 0$.

A proof of this theorem can be found in the appendix.

The denominator on the right hand side of Equation (6.4) quantifies how many edges are adjacent to the average edge, and thus controls how many 2-stars are in the graph. It can be crudely bounded with the maximum expected degree over the latent ξ_i. For

$$\lambda_n^{\max} = \max_{\xi_i} \; E(d_i | \xi_i),$$

it follows that

$$P(A_{ij} = 1 | A_{i\ell} = 1) \leq \lambda_n^{\max}/n.$$

Corollary 6.5. *Under the exchangeable random graph model (Definition 6.1), define* $\lambda_n = E(d_i)$ *as the expected node degree. If* $\lambda_n \to \infty$, $\lambda_n = o(n)$, *and*

$$p_{\max} = o\left(\frac{\lambda_n}{\lambda_n^{\max}}\right),$$

then $trans(A) \xrightarrow{P} 0$.

So, ensuring sparsity by sending $p_{\max}$ to zero removes transitivity both from the model and from the sampled network. The next subsection investigates the implications of restricting $p_{\max} > \epsilon > 0$ in sparse networks.

Implications of non-vanishing $p_{\max}$ in the Stochastic Blockmodel

It is easiest to consider a simplified parameterization of the Stochastic Blockmodel. The following parameterization is also called the planted partition model.

Definition 6.6. *The **four parameter Stochastic Blockmodel** is a Stochastic Blockmodel with K blocks, exactly s nodes in each block, $\Theta_{ii} = p$, and $\Theta_{ij} = r$ for $i \neq j$.*

In this model, (1) $n = Ks$, (2) the "in-block" probabilities are equal to p, and (3) the "out-of-block" probabilities equal to r. Moreover, the expected degree[2] of each node is

$$\text{Expected degree under the four parameter model} = sp + (n - s)r. \tag{6.5}$$

Define this quantity as λ_n. Under the four parameter model, p is analogous to $p_{\max}$.

Note that $sp \leq \lambda_n$ and

$$n\frac{p}{\lambda_n} \leq K.$$

In sparse and transitive graphs, λ_n is bounded and p is non-vanishing. In this regime, K grows proportionally to n. The following proposition states this fact in terms of s, the population of each block.

Proposition 6.7. *Under the four parameter Stochastic Blockmodel, if p is bounded from below, then*

$$s = O(\lambda_n)$$

where s is the population of each block and λ_n is the expected node degree.

The following Theorem shows that graphs sampled from this parameterization are asymptotically transitive.

[2]For ease of exposition, this formula allows self-loops.

Theorem 6.8. *Suppose that A is the adjacency matrix sampled from the four parameter Stochastic Blockmodel (Definition 6.6) with n nodes. If $p > \epsilon > 0$, $r = O(n^{-1})$, and $s \geq 3$, then as $n \to \infty$*

$$trans(A) \xrightarrow{P} c > 0,$$

where c is a constant which depends on p, r, s.

Remark: If $r = \frac{c_0}{n}$, for some constant c_0, then

$$c \approx \frac{p^3 s^2}{p^2 s^2 + c_0^2 + 2spc_0}.$$

The appendix contains the proof for Theorem 6.8.

The fixed block size asymptotics in Theorem 6.8 align with two pieces of previous empirical research suggesting the "best" clusters in massive networks are small. Leskovec et al. (2009) found that in a large corpus of empirical networks, the tightest clusters (as judged by several popular clustering criteria) were no larger than 100 nodes, even though some of the networks had several million nodes. This result is consistent with findings in Physical Anthropology. Dunbar (1992) took various measurements of brain size in 38 different primates and found that the size of the neocortex divided by the size of the rest of the brain had a log-linear relationship with the size of the primates natural communities. In humans, the neocortex is roughly four times larger than the rest of the brain. Extrapolating the log-linear relationship estimated from the 38 other primates, Dunbar (1992) suggests that the average human does not have the social intellect to maintain a stable community larger than roughly 150 people (colloquially referred to as Dunbars number). Leskovec et al. (2009) found a similar result in several other networks that were not composed of humans. The research of Leskovec et al. (2009) and Dunbar (1992) suggests that the block sizes in the Stochastic Blockmodel should not grow asymptotically. Rather, block sizes should remain fixed (or grow very slowly).

Implications for the exchangeable model

The interaction between sparsity and transitivity also has surprising implications in the more general exchangeable random graph model. To see this, note that in the exchangeable model, it is sufficient to assume that $\xi_1, \ldots, \xi_n$ are $i.i.d.$ Uniform$(0, 1)$ (Kallenberg, 2005). Then, the conditional density of ξ_i and ξ_j given $A_{ij} = 1$ is

$$c(\xi_i, \xi_j) = \frac{P(A_{ij} = 1 | \xi_i, \xi_j)}{P(A_{ij} = 1)} \tag{6.6}$$

When $p_{\max}$ does not converge to zero, there exist values of ξ_i^* and ξ_j^* such that $P(A_{ij} = 1|\xi_i, \xi_j)$ does not converge to zero. However, in a sparse graph, the edge density $P(A_{ij} = 1)$ (in the denominator of Equation (6.6)) converges to zero. So, $c(\xi_i^*, \xi_j^*)$ is asymptotically unbounded. For example, in the popular $P(A_{ij} = 1) = O(1/n)$ limit, $c(\xi_i^*, \xi_j^*)$ is proportional to n. In a sense, as a sparse and transitive network grows, each edge becomes more informative. This is the blessing of transitivity in sparse and stochastic networks.

This asymptotic setting, where $p_{\max}$ is bounded from below, makes for an entirely different style of asymptotic proof; the asymptotic power comes from the fact that each edge becomes increasingly informative in the asymptote. Previous consistency proofs rely on concentration of measure for functions of several independent random variables (i.e. several edges). In the sparse and transitive asymptotic setting, concentration follows from the blessing of transitivity, allowing asymptotic results with fixed block sizes and bounded degrees. For example, in Theorems 6.11 and 6.13 in the next section, neither the block size nor node degree grows in the asymptote.

6.3 Local (model + algorithm + results)

This section investigates clustering, or community detection, in sparse and transitive networks. Following the results of the last section, sparse and transitive communities are small. As such, this section is focused on finding small clusters of nodes. In an attempt to strip away as many assumptions as possible from the Stochastic Blockmodel, this section

1. proposes a "localized" model with a small and transitive cluster embedded in a large and sparse graph (that could be chosen by an adversary),

2. introduces a novel local clustering algorithm that explicitly leverages the graphs transitivity, and

3. shows that this local algorithm will discover the cluster in the localized model with high probability.

Similarly to the last section, the interaction between sparsity and transitivity provides for these results, enabling both the fast algorithm and the fixed block asymptotics.

The local Stochastic Blockmodel

The "local" Stochastic Blockmodel (defined below) presumes that a small set of nodes S_* constitute a single block and the model parameterizes how these nodes relate to each other and how they relate to the rest of the network.

Definition 6.9. *Suppose $A \in \{0,1\}^{(n+s)\times(n+s)}$ is an adjacency matrix on $n+s$ nodes. If there is a set of nodes S_* with $|S_*| = s$ and*

1. *$i, j \in S_*$ implies $P(A_{ij} = 1) \geq p_{in}$,*
2. *$i \in S_*$ and $j \in S_*^c$ implies $P(A_{ij} = 1) \leq p_{out}$,*
3. *the random variables $\{A_{ij} : \forall i \in S_* \text{ and } \forall j\}$ are both mutually independent and also independent of the rest of the graph*

*then A follows the **local Stochastic Blockmodel** with parameters S_*, p_{in}, p_{out}.*

The only assumption that this definition makes about edges outside of S_* (that is, (i, j) with $i, j \notin S_*$) is that they are independent of the edges that connect to at least one node in S_*. So, the edges outside of S_* could be chosen by an adversary, as long as the adversary does not observe the rest of the graph. The theorems below will add an additional assumption that the average degree (within S_*^c) must be not too large (i.e. it must be sparse).

Local clustering with transitivity

A local algorithm searches around a seed node for a tight community that includes this seed node. Several papers have demonstrated the computational advantages of local algorithms for massive networks (Priebe et al., 2005; Spielman and Teng, 2008). In addition to fast running times and small memory requirements, the local results are often more easily interpretable (Priebe et al., 2005) and yield what appear to be "statistically regularized" results when compared to other, non-local techniques (Leskovec et al., 2009; Clauset, 2005; Liao et al., 2009). Andersen et al. (2006); Andersen and Chung (2007); Andersen and Peres (2009) have studied the running times and given perturbation bounds showing that local algorithms can approximate the graph conductance. With the exception of Rukhin and Priebe (2012) and Wang et al. (2013), the previous literature has not addressed the statistical properties of local graph algorithms under statistical models.

Given an adjacency matrix A and a seed node i, this section defines an algorithm that finds a clusters around node i. This algorithm has a single tuning parameter

cut that balances the size of the cluster with the tightness of the cluster. Smaller values of *cut* return looser clusters. The algorithm initializes the cluster with the seed node $S = \{i\}$. It then repeats the following step: For every edge between a node in S ($j \in S$) and a node not in S ($\ell \in S^c$), add ℓ to S if there are at least *cut* nodes that connect to both ℓ and j (this ensures that (i, j) is contained in at least *cut*-many triangles). Stop the algorithm if all edges across the boundary of S are contained in fewer than *cut*-many triangles.

Algorithm 6.1 LocalTrans(A, i, cut)

1. Initialize set S to contain node i.
2. For each edge (i, ℓ) on the boundary of S ($i \in S$ and $\ell \notin S$) calculate $T_{i\ell}$:

$$T_{i\ell} = \sum_k A_{ik} A_{k\ell}.$$

3. If there exists any edge(s) (i, ℓ) on the boundary of S with $T_{i\ell} \geq cut$, then add the corresponding node(s) ℓ to S and return to step 2.
4. Return S.

Consider $\texttt{LocalTrans}(A, j, \tau)$ as a function that returns a set of nodes, then

$$i \in \texttt{LocalTrans}(A, j, cut) \implies j \in \texttt{LocalTrans}(A, i, cut).$$

Moreover, if $cut^+ > cut$ then,

$$\texttt{LocalTrans}(A, i, cut^+) \subset \texttt{LocalTrans}(A, i, cut).$$

This shows that the results of $\texttt{LocalTrans}(A, i, cut)$, for every node i and every parameter *cut*, can be arranged into a dendogram. `LocalTrans` only finds one branch of the tree. A simple and fast algorithm can find the entire tree.

To compute the entire dendogram, apply single linkage hierarchical clustering[3] to the similarity matrix

$$T = (AA) \cdot A, \text{ where } \cdot \text{ is element-wise multiplication.} \tag{6.7}$$

The computational bottleneck of this algorithm is computing T, which can be computed in $O(|E|^{3/2})$. Techniques using fast matrix multiplication can slightly

[3]This is equivalent to finding the maximum spanning tree

decrease this exponent (Alon et al., 1997).

When A contains no self loops, T_{ij} equals the number of triangles that contain both nodes i and j. We propose *single* linkage because it is the easiest to analyze and it yields good theoretical results. However, in some simulation, *average* linkage has performed better than single linkage. One could also state a local algorithm in terms of average linkage.

Algorithm 6.2 GlobalTrans(A, τ)

1. Compute the similarity matrix $T = [AA] \cdot A$, where $\cdot$ is element-wise multiplication.
2. Run single linkage hierarchical clustering on similarity matrix T, i.e. grow a maximum spanning tree.
3. Cut the dendogram at level τ, i.e. delete any edges in the spanning tree with weight smaller than τ.
4. Return the connected components.

Proposition 6.10. *Viewing* `LocalTrans` *as a function that returns a set of nodes and* `GlobalTrans` *as functions that returns a set of sets,* $\texttt{LocalTrans}(A, i, \tau) \subset \texttt{GlobalTrans}(A, \tau)$. *Moreover,*

$$\bigcup_i \texttt{LocalTrans}(A, i, \tau) = \texttt{GlobalTrans}(A, \tau).$$

Proof. Nodes i and j are in the same cluster in both $\texttt{LocalTrans}(A, i, \tau)$ and $\texttt{GlobalTrans}(A, \tau)$ if and only if there exists a path from i to j such that every edge in the path is in at least τ triangles. □

Local Inference

The next theorem shows that `LocalTrans` estimates the local block in the Local Stochastic Blockmodel with high probability.

Theorem 6.11. *Under the local Stochastic Blockmodel (Definition 6.9), if*

$$\sum_{i,j \in S_*^c} A_{ij} \leq n\lambda,$$

then

1. ***cut = 1:*** *for all* $i \in S_*$, $LocalTrans(A, i, cut = 1) = S_*$ *with probability greater than*
$$1 - \left(\frac{1}{2}s^2(1-p_{in}^2)^{s-2} + O(p_{out}^2 ns(s+\lambda))\right).$$

2. ***cut = 2:*** *for all* $i \in S_*$, $LocalTrans(A, i, cut = 2) = S_*$ *with probability greater than*
$$1 - \left(s^3(1-p_{in}^2)^{s-3} + O(p_{out}^3 ns(s+\lambda)^2)\right).$$

See the appendix for the proof of this theorem.

For the local Stochastic Blockmodel to create a transitive block with bounded expected degrees, it is necessary for p_{in} to be bounded from below and for s to be bounded from above. Then, $p_{out} = O(1/n)$ is a sufficient condition for bounded expected degrees. However, because of the inequality in bullet (2) of Definition 6.9, $p_{out} = O(1/n)$ is not a necessary condition for sparsity. As such, the restriction on p_{out} is particularly relevant. It is also fundamental to the bound in Theorem 6.11.

The approximation terms $O(p_{out}^2 ns(s+\lambda))$ and $O(p_{out}^3 ns(s+\lambda)^2)$ bound the probability of a connection in T across the boundary of S_*. The strength of these terms come from the fact that a connection across the boundary requires $cut + 1$ simultaneous edges across the boundary of S_*. Figure 6.2 gives a graphical explanation for the "+1". As such, p_{out} is raised to the *cut +1* power. When s and λ are fixed, then $p_{out} = o(n^{-1/2})$ and $p_{out} = o(n^{-1/3})$ make the approximation term asymptotically negligible for the case $cut = 1$ and $cut = 2$, respectively. Both of these settings allow the nodes in S_* have (potentially) large degrees. If $p_{out} = O(n^{-1})$, then the approximation terms become $O(n^{-1})$ and $O(n^{-2})$ respectively.

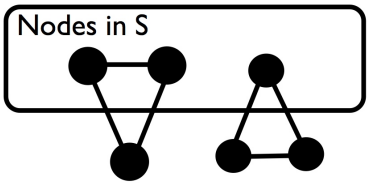

Figure 6.2: This figure illustrates the two types of triangles that contain nodes in both S_* and S_*^c. To make *one* triangle that crosses the boundary of S requires *two* edges to cross the boundary.

Theorem 2 and the algorithms `LocalTrans` and `GlobalTrans` all leverage the interaction between transitivity and sparsity, making the task of computing S and estimating S_* both algorithmically tractable and statistically feasible.

Preliminary Data Analysis

This section applies `GlobalTrans` to an online social network from the website slashdot.org, demonstrating the shortcomings of the proposed algorithms with input A and motivating the next section that uses the graph Laplacian as the input. The slashdot network contains 77 360 nodes with an average degree of roughly 12 (Leskovec et al., 2009).[4] This network is particularly interesting because it has a smaller transitivity ratio (.024) than the typical social network.

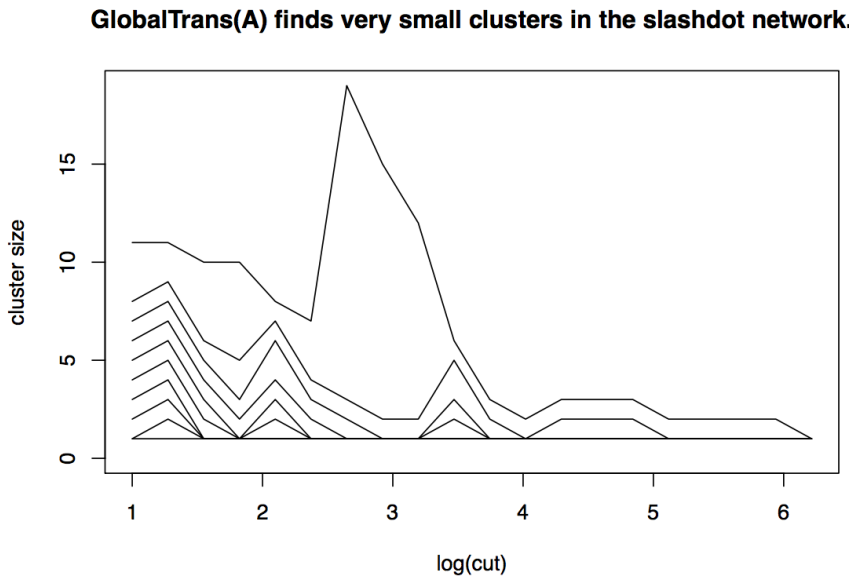

Figure 6.3: This plots the number of nodes in the largest ten clusters (ignoring a single giant cluster) found by `GlobalTrans`(A, cut) in the slashdot social network. These clusters are very small, and probably too small for many applications. Moreover, there are not that many of them.

Figure 6.3 plots the size of the ten largest clusters returned by `GlobalTrans`(A, cut) as a function of *cut* (excluding the largest cluster that consists of the majority of the graph). The values of *cut* range from 3 to 500 and they are plotted on the $\log_e$ scale. Over this range of *cut*, only two times does a cluster exceed ten nodes. While we motivated the local techniques as searching for small clusters, these clusters are

[4]This data can be downloaded at `http://snap.stanford.edu/data/soc-Slashdot0811.html`

perhaps *too* small. It suggests that there are no clusters that are adequately described by the local Stochastic Blockmodel.

One potential reason for this failure is that under the Local Stochastic Blockmodel, the probability of a connection between a node in S_* and a node in S_*^c is uniformly bounded by some value, p_{out}. The slashdot social network, like many other empirical networks, has a long tailed degree distribution. A more realistic model might allow the nodes in S_* to be more highly connected to the high degree nodes in S_*^c. The next subsection (1) proposes a "degree-corrected" local Stochastic Blockmodel, (2) proves that `LocalTrans` with a simple adjustment can estimate S_* in the degree corrected model, and (3) demonstrates how this new version of the algorithm improves the results on the slashdot social network.

6.4 The Degree-corrected Local Stochastic Blockmodel

Inspired by Karrer and Newman (2011), the degree-corrected model in Definition 6.12 makes the probability of a connection between a node $i \in S_*$ and a node $j \notin S_*$ scale with the degree of node j on the subgraph induced by S_*^c.

$$d_j^* = \sum_{\ell \in S_*^c} A_{\ell j} \tag{6.8}$$

For the following definition to make sense, we presume that d_j^* is fixed for all $j \in S_*^c$.

Definition 6.12. *Suppose $A \in \{0,1\}^{(n+s)\times(n+s)}$ is an adjacency matrix and S_* is a set of nodes with $|S_*| = s$. For $j \in S_*^c$, define d_j^* as in Equation (6.8). If*

1. *$i \in S_*$ and $j \in S_*^c$ implies*
$$P(A_{ij} = 1) \leq \frac{d_j^*}{n},$$
2. *$i, j \in S_*$ implies $P(A_{ij} = 1) \geq p_{in}$,*
3. *$\{A_{ij} : \forall j \text{ and } \forall i \in S_*\}$ are mutually independent*

*then A follows the **local degree-corrected Stochastic Blockmodel** with parameters S_*, p_{in}.*

The fundamental difference between the previous local model and this degree corrected version is the assumption that if $i \in S_*$ and $j \in S_*^c$, then

$$P(A_{ij} = 1) \leq \frac{d_j^*}{n}.$$

In the previous model, $P(A_{ij} = 1) \leq p_{out}$. This new condition can be interpreted as $P(A_{ij} = 1) \leq p_{out}\, d_j^*$ for $p_{out} = 1/n$. In this degree-corrected model, the nodes in S_* connect to more high degree nodes than they do under the previous local model.

The degree corrected model creates two types of problems for $\texttt{LocalTrans}(A, i, \tau)$. Because the high degree nodes in S_*^c create many connections to the nodes in S_*, it is more likely to create triangles with two nodes in S_*. Additionally, by definition, the high degree nodes outside of S_* have several neighbors outside of S_*. As such, it is more likely to create triangles with one node in S_* and two nodes outside of S_*. In essence, the high degree nodes create several triangles in the graph, washing out the clusters that $\texttt{LocalTrans}(A, i, \tau)$ can detect. To confront this difficulty, it is necessary to down weight the triangles that contain high degree nodes.

The graph Laplacian

Similarly to the adjacency matrix, the normalized graph Laplacian represents the graph as a matrix. In both spectral graph theory and in spectral clustering, the graph Laplacian offers several advantages over the adjacency matrix (Chung, 1997; Von Luxburg, 2007). The spectral clustering algorithm uses the eigenvectors of the normalized graph Laplacian, not the adjacency matrix, because the normalized Laplacian is robust to high degree nodes (Von Luxburg, 2007).

For adjacency matrix A, define the diagonal matrix D and the normalized graph Laplacian L, both elements of $R^{n \times n}$, in the following way

$$\begin{aligned} D_{ii} &= d(i) \\ L_{ij} &= [D^{-1/2} A D^{-1/2}]_{ij} = \frac{A_{ij}}{\sqrt{D_{ii} D_{jj}}}. \end{aligned} \tag{6.9}$$

Some readers may be more familiar defining L as $I - D^{-1/2} W D^{-1/2}$. For our purposes, it is necessary to drop the $I-$.

The last section utilized the matrix $T = [AA] \cdot A$ to find the triangles in the graph. To confront the degree corrected model, the next theorem uses $[LL] \cdot L$ instead. The interpretation of this matrix is similar to T. It differs because it down weights the contribution of each triangle by the inverse product of the node degrees. For

example, where a triangle between nodes i, j, k would add 1 to element T_{ij}, it would add $(d(i)d(j)d(k))^{-1}$ to the i, jth element of $[LL] \cdot L$.

Some versions of spectral clustering use the random walk graph Laplacian, an alternative form of the normalized graph Laplacian.

$$L_{RW} = D^{-1}A$$

While the algorithmic results from spectral clustering can be depend on the choice of graph Laplacian, `LocalTrans` returns exactly the same results with L as it does with L_{RW}. To see this, first imagine that if the graph is directed, then A is asymmetric, and for T to correspond to *directed* cycles of length three, it is necessary to take the transpose of the final A, that is $[AA] \cdot A^T$. Since L_{RW} is asymmetric, it is reasonable to use the additional transpose from the directed formulation. It is easy to show that

$$[LL] \cdot L = [L_{RW} L_{RW}] \cdot L_{RW}^T. \tag{6.10}$$

Chaudhuri et al. (2012) and Chen et al. (2012a) have recently proposed a "regularized" graph Laplacian. Chaudhuri et al. (2012) propose replacing D with $D_\tau = D + \tau I$, where $\tau > 0$ is a regularization constant. They show that a spectral algorithm with

$$L_\tau = D_\tau^{-1/2} A D_\tau^{-1/2}$$

has superior performance on sparse graphs. Similarly, it will help to use L_τ with `LocalTrans`. (Note that the equivalence in Equation (6.10) still holds with the regularized versions of the Laplacians.)

The next theorem shows that under the local degree-corrected model — with the regularized graph Laplacian, a specified choice of tuning parameter *cut*, and $i \in S_*$ — the estimate $\texttt{LocalTrans}(L_\tau, i, cut) = S_*$ with high probability. Importantly, using L_τ instead of A allows for reasonable results under the degree-corrected model.

Theorem 6.13. *Let A come from the local degree-corrected Stochastic Blockmodel. Define λ such that*

$$\sum_{i,j \in S_*^c} A_{ij} \leq n\lambda. \tag{6.11}$$

Set $cut = [2(s-1)p_{in} + 2\lambda + \tau]^{-3}$. If

$$n \geq 3\left(2(s-1)p_{in} + 2\lambda + \tau\right)^{3/\epsilon} \tau^{-1/\epsilon},$$

and $s \geq 3$*, then for any* $i \in S_*$*,*

$$\texttt{LocalTrans}(L_\tau, i, cut) = S_*$$

with probability at least

$$1 - \left(1/2\ s^2(1-p_{in}^2)^{s-2} + s\exp\left(-1/4\ sp_{in} + \lambda\right) + O(n^{3\epsilon-1})\right).$$

A proof of Theorem 6.13 can be found in the Appendix.

Because simple summary statistics (of sparsity and transitivity) on empirical networks contradict the types of models studied in the literature, Theorem 6.13 tries to minimizes the assumptions on the "global" structure of the graph. It only assumes that the graph outside of S_*, i.e. the induced subgraph on S_*^c, is sparse. There are no other assumptions on this part of the graph.

This result is asymptotic in $n = |S_*^c|$, with S_* fixed and containing nodes with bounded expected degree; the assumption in Equation 6.11 and the definition of d_j^* imply that the nodes in S_* have expected degree less than $s + \lambda$.

Preliminary Data Analysis

Recall that Figure 6.3 illustrates how $\texttt{GlobalTrans}(A, cut)$ fails to find any clusters larger than twenty nodes in the slashdot social network. Figure 6.4 shows that using L_τ instead of A corrects for the problems observed in Figure 6.3. It plots the size of the largest ten clusters in the slashdot social network found by $\texttt{GlobalTrans}(L_{12}, cut)$ for values of *cut* between $3 * 10^{-6}$ and $500 * 10^{-6}$. It finds several clusters that exceed twenty nodes. In this analysis, and all other analyses using L_τ, the regularization constant τ is set equal to the average node degree (as suggested in Chaudhuri et al. (2012)). In this case, $\tau \approx 12$.

Figure 6.5 shows some of the clusters from the slashdot social network. Specifically, it plots twenty-four of the induced subgraphs from $\texttt{GlobalTrans}(L_{\tau=12}, cut = 32 * 10^{-6})$. Because the clusters are not so large, the sub-graphs are easily visualized and it is easy to see how these clusters have several different structures. Some are nearly planar; others appear as densely connected, "clique-like" sub-graphs; other clusters are a collection of several smaller clusters, weakly strung together. Figure 6.1 in the introduction gives a similar plot for the epinions social network. These visualizations were created using the graph visualization tool in the igraph package in R (Csardi and Nepusz, 2006).

Figure 6.6 illustrates how $\texttt{LocalTrans}(L_{\tau=12}, i, cut)$ changes as a function of *cut* for a certain node in the epinions social network. Each of the four panels displays the

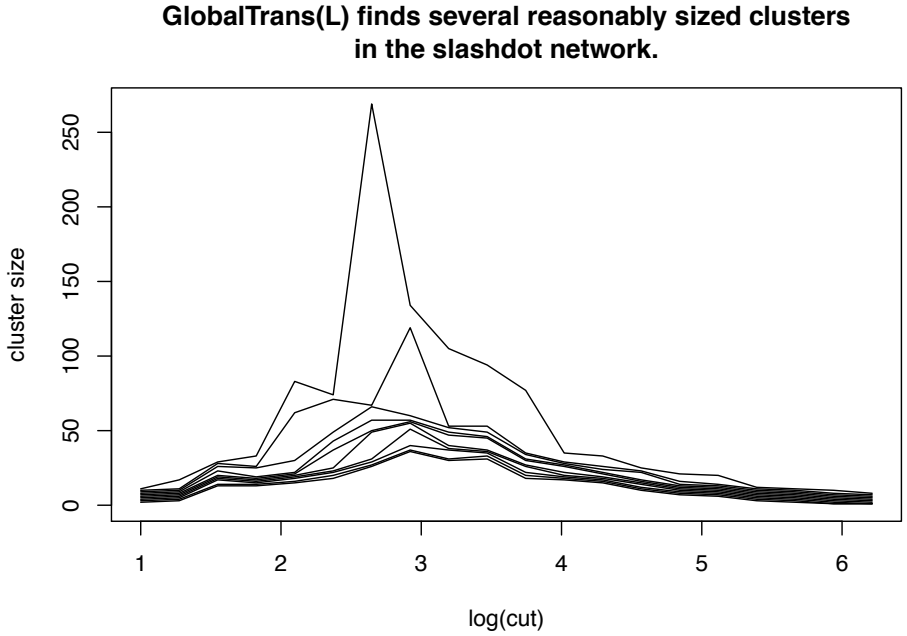

Figure 6.4: A plot of the number of nodes in the largest ten clusters (ignoring one very large cluster) found by `GlobalTrans`(L_τ, cut) in the slashdot social network. `GlobalTrans` with L_τ instead of A finds much larger clusters.

network for $cut = 58 * 10^{-6}$. In each of the four panels, solid nodes are the nodes that are included in `LocalTrans`$(L_{\tau=12}, i, cut)$ for four different values of *cut*. This seed node was selected because the local cluster is slowly growing as *cut* decreases and you can see in this in Figure 6.6.[5] The left most panel displays the results for the largest value of *cut*. This returns the smallest cluster and not surprisingly, the igraph package plots these nodes in the center of the larger graph. Moving to the right, the clusters grow larger and the additional nodes start to extend to the periphery of the visualization. While the clusters for this node grow slowly, for many other nodes, the transitions are abrupt. For example, the nodes that join the cluster in the last panel in Figure 6.6 jump from cluster sizes of one or two into this bigger cluster. Then, decreasing *cut* a little bit more, this cluster becomes part of a giant component.

[5]In particular, it was chosen as the "slowest growing" from a randomly chosen set of 200 nodes.

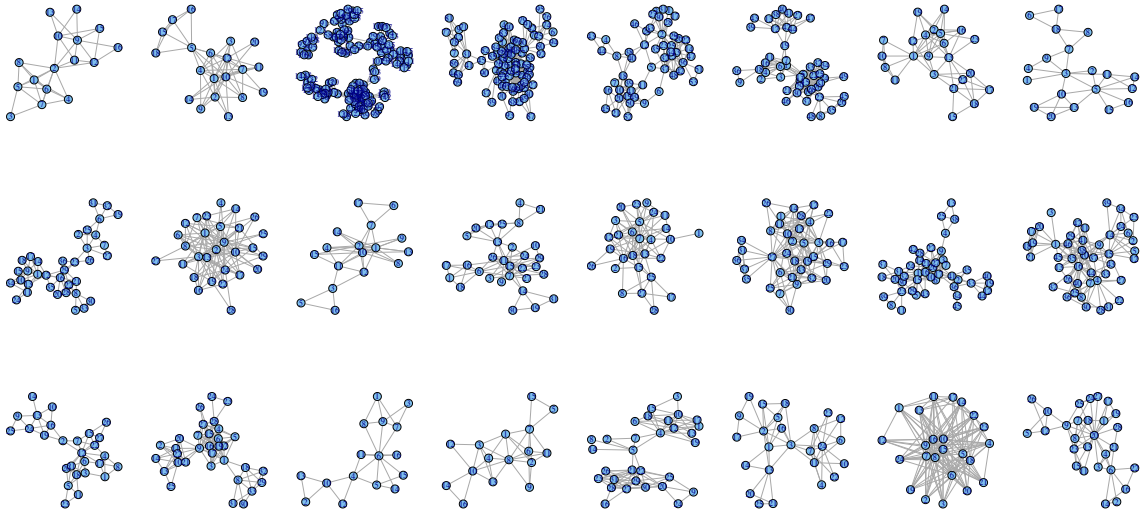

Figure 6.5: Twenty-four small clusters from the slashdot data set. Because `GlobalTrans` discovers small clusters, one can easily plot and visualize the clusters with a standard graph visualization tool (Csardi and Nepusz, 2006). The point of this figure is to show the variability in cluster structures; some are tight, clique-like clusters; others are small lattice-like clusters; others are "stringy" collections of three or four tight clusters. This highlights the ease of visualizing the results of local clustering.

6.5 Discussion

The tension between transitivity and sparsity in networks that implies that there are local regions of the graph that are dense and transitive. This leads to the blessing of dimensionality, which says that edges (in sparse and transitive graphs) become asymptotically more informative. For example, under the exchangeable model, if the model is sparse and transitive, then the conditional density of the latent variables ξ_i, ξ_j, given $A_{ij} = 1$, is asymptotically unbounded, concentrating on the values of ξ_i, ξ_j that are consistent with the local structure in the model. This has important implications for statistical models, methods, and estimation theory.

In sparse and non-transitive Stochastic Blockmodels, the block structure is *not* revealed in the local structure of the network. Rather, the blocks are revealed by comparing the edge density of various partitions. However, under transitive models, the local structure of the network can reveal the block structure. As such, these

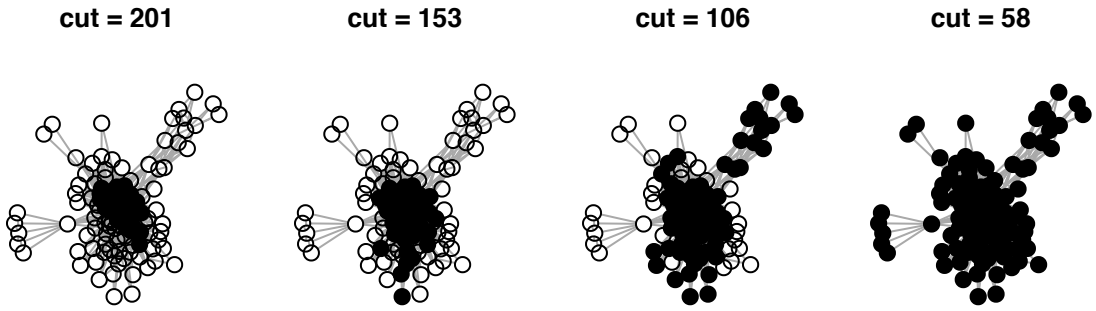

Figure 6.6: Starting from a seed node, this figure demonstrates how `LocalTrans`$(L_{\tau=12}, i, cut)$ grows as *cut* decreases. In each panel, the graph is drawn for the smallest value of *cut*, and the solid nodes correspond to the nodes returned by `LocalTrans`$(L_{\tau=12}, i, cut)$, where the value of *cut* is given above the graph in the units 10^{-6}. Moving from left to right, the clusters grow larger, and the additional nodes start to extend to the periphery of the visualization.

blocks can be estimated by fast local algorithms. Theorems 6.11 and 6.13 show that `LocalTrans` performs well under a *local* Stochastic Blockmodel that makes minimal assumptions on the nodes outside of the true cluster; this is the first statistical result to demonstrate how local clustering algorithms can be robust to vast regions of the graph.

This paper studies small clusters because (1) they can create sparse *and* transitive Stochastic Blockmodels, (2) they are relatively easy to find, both computationally and statistically, and (3) they are easy to plot and visualize. In future research, we will study how these ideas can be used to find large partitions in networks. Sparse and transitive models do not preclude large partitions, as long as some type of local structure exists within each partition. It is not yet clear how global algorithms like spectral clustering might leverage this transitive structure in a stochastic model; this is one area for future research.

Appendix A

Appendix for Chapter 2

Proof of Lemma 3.3

Proof. Recall that $\mathscr{D}_{ii} = \theta_i[D_B]_{z_i}$ and $[\Theta_\tau]_{ii} = \theta_i \frac{\mathscr{D}_{ii}}{\mathscr{D}_{ii}+\tau}$. The ij'th element of $\mathscr{L}_\tau$:

$$[\mathscr{L}_\tau]_{ij} = \frac{\mathscr{A}_{ij}}{\sqrt{(\mathscr{D}_{ii}+\tau)(\mathscr{D}_{jj}+\tau)}} = \frac{\theta_i\theta_j B_{z_iz_j}}{\sqrt{\mathscr{D}_{ii}\mathscr{D}_{jj}}}\sqrt{\frac{\mathscr{D}_{ii}}{\mathscr{D}_{ii}+\tau}\frac{\mathscr{D}_{jj}}{\mathscr{D}_{jj}+\tau}} = \frac{B_{z_iz_j}}{\sqrt{[D_B]_{z_i}[D_B]_{z_j}}} * \sqrt{[\Theta_\tau]_{ii}[\Theta_\tau]_{jj}}.$$

Hence,

$$\mathscr{L}_\tau = \Theta_\tau^{\frac{1}{2}} Z B_L Z^T \Theta_\tau^{\frac{1}{2}}.$$

□

Proof of Lemma 2.3

Proof. Let $C = (Z^T\Theta_\tau Z)^{1/2} B_L (Z^T\Theta_\tau Z)^{1/2}$. If $\theta_i > 0, i = 1, ..., N$, then $C \succ 0$ since $B \succ 0$ by assumption. Let $\lambda_1 \geq ... \geq \lambda_K > 0$ be the eigenvalues of C. Let $\Lambda \in \mathcal{R}^{K\times K}$ be a diagonal matrix with its ss'th element to be λ_s. Let $U \in \mathcal{R}^{K\times K}$ be an orthogonal matrix where its s'th column is the eigenvector of C corresponding $\lambda_s, s = 1, ..., K$. By eigen-decomposition, we have $C = U\Lambda U^T$. Define $\mathscr{X}_\tau = \Theta_\tau^{\frac{1}{2}} Z (Z^T\Theta_\tau Z)^{-1/2} U$, then

$$\mathscr{X}_\tau^T \mathscr{X}_\tau = U^T (Z^T\Theta_\tau Z)^{-1/2} (Z^T\Theta_\tau Z)(Z^T\Theta_\tau Z)^{-1/2} U = U^T U = I.$$

On the other hand,

$$\mathscr{X}_\tau \Lambda \mathscr{X}_\tau^T = \Theta_\tau^{\frac{1}{2}} Z (Z^T\Theta_\tau Z)^{-1/2} C (Z^T\Theta_\tau Z)^{-1/2} Z^T \Theta_\tau^{\frac{1}{2}} = \Theta_\tau^{\frac{1}{2}} Z B_L Z^T \Theta_\tau^{\frac{1}{2}} = \mathscr{L}_\tau.$$

Hence, $\lambda_s, s = 1, ..., K$ are $\mathscr{L}_\tau$'s positive eigenvalues and $\mathscr{X}_\tau$ contains $\mathscr{L}_\tau$'s eigenvectors corresponding to its nonzero eigenvalues. For part 2, notice that $||\mathscr{X}_\tau^i||_2 = (\frac{[\Theta_\tau]_{ii}}{[Z^T\Theta_\tau Z]_{z_i z_i}})^{1/2}$, then

$$[\mathscr{X}_\tau^*]^i = \frac{\mathscr{X}_\tau^i}{||\mathscr{X}_\tau^i||_2} = \frac{(\frac{[\Theta_\tau]_{ii}}{[Z^T\Theta_\tau Z]_{z_i z_i}})^{1/2} Z_i U}{||\mathscr{X}_\tau^i||_2} = Z_i U.$$

Therefore, $\mathscr{X}_\tau^* = ZU$. □

Proof of Theorem 2.4

Proof. We extend the proof of Theorem 2 in Chung and Radcliffe (2011) to the case of regularized graph laplacian. Let $H = \mathscr{D}_\tau^{-1/2} A \mathscr{D}_\tau^{-1/2}$. Then $||L_\tau - \mathscr{L}_\tau|| \leq ||H - \mathscr{L}_\tau|| + ||L_\tau - H||$. We bound the two terms separately.

For the first term, we apply the concentration inequality for matrix:

Lemma A.1. *Let $X_1, X_2, ..., X_m$ be independent random $N \times N$ Hermitian matrices. Moreover, assunme that $||X_i - \mathrm{E}(X_i)|| \leq M$ for all i, and put $v^2 = ||\sum var(X_i)||$. Let $X = \sum X_i$. Then for any $a > 0$,*

$$pr(||X - \mathrm{E}(X)|| \geq a) \leq 2N \exp\big(-\frac{a^2}{2v^2 + 2Ma/3}\big).$$

Notice that $||H - \mathscr{L}_\tau|| = \mathscr{D}_\tau^{-1/2}(A - \mathscr{A})\mathscr{D}_\tau^{-1/2}$. Let $E^{ij} \in \mathcal{R}^{N \times N}$ be the matrix with 1 in the ij and ji'th positions and 0 everywhere else. Let

$$\begin{aligned} X_{ij} &= \mathscr{D}_\tau^{-1/2}((A_{ij} - p_{ij})E^{ij})\mathscr{D}_\tau^{-1/2} \\ &= \frac{A_{ij} - p_{ij}}{\sqrt{(\mathscr{D}_{ii} + \tau)(\mathscr{D}_{jj} + \tau)}} E^{ij}. \end{aligned}$$

$H - \mathscr{L}_\tau = \sum X_{ij}$. Then we can apply the matrix concentration theorem on $\{X_{ij}\}$. By similar argument as in Chung and Radcliffe (2011), we have

$$||X_{ij}|| \leq [(\mathscr{D}_{ii} + \tau)(\mathscr{D}_{jj} + \tau)]^{-1/2} \leq \frac{1}{\delta + \tau}, \qquad v^2 = ||\sum E(X_{ij}^2)|| \leq \frac{1}{\delta + \tau}.$$

Take $a = \sqrt{\frac{3\ln(4N/\epsilon)}{\delta + \tau}}$. By assumption $\delta + \tau > 3\ln N + 3\ln(4/\epsilon)$, it implies $a < 1$.

Applying Lemma B.2, we have

$$\begin{aligned} pr(||H-\mathscr{L}_\tau|| \geq a) &\leq 2N \exp\left(- \frac{\frac{3\ln(4N/\epsilon)}{\delta+\tau}}{2/(\delta+\tau)+2a/[3(\delta+\tau)]}\right) \\ &\leq 2N\exp(-\frac{3\ln(4N/\epsilon)}{3}) \\ &\leq \epsilon/2. \end{aligned}$$

For the second term, first we apply the two sided concentration inequality for each i, (see for example Chung and Lu (2006, chap. 2))

$$pr(|D_{ii}-\mathscr{D}_{ii}| \geq \lambda) \leq \exp\{-\frac{\lambda^2}{2\mathscr{D}_{ii}}\} + \exp\{-\frac{\lambda^2}{2\mathscr{D}_{ii}+\frac{2}{3}\lambda}\}$$

Let $\lambda = a(\mathscr{D}_{ii}+\tau)$, where a is the same as in the first part.

$$\begin{aligned} pr(|D_{ii}-\mathscr{D}_{ii}| \geq a(\mathscr{D}_{ii}+\tau)) &\leq \exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)^2}{2\mathscr{D}_{ii}}\} + \exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)^2}{2\mathscr{D}_{ii}+\frac{2}{3}a(\mathscr{D}_{ii}+\tau)}\} \\ &\leq 2\exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)^2}{(2+\frac{2}{3}a)(\mathscr{D}_{ii}+\tau)}\} \\ &\leq 2\exp\{-\frac{a^2(\mathscr{D}_{ii}+\tau)}{3}\} \\ &\leq 2\exp\{-\ln(4N/\epsilon)\frac{(\mathscr{D}_{ii}+\tau)}{\delta+\tau}\} \\ &\leq 2\exp\{-\ln(4N/\epsilon)\} \\ &\leq \epsilon/2N. \end{aligned}$$

$$||\mathscr{D}_\tau^{-1/2}D_\tau^{1/2}-I|| = max_i|\sqrt{\frac{D_{ii}+\tau}{\mathscr{D}_{ii}+\tau}}-1| \leq max_i|\frac{D_{ii}+\tau}{\mathscr{D}_{ii}+\tau}-1|.$$

$$\begin{aligned} pr(||\mathscr{D}_\tau^{-1/2}D_\tau^{1/2}-I|| \geq a) &\leq pr(max_i|\frac{D_{ii}+\tau}{\mathscr{D}_{ii}+\tau}-1| \geq a) \\ &\leq pr(\cup_i\{|(D_{ii}+\tau)-(\mathscr{D}_{ii}+\tau)| \geq b(\mathscr{D}_{ii}+\tau)\}) \\ &\leq \epsilon/2. \end{aligned}$$

Note that $||L_\tau|| \leq 1$, therefore, with probability at least $1 - \epsilon/2$, we have

$$\begin{aligned}
||L_\tau - H|| &= ||D_\tau^{-1/2} A D_\tau^{-1/2} - \mathscr{D}_\tau^{-1/2} A \mathscr{D}_\tau^{-1/2}|| \\
&= ||L_\tau - \mathscr{D}_\tau^{-1/2} D_\tau^{1/2} L_\tau D_\tau^{1/2} \mathscr{D}_\tau^{-1/2}|| \\
&= ||(I - \mathscr{D}_\tau^{-1/2} D_\tau^{1/2}) L_\tau D_\tau^{1/2} \mathscr{D}_\tau^{-1/2} + L_\tau (I - D_\tau^{1/2} \mathscr{D}_\tau^{-1/2})|| \\
&\leq ||\mathscr{D}_\tau^{-1/2} D_\tau^{1/2} - I|| ||\mathscr{D}_\tau^{-1/2} D_\tau^{1/2}|| + ||\mathscr{D}_\tau^{-1/2} D_\tau^{1/2} - I|| \\
\leq a^2 + 2a.
\end{aligned}$$

Combining the two part, we have that with probability at least $1 - \epsilon$,

$$||L_\tau - \mathscr{L}_\tau|| \leq a^2 + 3a \leq 4a,$$

where $a = \sqrt{\frac{3 \ln(4N/\epsilon)}{\delta + \tau}}$. □

Proof of Theorem 2.5

Proof. First we apply a lemma from McSherry (2001):

Lemma A.2. *For any matrix A, let P_A denotes the projection onto the span of A's first K left sigular vectors. Then $P_A A$ is the optimal rank K approximation to A in the following sense. For any rank K matrix X, $||A - P_A A|| \leq ||L - X||$. Further, for any rank K matrix B,*

$$||P_A A - B||_F^2 \leq 8K ||A - B||^2. \tag{A.1}$$

Let $W \in \mathcal{R}^{K \times K}$ be a diagonal matrix that contains the K largest eigenvalues of L_τ, $w_1 \geq w_2 \geq ... \geq w_K$. Let $\Lambda \in R^{K \times K}$ be the diagonal matrix that contains all positive eigenvalues of $\mathscr{L}_\tau$. Take $A = L_\tau$ and $B = \mathscr{L}_\tau$ in Lemma A.2. then $P_{L_\tau} L_\tau = X_\tau W X_\tau^T$ and the previous inequality can be rewritten as

$$||P_{L_\tau} L_\tau - \mathscr{L}_\tau||_F^2 = ||X_\tau W X_\tau^T - \mathscr{X}_\tau \Lambda \mathscr{X}_\tau^T||_F^2 \leq 8K ||L_\tau - \mathscr{L}_\tau||^2.$$

Then we apply a modified version of the Davis-Kahan theorem (Rohe et al. (2011)) to $\mathscr{L}_\tau$.

Proposition A.3. *Let $S \subset \mathcal{R}$ be an interval. Denote $\mathscr{X}_\tau$ as an orthonormal matrix whose column space is equal to the eigenspace of $\mathscr{L}_\tau$ corresponding to the eigenvalues in $\lambda_S(\mathscr{L}_\tau)$ (more formally, the column space of $\mathscr{X}_\tau$ is the image of the spectral projection of $\mathscr{L}_\tau$ induced by $\lambda_S(\mathscr{L}_\tau)$). Denote by X_τ the analogous quantity for $P_{L_\tau} L_\tau$. Define*

the distance between S and the spectrum of $\mathscr{L}_\tau$ outside of S as

$$\Delta = \min\{|\lambda - s|; \lambda \text{ eigenvalue of } \mathscr{L}_\tau,\ \lambda \notin S,\ s \in S\}.$$

if $\mathscr{X}_\tau$ and X_τ are of the same dimension, then there is an orthogonal matrix $\mathscr{O}$, that depends on $\mathscr{X}_\tau$ and X_τ, such that

$$||X_\tau - \mathscr{X}_\tau \mathscr{O}||_F^2 \leq \frac{2||P_{L_\tau} L_\tau - \mathscr{L}_\tau||_F^2}{\Delta^2}.$$

Take $S = (\lambda_K/2, 2)$, then $\Delta = \lambda_K/2$. By assumption (a) $\sqrt{\frac{K \ln(4N/\epsilon)}{\delta+\tau}} \leq \frac{1}{8\sqrt{3}}\lambda_K$, we have that when N is sufficiently large, with probability at least $1-\epsilon$,

$$|\lambda_K - w_K| \leq ||L_\tau - \mathscr{L}_\tau|| \leq 4\sqrt{\frac{3\ln(4N/\epsilon)}{\delta+\tau}} \leq \lambda_K/2.$$

Hence $w_K \in S$. X and $\mathscr{X}$ are of the same dimension.

$$\begin{aligned} ||X_\tau - \mathscr{X}_\tau \mathscr{O}||_F &\leq \frac{\sqrt{2}||P_{L_\tau} L_\tau - \mathscr{L}_\tau||_F}{\Delta} \leq \frac{2\sqrt{2}||P_{L_\tau} L_\tau - \mathscr{L}_\tau||_F}{\lambda_K} \\ &\leq \frac{8\sqrt{K}||L_\tau - \mathscr{L}_\tau||}{\lambda_K} \\ &\leq \frac{C}{\lambda_K}\sqrt{\frac{K\ln(4N/\epsilon)}{\delta+\tau}}. \end{aligned}$$

holds for $C = 32\sqrt{3}$ with probability at least $1-\epsilon$.

For part 2, note that for any i,

$$||[X_\tau^*]^i - [\mathscr{X}_\tau^*]^i \mathscr{O}||_2 \leq 2\frac{||X_\tau^i - \mathscr{X}_\tau^i \mathscr{O}||_2}{max\{||X_\tau^i||_2, ||\mathscr{X}_\tau^i||_2\}},$$

We have that

$$||X_\tau^* - \mathscr{X}_\tau^* \mathscr{O}||_F \leq \frac{||X_\tau - \mathscr{X}_\tau \mathscr{O}||_F}{m},$$

where $m = min_i\{||\mathscr{X}_\tau^i||_2\}$. □

Proof of Theorem 2.7

Proof. Recall that the set of misclustered nodes is defined as:

$$\mathscr{M} = \{i : \exists j \neq i, s.t. ||C_i\mathscr{O}^T - \mathcal{C}_i||_2 > ||C_i\mathscr{O}^T - \mathcal{C}_j||_2\}.$$

Note that Lemma 3.3 implies that the population centroid corresponding to i'th row of $\mathscr{X}_\tau^*$

$$\mathcal{C}_i = Z_i U.$$

Since all population centroids are of unit length and are orthogonal to each other, a simple calculation gives a sufficient condition for one observed centroid to be closest to the population centroid:

$$||C_i\mathscr{O}^T - \mathcal{C}_i||_2 < 1/\sqrt{2} \Rightarrow ||C_i\mathscr{O}^T - \mathcal{C}_i||_2 < ||C_i\mathscr{O}^T - \mathcal{C}_j||_2 \quad \forall Z_j \neq Z_i.$$

Define the following set of nodes that do not satisfy the sufficient condition,

$$\mathscr{U} = \{i : ||C_i\mathscr{O}^T - \mathcal{C}_i||_2 \geq 1/\sqrt{2}\}.$$

The mis-clustered nodes $\mathscr{M} \in \mathscr{U}$.

Define $Q \in \mathcal{R}^{N\times K}$, where the i'th row of Q is C_i, the observed centroid of node i from k-means. By definition of k-means, we have

$$||X_\tau^* - Q||_2 \leq ||X_\tau^* - \mathscr{X}_\tau^*\mathscr{O}||_2.$$

By triangle inequality,

$$||Q - ZU\mathscr{O}||_2 = ||Q - \mathscr{X}_\tau^*\mathscr{O}||_2 \leq ||X_\tau^* - Q||_2 + ||X_\tau^* - \mathscr{X}_\tau^*\mathscr{O}||_2 \leq 2||X_\tau^* - \mathscr{X}_\tau^*\mathscr{O}||_2.$$

We have with probability at least $1-\epsilon$,

$$\begin{aligned}
\frac{|\mathscr{M}|}{N} &\leq \frac{|\mathscr{U}|}{N} = \frac{1}{N}\sum_{i\in\mathscr{U}} 1 \\
&\leq \frac{2}{N}\sum_{i\in\mathscr{U}} ||C_i\mathscr{O}^T - \mathcal{C}_i||_2^2 \\
&= \frac{2}{N}\sum_{i\in\mathscr{U}} ||C_i - Z_iU\mathscr{O}||_2^2 \\
&\leq \frac{2}{N}||Q - ZU\mathscr{O}||_F^2 \\
&\leq \frac{8}{N}||X_\tau^* - \mathscr{X}_\tau^*\mathscr{O}||_F^2 \\
&\leq c_1\frac{K\ln(N/\epsilon)}{Nm^2(\delta+\tau)\lambda_K^2}.
\end{aligned}$$

□

Appendix B

Appendix for Chapter 3

B.1 Convergence of Singular Vectors

The classical spectral clustering algorithm above can be divided into two steps: (1) find the eigendecomposition of L and (2) run k-means. Several previous papers have studied the estimation performance of the classical spectral clustering algorithm under a standard social network model. However, due to the asymmetry of A, previous proof techniques can not be directly applied to study the singular vectors for DI-SIM. In this analysis, we (a) symmetrize the graph Laplacian, (b) apply modern matrix concentration techniques to this symmetrized version of the graph Laplacian, and (c) apply an updated version of the Davis-Kahn theorem to bound the distance between the singular spaces of the empirical and population Laplacian.

For simplicity, from now on let L denote the regularized graph Laplacian.

Define the symmetrized version of L and $\mathscr{L}$ as

$$\tilde{L} = \begin{pmatrix} 0 & L \\ L^T & 0 \end{pmatrix}, \qquad \tilde{\mathscr{L}} = \begin{pmatrix} 0 & \mathscr{L} \\ \mathscr{L}^T & 0 \end{pmatrix}.$$

The next theorem gives a sharp bound between $\tilde{L}$ and $\tilde{\mathscr{L}}$.

Theorem B.1. *(Concentration of L) Let G be a random graph, with independent edges and $pr(v_i \sim v_j) = p_{ij}$. Let δ be the minimum expected row and column degree of G, that is $\delta = \min(\min_i \mathscr{O}_{ii}, \min_j \mathscr{P}_{jj})$. For any $\epsilon > 0$, if $\delta + \tau > 3\ln(N_r + N_c) + 3\ln(4/\epsilon)$, then with probability at least $1 - \epsilon$,*

$$\|\tilde{L} - \tilde{\mathscr{L}}\| \leq 4\sqrt{\frac{3\ln(4(N_r + N_c)/\epsilon)}{\delta + \tau}}. \tag{B.1}$$

Proof. Let $C = \mathscr{P}_\tau^{-\frac{1}{2}} A \mathscr{O}_\tau^{-\frac{1}{2}}$ and define $\tilde{C}$ in the same way as $\tilde{L}$. Then $\|\tilde{L} - \tilde{\mathscr{L}}\| \leq \|\tilde{C} - \tilde{\mathscr{L}}\| + \|\tilde{L} - \tilde{C}\|$. We bound the two terms separately.

For the first term, we apply the following concentration inequality for matrices, see for example Chung and Radcliffe (2011).

Lemma B.2. *Let $X_1, X_2, ..., X_m$ be independent random $N \times N$ Hermitian matrices. Moreover, assume that $\|X_i - \mathrm{E}(X_i)\| \leq M$ for all i, and $v^2 = \|\sum var(X_i)\|$. Let $X = \sum X_i$. Then for any $a > 0$,*

$$pr(\|X - \mathrm{E}(X)\| \geq a) \leq 2N \exp\left(- \frac{a^2}{2v^2 + 2Ma/3} \right).$$

Let E^{ij} be the matrix with 1 in the i, j and j, i positions and 0 everywhere else. Let $p_{ij} = \mathscr{A}_{ij}$. To use this inequality, express $\tilde{C} - \tilde{\mathscr{L}}$ as the sum of the matrices $Y_{i,m+j}$,

$$Y_{i,m+j} = \frac{1}{\sqrt{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}} (A_{ij} - p_{ij}) E^{i,m+j}, i = 1, ..., m, j = 1, ..., n.$$

Note that

$$\|\tilde{C} - \tilde{\mathscr{L}}\| = \| \sum_{i=1}^{m} \sum_{j=1}^{n} Y_{i,m+j} \|,$$

and

$$\|Y_{i,m+j}\| \leq \frac{1}{\sqrt{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)}} \leq (\delta + \tau)^{-1}.$$

Moreover,

$$\mathrm{E}[Y_{i,m+j}] = 0 \quad \textit{and} \quad \mathrm{E}[Y_{i,m+j}^2] = \frac{1}{(\mathscr{O}_{ii} + \tau)(\mathscr{P}_{jj} + \tau)} (p_{ij} - p_{ij}^2)(E^{ii} + E^{m+j,m+j}).$$

Then,

$$
\begin{aligned}
v^2 &= \| \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{E}[Y_{i,m+j}^2] \| = \| \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}(p_{ij}-p_{ij}^2)(E^{ii}+E^{m+j,m+j}) \| \\
&= \| \sum_{i=1}^{m}[\sum_{j=1}^{n} \frac{1}{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}(p_{ij}-p_{ij}^2)]E^{ii} + \sum_{j=1}^{n}[\sum_{i=1}^{m} \frac{1}{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}(p_{ij}-p_{ij}^2)]E^{m+j,m+j} \| \\
&= \max\left\{ \max_{i=1,\ldots,m}(\sum_{j=1}^{n} \frac{1}{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}(p_{ij}-p_{ij}^2)), \max_{j=1,\ldots,n}(\sum_{i=1}^{m} \frac{1}{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}(p_{ij}-p_{ij}^2)) \right\} \\
&\le \max\left\{ \max_{i=1,\ldots,m} \frac{1}{\delta+\tau}\sum_{j=1}^{n} \frac{p_{ij}}{\mathscr{O}_{ii}+\tau}, \max_{j=1,\ldots,n} \frac{1}{\delta+\tau}\sum_{i=1}^{m} \frac{p_{ij}}{\mathscr{P}_{jj}+\tau} \right\} \\
&= (\delta+\tau)^{-1}.
\end{aligned}
$$

Take

$$
a = \sqrt{\frac{3\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}}.
$$

By assumption, $\delta+\tau > 3\ln(N_r+N_c)+3\ln(4/\epsilon)$. So $a<1$. Applying Lemma B.2,

$$
\begin{aligned}
pr(\|\tilde{C}-\tilde{\mathscr{L}}\| \ge a) &\le 2(N_r+N_c)\exp\left(-\frac{\frac{3\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}}{2/(\delta+\tau)+2a/[3(\delta+\tau)]} \right) \\
&\le 2N\exp(-\frac{3\ln(4(N_r+N_c)/\epsilon)}{3}) \\
&\le \epsilon/2.
\end{aligned}
$$

For the second term $\|\tilde{L}-\tilde{C}\|$, define

$$
D_\tau = \begin{pmatrix} O_\tau & 0 \\ 0 & P_\tau \end{pmatrix}, \qquad \mathscr{D}_\tau = \begin{pmatrix} \mathscr{O}_\tau & 0 \\ 0 & \mathscr{P}_\tau \end{pmatrix}, \quad D = D_0, \text{ and } \mathscr{D} = \mathscr{D}_0.
$$

Apply the two sided concentration inequality for each i, $1 \le i \le N_r+N_c$, (see for example Chung and Lu (2006, chap. 2))

$$
pr(|D_{ii}-\mathscr{D}_{ii}| \ge \lambda) \le \exp\{-\frac{\lambda^2}{2\mathscr{D}_{ii}}\} + \exp\{-\frac{\lambda^2}{2\mathscr{D}_{ii}+\frac{2}{3}\lambda}\}.
$$

Let $\lambda = a(\mathscr{D}_{ii} + \tau)$, where a is as before.

$$\begin{aligned} pr\bigg(|D_{ii} - \mathscr{D}_{ii}| \geq a(\mathscr{D}_{ii} + \tau)\bigg) &\leq \exp\{-\frac{a^2(\mathscr{D}_{ii} + \tau)^2}{2\mathscr{D}_{ii}}\} + \exp\{-\frac{a^2(\mathscr{D}_{ii} + \tau)^2}{2\mathscr{D}_{ii} + \frac{2}{3}a(\mathscr{D}_{ii} + \tau)}\} \\ &\leq 2\exp\{-\frac{a^2(\mathscr{D}_{ii} + \tau)^2}{(2 + \frac{2}{3}a)(\mathscr{D}_{ii} + \tau)}\} \\ &\leq 2\exp\{-\frac{a^2(\mathscr{D}_{ii} + \tau)}{3}\} \\ &\leq 2\exp\{-\ln(4(N_r + N_c)/\epsilon)\frac{(\mathscr{D}_{ii} + \tau)}{\delta + \tau}\} \\ &\leq 2\exp\{-\ln(4(N_r + N_c)/\epsilon)\} \\ &\leq \epsilon/2(N_r + N_c). \end{aligned}$$

Because

$$\|\mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}} - I\| = max_i\left|\sqrt{\frac{D_{ii} + \tau}{\mathscr{D}_{ii} + \tau}} - 1\right| \leq max_i\left|\frac{D_{ii} + \tau}{\mathscr{D}_{ii} + \tau} - 1\right|,$$

It follows that

$$\begin{aligned} pr(\|\mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}} - I\| \geq a) &\leq pr(max_i\left|\frac{D_{ii} + \tau}{\mathscr{D}_{ii} + \tau} - 1\right| \geq a) \\ &\leq pr(\cup_i\{|(D_{ii} + \tau) - (\mathscr{D}_{ii} + \tau)| \geq a(\mathscr{D}_{ii} + \tau)\}) \\ &\leq \epsilon/2. \end{aligned}$$

Note that $\|\tilde{L}_\tau\| \leq 1$. Therefore, with probability at least $1 - \epsilon/2$,

$$\begin{aligned} \|\tilde{L}_\tau - C\| &= \|D_\tau^{-\frac{1}{2}} \tilde{A} D_\tau^{-\frac{1}{2}} - \mathscr{D}_\tau^{-\frac{1}{2}} \tilde{A} \mathscr{D}_\tau^{-\frac{1}{2}}\| \\ &= \|\tilde{L}_\tau - \mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}} \tilde{L}_\tau D_\tau^{\frac{1}{2}} \mathscr{D}_\tau^{-\frac{1}{2}}\| \\ &= \|(I - \mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}})\tilde{L}_\tau D_\tau^{\frac{1}{2}} \mathscr{D}_\tau^{-\frac{1}{2}} + \tilde{L}_\tau(I - D_\tau^{\frac{1}{2}} \mathscr{D}_\tau^{-\frac{1}{2}})\| \\ &\leq \|\mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}} - I\|\|\mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}}\| + \|\mathscr{D}_\tau^{-\frac{1}{2}} D_\tau^{\frac{1}{2}} - I\| \\ \leq a^2 + 2a. \end{aligned}$$

Combining the two parts yields

$$\|\tilde{L}_\tau - \tilde{\mathscr{L}_\tau}\| \leq a^2 + 3a \leq 4a,$$

with probability at least $1-\epsilon$. □

The next theorem bounds the difference between the empirical and population singular vectors in terms of the Frobenius norm.

Theorem B.3. *(Concentration of Singular Space) Let A be the adjacency matrix generated from the DC-ScBM with parameters $\{\boldsymbol{B}, Y, Z, \Theta_Y, \Theta_Z\}$. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$ be the positive singular values of $\mathscr{L}_\tau$.*

Let $X_L(X_R)$ and $\mathscr{X}_L(\mathscr{X}_R)$ contain the top K left(right) singular vectors of L_τ and $\mathscr{L}_\tau$ respectively. For any $\epsilon > 0$ and sufficiently large N_r and N_r, if $\delta > 3\ln(N_r + N_c) + 3\ln(4/\epsilon)$, then with probability at least $1-\epsilon$

$$\|X_L - \mathscr{X}_L\mathscr{R}_L\|_F \leq \frac{8\sqrt{6}}{\lambda_K}\sqrt{\frac{K\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}} \tag{B.2}$$

$$\text{and } \|X_R - \mathscr{X}_R\mathscr{R}_R\|_F \leq \frac{8\sqrt{6}}{\lambda_K}\sqrt{\frac{K\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}}, \tag{B.3}$$

for some orthogonal matrices $\mathscr{R}_L, \mathscr{R}_R \in \mathbb{R}^{K\times K}$.

Proof. Define

$$\mathscr{X}^{\tilde{}} = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathscr{X}_L \\ \mathscr{X}_R \end{pmatrix}.$$

A simple calculation shows that $\mathscr{X}^{\tilde{}} \in \mathbb{R}^{(N_r+N_c)\times K}$ contains the top K eigenvectors of $\tilde{L}$ corresponding to its top K eigenvalues.

We apply an improved version of Davis Kahn theorem from **?**). By a slightly modified proof of Lemma 5.1 in **?**), it can be shown that

$$\|\tilde{X}\tilde{X}^T - \mathscr{X}^{\tilde{}}\,\mathscr{X}^{\tilde{}T}\|_F \leq \frac{\sqrt{2K}}{\lambda_K}\|\tilde{L}_\tau - \mathscr{L}^{\tilde{}}_\tau\|.$$

Combining it with Theorem B.1 and its assumptions,

$$\|\tilde{X}\tilde{X}^T - \mathscr{X}^{\tilde{}}\,\mathscr{X}^{\tilde{}T}\|_F \leq \frac{4\sqrt{6}}{\lambda_K}\sqrt{\frac{K\ln(4(N_r+N_c)/\epsilon)}{\delta+\tau}},$$

with probability at lease $1-\epsilon$. By definition of $\tilde{\mathscr{X}}$ and $\tilde{X}$,

$$\begin{aligned}\|\tilde{X}\tilde{X}^T - \tilde{\mathscr{X}}\tilde{\mathscr{X}}^T\|_F &= \left\| \begin{pmatrix} \frac{1}{2}(X_L X_L^T - \mathscr{X}_L \mathscr{X}_L^T) & \frac{1}{2}(X_L X_R^T - \mathscr{X}_L \mathscr{X}_R^T) \\ \frac{1}{2}(X_R X_L^T - \mathscr{X}_R \mathscr{X}_L^T) & \frac{1}{2}(X_R X_R^T - \mathscr{X}_R \mathscr{X}_R) \end{pmatrix} \right\|_F \\ &\geq \frac{1}{2}\|X_L X_L^T - \mathscr{X}_L \mathscr{X}_L^T\|_F \\ &\geq \frac{1}{2}\|X_L - \mathscr{X}_L \mathscr{R}_L\|_F.\end{aligned}$$

Similarly $\|\tilde{X}\tilde{X}^T - \tilde{\mathscr{X}}\tilde{\mathscr{X}}^T\|_F \geq \frac{1}{2}\|X_R - \mathscr{X}_R\mathscr{R}_R\|_F$. This proves the above theorem. □

B.2 Proof of Theorem 3.7

To rigorously discuss the asymptotic estimation properties of DI-SIM, the next subsections examine the behavior of DI-SIM applied to a population version of the graph Laplacian $\mathscr{L}$, and compare this to DI-SIM applied to the observed graph Laplacian L.

The population version of di-sim

This subsection shows that DI-SIM applied to $\mathscr{L}$ can perfectly identify the blocks in the Stochastic co-Blockmodel. Recall DI-SIM applied to L.

1. Find the left singular vectors $X_L \in \mathbb{R}^{N_r \times k_y}$.
2. Normalize each row of X_L to have unit length. Denote the normalized rows of X_L as $u_1, \ldots, u_{N_r} \in \mathbb{R}^{k_y}$ with and $\|u_i\|_2 = 1$.
3. Run $(1+\alpha)$-approximate k-means on $u_1, \ldots, u_{N_r}$ with k_y clusters.
4. Repeat steps (a), (b), and (c) for the the right singular vectors $X_R \in \mathbb{R}^{N_c \times k_y}$ with k_z clusters.

k-means clusters points $u_1, \ldots, u_n$ in Euclidean space by optimizing the following objective function (Steinhaus (1956)),

$$\min_{\{m_1,\ldots,m_{k_y}\} \subset \mathbb{R}^{k_y}} \sum_i \min_g \|u_i - m_g\|_2^2. \tag{B.4}$$

Define the *centroids* as the arguments $m_1^*, \ldots, m_{k_y}^*$ that optimize (B.4). Finding $m_1^*, \ldots, m_{k_y}^*$ is NP-hard. DI-SIM uses a linear time algorithm, $(1+\alpha)$-approximate

k-means (Kumar et al. (2004)). That is, the algorithm computes $\hat{m}_1, \ldots, \hat{m}_{k_y}$ such that

$$\sum_i \min_g \|u_i - \hat{m}_g\|_2^2 \leq (1+\alpha) \sum_i \min_g \|u_i - m_g^*\|_2^2.$$

To study DI-SIM applied to $\mathscr{L}$, Lemma 3.3 gives an explicit form as a function of the parameters of the DC-ScBM. Recall that $\mathscr{A} = E(A)$ and under the DC-ScBM,

$$\mathscr{A} = \Theta_y Y \mathbf{B} Z^T \Theta_z,$$

where $Y \in \{0,1\}^{N_r \times k_y}, Z \in \{0,1\}^{N_c \times k_z}$, and $B \in [0,1]^{k_y \times k_z}$. Assume that $k_y \leq k_z$, without loss of generality. Moreover, recall that the regularized population versions of O, P, and L are defined as

$$\begin{aligned} &\mathscr{P}_{jj} = \textstyle\sum_k \mathscr{A}_{kj} \\ &\mathscr{O}_{ii} = \textstyle\sum_k \mathscr{A}_{ik} \\ &\mathscr{O}_\tau = \mathscr{O} + \tau I, \qquad \mathscr{P}_\tau = \mathscr{P} + \tau I \\ &\mathscr{L} = \mathscr{O}_\tau^{-\frac{1}{2}} \mathscr{A} \mathscr{P}_\tau^{-\frac{1}{2}} \end{aligned} \tag{B.5}$$

where $\mathscr{O}_\tau$ and $\mathscr{P}_\tau$ are diagonal matrices.

The following proves Lemma 3.3.

Proof. Recall that $\mathscr{O}_{ii} = \theta_i^Y [P_B]_{y_i y_i}$ and $\mathscr{P}_{jj} = \theta_j^Z [O_B]_{z_j z_j}$. In addition,

$$[\Theta_{Y,\tau}]_{ii} = \theta_i^Y \frac{\mathscr{O}_{ii}}{\mathscr{O}_{ii} + \tau} \quad \text{and} \quad [\Theta_{Z,\tau}]_{jj} = \theta_j^Z \frac{\mathscr{P}_{jj}}{\mathscr{P}_{jj} + \tau}.$$

The ij'th element of $\mathscr{L}_\tau$ is

$$[\mathscr{L}]_{ij} = \frac{\mathscr{A}_{ij}}{\sqrt{(\mathscr{O}_{ii}+\tau)(\mathscr{P}_{jj}+\tau)}} = \frac{\theta_i^Y \theta_j^Z B_{y_i z_j}}{\sqrt{\mathscr{O}_{ii}\mathscr{P}_{jj}}} \sqrt{\frac{\mathscr{O}_{ii}}{\mathscr{O}_{ii}+\tau} \frac{\mathscr{P}_{jj}}{\mathscr{P}_{jj}+\tau}} = \frac{B_{z_i z_j}}{\sqrt{[P_B]_{y_i}[O_B]_{z_j}}} \sqrt{[\Theta_{Y,\tau}]_{ii}[\Theta_{Z,\tau}]_{jj}}.$$

Hence,

$$\mathscr{L} = \Theta_{Y,\tau}^{\frac{1}{2}} Z B_L Z^T \Theta_{Z,\tau}^{\frac{1}{2}},$$

where B_L is defined as

$$B_L = O_B^{-1/2} B P_B^{-1/2}. \tag{B.6}$$

□

Recall that $\mathscr{A} = \Theta_Y Y \mathbf{B} Z^T \Theta_Z$. Lemma 3.3 demonstrates that $\mathscr{L}$ has a similarly simple form that separates the block-related information (B_L) and node specific information (Θ_Y and Θ_Z).

Assume that $rank(B_L) = K, 0 < K = k_y \leq k_z$. Recall $H = (Y^T \Theta_{Y,\tau} Y)^{\frac{1}{2}} B_L (Z^T \Theta_{Z,\tau} Z)^{\frac{1}{2}}$. Singular value decomposition of H gives

$$H = U \Lambda V^T.$$

where $U \in \mathbb{R}^{k_y \times K} / V \in \mathbb{R}^{k_z \times K}$ is the left/right singular vector of H and $\Lambda \in \mathbb{R}^{K \times K}$ is diagonal containing the positive singular values of H, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$. The proof of the next lemma shows that H and $\mathscr{L}$ share the same nonzero singular values.

The next lemma gives the explicit form of the left and right population singular vectors and further shows that their normalized versions are block constant.

Lemma B.4. *(Singular value decomposition for $\mathscr{L}$) Under the DC-ScBM with parameters $\{\boldsymbol{B}, Y, Z, \Theta_Y, \Theta_Z\}$, Let $\mathscr{X}_L \in \mathbb{R}^{N_r \times K} (\mathscr{X}_R \in \mathbb{R}^{N_c \times K})$ contain the left/right singular vectors of $\mathscr{L}_\tau$. Define $\mathscr{X}_L^* / \mathscr{X}_R^*$ to be the row-normalized $\mathscr{X}_L / \mathscr{X}_R$. Then*

1. $\mathscr{X}_L = \Theta_{Y,\tau}^{\frac{1}{2}} Y (Y^T \Theta_{Y,\tau} Y)^{-\frac{1}{2}} U$,
2. $\mathscr{X}_R = \Theta_{Z,\tau}^{\frac{1}{2}} Z (Z^T \Theta_{Z,\tau} Z)^{-\frac{1}{2}} V$.
3. $\mathscr{X}_L^* = YU$, $Y_i \neq Y_j \Leftrightarrow Y_i U \neq Y_j U$.
4. $\mathscr{X}_R^* = ZV^*$, *where* $V_j^* = V_j / \|V_j\|_2$.

Proof. Recall that $H = (Y^T \Theta_{Y,\tau} Y)^{\frac{1}{2}} B_L (Z^T \Theta_{Z,\tau} Z)^{\frac{1}{2}}$ and singular value decompositon of H gives $H = U \Lambda V^T$.

Define $\mathscr{X}_L = \Theta_{Y,\tau}^{\frac{1}{2}} Y (Y^T \Theta_{Y,\tau} Y)^{-\frac{1}{2}} U$, and $\mathscr{X}_R = \Theta_{Z,\tau}^{\frac{1}{2}} Z (Z^T \Theta_{Z,\tau} Z)^{-\frac{1}{2}} V$. It is easy to check that $\mathscr{X}_L^T \mathscr{X}_L = I$ and $\mathscr{X}_R^T \mathscr{X}_R = I$.

On the other hand,

$$\mathscr{X}_L \Lambda \mathscr{X}_R^T = \Theta_{Y,\tau}^{\frac{1}{2}} Y B_L Z^T \Theta_{Z,\tau}^{\frac{1}{2}} = \mathscr{L}.$$

Hence, $\lambda_s, s = 1, ..., r$ are $\mathscr{L}_\tau$'s nonzero singular values and $\mathscr{X}_L / \mathscr{X}_R$ contains $\mathscr{L}_\tau$'s left/right singular vectors corresponding to its nonzero singular values.

Let $\mathscr{X}_L^i$ denote the i'th row of $\mathscr{X}_L$. For part (c), notice that

$$\|\mathscr{X}_L^i\|_2 = \left(\frac{[\Theta_{Y,\tau}]_{ii}}{[Y^T \Theta_{Y,\tau} Y]_{y_i y_i}}\right)^{\frac{1}{2}}.$$

So,

$$[\mathscr{X}_L^*]^i = \frac{\mathscr{X}_L^i}{\|\mathscr{X}_L^i\|_2} = Y_i U.$$

Therefore, $\mathscr{X}_L^* = YU$. For (d), notice that

$$\|\mathscr{X}_R^j\|_2 = (\frac{[\Theta_{Z,\tau}]_{jj}\|V_{Z_j}\|^2}{[Z^T\Theta_{Z,\tau}Z]_{z_j z_j}})^{\frac{1}{2}}.$$

Hence,

$$[\mathscr{X}_R^*]^j = \frac{\mathscr{X}_R^j}{\|\mathscr{X}_R^j\|_2} = Z_j V^*.$$

□

Comparing the population and observed clusters

The first part of the section proves the bound of misclustering rate for row nodes.

Clustering for Y

Proof. Recall that the set of misclustered row nodes is defined as:

$$\mathscr{M}_y = \left\{i : \|c_i^L - y_i\mu^y\mathscr{R}_L\|_2 > \|c_i^L - y_j\mu^y\mathscr{R}_L\|_2 \text{ for any } y_j \neq y_i\right\}.$$

Let $\mathcal{C}_i$ denote $y_i\mu^y$. Note that Lemma B.4 implies that the population centroid corresponding to the i'th row of $\mathscr{X}_L^*$ is

$$\mathcal{C}_i = y_i\mu^y = y_i U.$$

Since all population centroids are of unit length and are orthogonal to each other, a simple calculation gives a sufficient condition for one observed centroid to be closest to the population centroid:

$$\|c_i^L\mathscr{R}_L^T - \mathcal{C}_i^L\|_2 < 1/\sqrt{2} \Rightarrow \|c_i^L\mathscr{R}_L^T - \mathcal{C}_i^L\|_2 < \|c_i^L\mathscr{R}_L^T - \mathcal{C}_j^L\|_2, \quad \forall j \neq i.$$

Define the following set of nodes that do not satisfy the sufficient condition,

$$\mathscr{B}_y = \{i : \|c_i^L\mathscr{R}_L^T - \mathcal{C}_i^L\|_2 \geq 1/\sqrt{2}\}.$$

The mis-clustered nodes $\mathscr{M}_y \subset \mathscr{B}_y$.

Define $C_L \in \mathbb{R}^{N_r \times K}$, where the i'th row of C_L is c_i^L, the observed centroid of node i from the $(1+\alpha)$-approximate k-means. Define $M_L \in \mathbb{R}^{N_r \times K}$ to be the global solution of k-means. By definition,

$$\|X_L^* - C_L\|_F \leq (1+\alpha)\|X_L^* - M_L\|_F \leq (1+\alpha)\|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F.$$

Further, by the triangle inequality,

$$\|C_L - YU\mathscr{R}_L\|_F = \|C_L - \mathscr{X}_L^* \mathscr{R}_L\|_F \leq \|X_L^* - C_L\|_F + \|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F \leq (2+\alpha)\|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F.$$

Thus,

$$\begin{aligned}
\frac{|\mathscr{M}_y|}{N_r} \leq \frac{|\mathscr{B}_y|}{N_r} &= \frac{1}{N_r} \sum_{i \in \mathscr{B}_y} 1 \\
&\leq \frac{2}{N_r} \sum_{i \in \mathscr{B}_y} \|c_i^L \mathscr{R}_L^T - \mathcal{C}_i^L\|_2^2 \\
&= \frac{2}{N_r} \|C_L - YU\mathscr{R}_L\|_F^2 \\
&\leq \frac{2(2+\alpha)^2}{N_r} \|X_L^* - \mathscr{X}_L^* \mathscr{R}_L\|_F^2 \\
&\leq \frac{8(2+\alpha)^2}{N_r m_y^2} \|X_L - \mathscr{X}_L \mathscr{R}_L\|_F^2.
\end{aligned}$$

The last inequality is due to the following fact.

Lemma B.5. *For two non-zero vectors v_1, v_2 of the same dimension, we have*

$$\|\frac{v_1}{\|v_1\|_2} - \frac{v_2}{\|v_2\|_2}\|_2 \leq 2\frac{\|v_1 - v_2\|_2}{max(\|v_1\|_2, \|v_2\|_2)}.$$

By Theorem B.3, we have, with probability at least $1 - \epsilon$,

$$\frac{|\mathscr{M}_y|}{N_r} \leq c_0(\alpha) \frac{K \ln(4(N_r + N_c)/\epsilon)}{N_r \lambda_K^2 m_y^2 (\delta + \tau)}.$$

□

The second part proves the bound of the misclustering rate for column nodes.

Clustering for Z

Because $k_y \leq k_z$, it is slightly more challenging to bound $\mathscr{M}_z$.

Proof. Recall that $H = (Y^T\Theta_{Y,\tau}Y)^{\frac{1}{2}}B_L(Z^T\Theta_{Z,\tau}Z)^{\frac{1}{2}}$ and $H = U\Lambda V^T$. Left multiply by $\Lambda^{-1}U^T$, we have

$$V = H^T U\Lambda^{-1}.$$

Hence

$$\|V_i - V_j\|_2 \geq \frac{1}{\lambda_1}\|H_{\cdot i}U - H_{\cdot j}U\|_2 \geq \|H_{\cdot i} - H_{\cdot j}\|_2.$$

The second inequality is due to the facts that $\lambda_1 \leq 1$ and U is an orthogonal matrix. Recall that

$$\gamma_z = \min_{i\neq j}\|H_{\cdot i} - H_{\cdot j}\|_2 + (1-\kappa),$$

where $\kappa = \max_{i,j}\|V_i\|_2/\|V_j\|_2$. We have that, $\forall i \neq j$,

$$\|V_i^* - V_j^*\|_2 \geq \gamma_z.$$

This is because

$$\begin{aligned}
\|V_i^* - V_j^*\|_2 &= \|\frac{V_i - V_j}{\|V_j\|_2} + V_i(\frac{1}{\|V_i\|_2} - \frac{1}{\|V_j\|_2})\|_2 \\
&\geq \|V_i - V_j\|_2 + 1 - \frac{\|V_i\|_2}{\|V_j\|_2} \\
&\geq \|H_{\cdot i} - H_{\cdot j}\|_2 + (1-\kappa) \\
&\geq \gamma_z.
\end{aligned}$$

Recall that the set of misclustered row nodes is defined as:

$$\mathscr{M}_z = \left\{i : \|c_i^R - z_i\mu^z\mathscr{R}_R\|_2 > \|c_i^R - z_j\mu^z\mathscr{R}_R\|_2 \text{ for any } z_j \neq z_i\right\}.$$

Let $\mathcal{C}_i^R$ denote $z_i\mu^z$. Note that Lemma B.4 implies that the population centroid corresponding to the i'th row of $\mathscr{X}_R^*$ is

$$\mathcal{C}_i^R = z_i\mu^z = Z_iV^*.$$

Define the following set of column nodes,

$$\mathscr{B}_z = \{i : \|c_i^R\mathscr{R}_R^T - \mathcal{C}_i^R\|_2 \geq \gamma_z/2\}.$$

It is straightforward to show that $\mathscr{M}_z \in \mathscr{B}_z$.

Define $C_R \in \mathbb{R}^{N_c \times K}$, where the i'th row of M is c_i^R, the observed centroid of column node i from $(1+\alpha)$-approximate k-means. Define $M_R \in \mathbb{R}^{N_r \times K}$ to be the global solution of k-means. By definition, we have

$$\|X_R^* - C_R\|_F \leq (1+\alpha)\|X_R^* - M_R\|_F \leq (1+\alpha)\|X_R^* - \mathscr{X}_R^* \mathscr{R}_R\|_F.$$

Further, by the triangle inequality,

$$\|C_R - ZV^* \mathscr{R}_R\|_F = \|C_R - \mathscr{X}_R^* \mathscr{R}_R\|_F \leq \|X_R^* - C_R\|_F + \|X_R^* - \mathscr{X}_R^* \mathscr{R}_R\|_F \leq (2+\alpha)\|X_R^* - \mathscr{X}_R^* \mathscr{R}_R\|_F.$$

Putting all of these pieces together,

$$\begin{aligned}
\frac{|\mathscr{M}_z|}{N_c} \leq \frac{|\mathscr{B}_z|}{N_c} &= \frac{1}{N_c} \sum_{i \in \mathscr{B}_z} 1 \\
&\leq \frac{4}{N_c \gamma_z^2} \sum_{i \in \mathscr{B}_y} \|c_i^R \mathscr{R}_L^R - \mathcal{C}_i^R\|_2^2 \\
&= \frac{4}{N_c \gamma_z^2} \|C_R - ZV^* \mathscr{R}_R\|_F^2 \\
&\leq \frac{4(2+\alpha)^2}{N_c \gamma_z^2} \|X_R^* - \mathscr{X}_R^* \mathscr{R}_R\|_F^2 \\
&\leq \frac{16(2+\alpha)^2}{N_c \gamma_z^2 m_z^2} \|X_R - \mathscr{X}_R \mathscr{R}_R\|_F^2.
\end{aligned}$$

By Theorem B.3, we have with probability at least $1-\epsilon$,

$$\frac{|\mathscr{M}_z|}{N_c} \leq c_1(\alpha) \frac{K \ln(4(N_r + N_c))/\epsilon)}{N_r \lambda_K^2 m_z^2 \gamma_z^2 (\delta + \tau)}.$$

□

The following is a proof of Corollary 3.9.

Proof. Under the four parameter ScBM, presume that $\theta_i = 1/s$ for all i. From the proof of Lemma B.4, $\mathscr{L}$ has the same singular values as

$$H = (Y^T \Theta_{Y,\tau=0} Y)^{\frac{1}{2}} B_L (Z^T \Theta_{Z,\tau=0} Z)^{\frac{1}{2}} = B_L = O_B^{-\frac{1}{2}} B P_B^{-\frac{1}{2}} = \frac{1}{s^2(Kr+p)} (s^2 p I_K + s^2 r \mathbf{1}_K \mathbf{1}_K^T).$$

By inspection, the constant vector is an eigenvector of this matrix. It has eigenvalue

$$\lambda_1 = \frac{p + Kr}{Kr + p} = 1.$$

Any vector orthogonal to a constant vector is also an eigenvector. These eigenvectors have eigenvalue

$$\lambda_k = \frac{p}{Kr + p} = \frac{1}{K(r/p) + 1}.$$

The result follows from using $m_y^2 = K/n$ (see discussion after Theorem 3.7) and $\delta \propto N$. □

Appendix C

Appendix for Chapter 4

Proof of Theorem 4.4

The proof requires some additional definitions. After giving these definitions, we will outline the proof.

Define the expectation of $\hat{\boldsymbol{\theta}}^{(z)}$ and $\hat{\boldsymbol{\theta}}^{R,(z)}$ to be $\bar{\boldsymbol{\theta}}^{(z)}$ and $\bar{\boldsymbol{\theta}}^{R,(z)}$. Define the expectation of $L(A; z, \boldsymbol{\theta})$ to be

$$\bar{L}_P(z, \boldsymbol{\theta}) = E[L(A; z, \boldsymbol{\theta})] = \sum_{i<j} \{P_{ij} \log \theta_{z_i z_j} + (1 - P_{ij}) \log(1 - \theta_{z_i z_j})\}.$$

Let $\bar{L}_P(z)$ to be the maximizer of $\bar{L}_P(z, \boldsymbol{\theta})$ over Θ, and let $\bar{L}_P^R(z)$ to be the maximizer of $\bar{L}_P(z, \boldsymbol{\theta})$ over Θ^R. That is,

$$\bar{L}_P(z) = \bar{L}_P(z, \bar{\boldsymbol{\theta}}^{(z)}) = \sup_{\boldsymbol{\theta} \in \Theta} \bar{L}_P(z, \boldsymbol{\theta}), \tag{C.1}$$

$$\bar{L}_P^R(z) = \bar{L}_P(z, \bar{\boldsymbol{\theta}}^{R,(z)}) = \sup_{\boldsymbol{\theta} \in \Theta^R} \bar{L}_P(z, \boldsymbol{\theta}). \tag{C.2}$$

The proof of the main theorem is divided into five lemmas. The first step is to bound the difference between $\bar{L}_P(\tilde{z})$ and $\bar{L}_P^R(\hat{z}^R)$ (Lemma C.3). Lemma C.1 and Lemma C.2 are two building blocks of Lemma C.3 . Lemma C.1 establishes a union bound of $|L^R(A; z) - \bar{L}_P^R(z)|$ for any partition z. Lemma 2 shows that under the true partition $\tilde{z}$, the expectation of regularized likelihood is close to the expectation of the ordinary likelihood. Lemma C.3 divides $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$ into three parts and controls them respectively. We can see this as a bias-variance tradeoff; we sacrifice some bias $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z})$ to decrease the variance $\max_z |L^R(A; z) - \bar{L}_P^R(z)|$. After Lemma C.3,

it is necessary to develop the concept of regularized refinement, an extension of the refinement idea proposed in Choi et al. (2012). Using the concept of regularized refinement, we can bound the error rate $N_e(\hat{z}^R)/N$ with a function of $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$. Lemma C.5 and Lemma C.6 use a new concept of regularized refinement to connect the bounds on the log-likelihood with the error rate $N_e(\hat{z}^R)/N$. From here on, we write $\hat{\theta}$ and $\bar{\theta}$ instead of $\hat{\theta}^{(z)}$ and $\bar{\theta}^{(z)}$ when the choice of z is understood.

Lemma C.1. *Let M to be the total expected degree of A. That is, $M = \sum_{i<j} EA_{ij}$.*

$$\max_z |L^R(A; z) - \bar{L}_P^R(z)| = o_p(M). \tag{C.3}$$

This proof follows a similar argument made in Choi, Wolfe, and Airoldi (2012).

Proof. Let $H(p) = -p \log p - (1-p)\log(1-p)$, which is the entropy of a Bernoulli random variable with parameter p. Define $X = \sum_{i<j} A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$. Let n_{ab} denote the maximum number of possible edges between all different blocks.

$$\begin{aligned} L^R(A; z) - \bar{L}_P^R(z) &= -\sum_{a=1}^{K} n_{aa}(H(\hat{\theta}_{aa}) - H(\bar{\theta}_{aa})) - n_{out}(H(\hat{r}) - H(\bar{r})) \\ &= \sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r}) + X - E(X). \end{aligned}$$

For the first part $\sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r})$, by similar argument as in Choi et al. (2012), we have that for every regularized estimator $\hat{\theta}^R$:

$$pr(\hat{\theta}^R) \leq \exp\left\{ -\sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) - n_{out} D(\hat{r} \| \bar{r}) \right\}.$$

Let $\hat{\Theta}$ denote the range of $\hat{\theta}^R$ for fixed z. Then the total number of sets of values $\hat{\theta}^R$ can take is $|\hat{\Theta}| = (n_{out}+1) \cdot \Pi_{a=1}^{K}(n_{aa}+1)$. Notice that $\sum_{a=1}(n_{aa}+1) + (n_{out}+1) = \frac{N(N-1)}{2} + K + 1$, we have $|\hat{\Theta}| \leq (\frac{N(N-1)}{2(K-1)} + 1)^{K+1} \leq (\frac{N^2}{2K})^{(K+1)}$. Then $\forall \epsilon > 0$,

$$\begin{aligned} pr\left\{ \sum_{a=1}^{K} n_{aa} D(\hat{\theta}_{aa} \| \bar{\theta}_{aa}) + n_{out} D(\hat{r} \| \bar{r}) > \epsilon \right\} &\leq |\hat{\Theta}| e^{-\epsilon} \leq (\frac{N^2}{2K})^{(K+1)} e^{-\epsilon} \\ &\leq \exp\left\{ 2(K+1)\log N - (K+1)\log(2K) - \epsilon \right\}. \end{aligned}$$

For the second part $X - E(X)$, each $X_{ij} = A_{ij} \log\{\bar{\theta}_{z_i z_j}/(1-\bar{\theta}_{z_i z_j})\}$ is bounded in magnitude by $C = 2\log N$. By the following concentration inequality:

$$pr\{|X - E(X)| \geq \epsilon\} \leq 2\exp\left\{ - \frac{\epsilon^2}{2\sum_{i<j} E(X_{ij}^2) + (2/3)C\epsilon} \right\}.$$

Here $\sum_{i<j} E(X_{ij}^2) \leq 4M\log^2 N$. Finally, by a union bound inequality over all partition z, we have:

$$pr\{\max_z |L^R(A;z) - \bar{L}_P^R(z)| \geq 2\epsilon M\} \leq \exp\{N\log K + 2(K+1)\log N - (K+1)\log(2K) - M\epsilon\} + 2\exp\left\{ N\log K - \frac{\epsilon^2 M}{8\log^2 N + (4/3)\epsilon \log N} \right\}.$$

Notice that in this asymptotic setting , the total expected degree $M = \omega(N(\log N)^{3+\delta})$. Then, $\max_z |L^R(A;z) - \bar{L}_P^R(z)| = o_p(M)$. □

Lemma C.2. *Under the true partition $\tilde{z}$, $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z}) = o(M)$.*

Proof. When N is sufficiently large,

$$\begin{aligned}
\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z}) &= \sum_{a<b} n_{ab} D(\theta_{ab}\|\bar{r}) = \sum_{a<b,\{a,b\}\in Q} n_{ab} D(\theta_{ab}\|\bar{r}) + \sum_{a<b,\{a,b\}\notin Q} n_{ab} D(\theta_{ab}\|\bar{r}) \\
&\leq |Q|C_1 + (N(N-1)/2 - \sum_{a=1}^{K} n_{aa} - |Q|)\frac{Cf(N)}{N}(\log(CNf(N))) \\
&\leq |Q|C_1 + N^2\frac{Cf(N)}{N}(\log N + \log Cf(N)) = o(M).
\end{aligned}$$

Here $C_1 > 0$ is some constant. The last equality is due to the fact that $M = \Omega(Ns)$, which is directly implied by Definition 2.2. □

Lemma C.3. *Under the true partition $\tilde{z}$ and the RMLE $\hat{z}^R$, $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) = o_p(M)$.*

Proof. First notice that the left hand side is a nonnegative value since $\tilde{z}$ maximizes $\bar{L}_P(\cdot)$ and $\bar{L}_P(\hat{z}^R) \geq \bar{L}_P^R(\hat{z}^R)$.

By adding another positive term, and using Lemma C.1 and Lemma C.2:

$$\begin{aligned}\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) &\leq \bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) + L^R(A;\hat{z}^R) - L^R(A,\tilde{z}) \\ &\leq |\bar{L}_P(\tilde{z}) - L^R(A,\tilde{z})| + |\bar{L}_P^R(\hat{z}^R) - L^R(A;\hat{z}^R)| \\ &\leq |\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\tilde{z})| + |\bar{L}_P^R(\tilde{z}) - L^R(A,\tilde{z})| + |\bar{L}_P^R(\hat{z}^R) - L^R(A;\hat{z}^R)| \\ &= o_p(M).\end{aligned}$$

□

To make $N_e(z)$ mathematically tractable, Choi, Wolfe, and Airoldi (2012) introduced the concept of block refinements. The next paragraphs first reintroduce the definition. We then extend this definition to the regularized block refinement.

Partitions and refinements

The refinement is the key concept to connect $\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R)$ with the error rate $N_e(\hat{z}^R)/N$. For this subsection, we first review the concept of partition and refinement. Then, we give its regularized version. Second, we state the fact that a refinement's log-likelihood is no less than the original partition's log-likelihood. Then, the distance between log-likehood of its (regularized) refinement and log-likelihood of true $\tilde{z}$ can be bounded by the distance between (regularized) log-likelihood of arbitrary z partitions and log-likelihood true $\tilde{z}$. Finally, the connection between (regularized) refinement log-likelihood and the error rate is established (Lemma C.6.)

For positive integer N, define $[N]$ as the set $\{1,\ldots,N\}$. The partition log-likelihood $\bar{L}_P^*$ is defined for any partition Π of the indices of a lower triangular matrix,

$$\Pi : \{(i,j)\}_{i\in[N],j\in[N],i<j} \to (1,\ldots,L).$$

Define

$$S_\ell = \{(i,j) : \Pi(i,j) = \ell \text{ and } i < j\} \quad \text{and} \quad \bar{\theta}_\ell = |S_\ell|^{-1} \sum_{i<j:\Pi(i,j)=\ell} P_{ij}.$$

The partition log-likelihood is defined as

$$\bar{L}_P^*(\Pi) = \sum_{i<j} \{P_{ij} \log \bar{\theta}_{\Pi(i,j)} + (1 - P_{ij}) \log(1 - \bar{\theta}_{\Pi(i,j)})\}.$$

Notice that any class assignment z induces a corresponding partition Π^z,

$$\Pi^z(i,j) = \ell, \quad \text{where } \ell = z_i + (z_j - 1) \cdot K.$$

It is straightforward to show that $\bar{L}_P^*(\Pi^z) = \bar{L}_P(z)$.

A refinement Π' of partition Π further divides the partitions in Π into subgroups. Formally,

Definition C.4. *A **refinement** Π' of partition Π satisfies the following condition.*

$$\Pi'(i_1,j_1) = \Pi'(i_2,j_2) \quad \Longrightarrow \quad \Pi(i_1,j_1) = \Pi(i_2,j_2), \quad \textit{for any } i_1 < j_1 \textit{ and } i_2 < j_2.$$

From Lemma A2 in Choi et al. (2012),

$$\bar{L}_P^*(\Pi) \leq \bar{L}_P^*(\Pi') \tag{C.4}$$

This will be essential for for Lemma C.6.

To define Π^*, a specific refinement of partition Π^z, we first need to define a set of triples T. The following construction comes directly from Choi et al. (2012):

> "For a given membership class under z, partition the corresponding set of nodes into subclasses according to the true class assignment $\tilde{z}$ of each node. Then remove one node from each of the two largest subclasses so obtained, and group them together as a pair; continue this pairing process until no more than one nonempty subclass remains. Then, terminate. If pair (i,j) is chosen from the above procedure, then $z_i = z_j$ and $\tilde{z}_i \neq \tilde{z}_j$."

Define C_1 as the number of (i,j) pairs selected by the above routine. Notice that at least one of i or j is misclustered. In fact, $N_e(z)/2 \leq C_1 \leq N_e(z)$. This will be important for Lemma C.5 which connects the error rate $N_e(z)/N$ with the refinement.

Define the set T to contain the triple (i,j,k) if the pair (i,j) was tallied in C_1, and $k \in [N]$ satisfies

$$D\left(P_{ik} \| \frac{P_{ik} + P_{jk}}{2}\right) + D\left(P_{jk} \| \frac{P_{ik} + P_{jk}}{2}\right) \geq C\frac{MK}{N^2}.$$

From assuming Equation 4.3 , if (i,j) is tallied in C_1, then there exists at least one such k. Further, if $z_k = z_\ell$, then (i,j,ℓ) is also in T. The set T is essential to defining the refinement partition Π^* and later the refined regularized partition Π^{*R}.

For each $(i,j,k) \in T$, remove (i,k) and (j,k) from their previous subset under Π^z, and place them into their own, distinct two-element set. Define the resulting partition as Π^*. Notice that it is a refinement of Π^z.

Regularized partition and regularized refinement

To extend the analysis to the RMLE, we will define the regularized partition Π^{zR} and the associated refinement partition Π^{*R}. Π^{zR} partitions the nodes into $K+1$ groups; if $z_i = z_j$, then $\Pi^{zR}(i,j) = z_i$ and if $z_i \neq z_j$, then $\Pi^{zR}(i,j) = K+1$. It follows from the definition of $\bar{L}_p^*$ that $\bar{L}_p^R(z) = \bar{L}_p^*(\Pi^{zR})$.

Construct Π^{*R} in the following way: For each $(i,j,k) \in T$, remove (i,k) and (j,k) from their previous subset under Π^{zR}, and place them into their own, distinct two-element set. Define the resulting partition as Π^{*R}. Notice that Π^{*R} is constructed from Π^{zR} in the same way that Π^* is constructed from Π^z. Define R as the set of elements in the off-diagonal block partition that where not removed by the set T,

$$R = \left\{(q,k) \in [N] \times [N] : z_q \neq z_k,\ (q,x,k) \notin T,\ (x,q,k) \notin T,\ \text{ for any } x \in [N]\right\}.$$

Notice that R is one group in Π^{*R}. Make a refinement Π' by subdividing R into $\binom{K}{2}$ new groups:

$$\text{For } u < v, u \in [K], v \in [K],\ \text{define } G_{uv} = \left\{(i,j) \in R : z_i = u, z_j = v \ \text{ or } z_i = v, z_j = u\right\}.$$

It follows that $\Pi' = \Pi^*$. So, Π^* is a refinement of Π^{*R} and Π^{*R} is a refinement for Π^{zR}.

Lemma C.5. *(Theorem 3 in Choi, Wolfe, and Airoldi (2012)) For any partition z and Π^* being its refinement, if the size of the smallest block $s = \Omega(\frac{MK}{N^2})$, and for any distinct class pairs (a, b), there exists a class c such that Equation 4.3 holds, then*

$$\bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi^*) = \frac{N_e(z)}{N}\Omega(M). \tag{C.5}$$

Proof.

$$\bar{L}_p(\tilde{z}) - \bar{L}_p^*(\Pi^*) = \sum_{i<j} D(P_{ij} || \bar{\theta}_{\Pi(i,j)}) = C_1 \Omega\left(s\ \frac{MK}{N^2}\right) = \frac{N_e(z)}{N}\Omega(M)$$

□

Lemma C.6. *Let $\Pi^{\hat{z}^R}$ be the partition corresponding to $\hat{z}^R$ (the regularized block estimator). Let Π' be the refinement of $\Pi^{\hat{z}^R}$, and let Π'^R be the regularized refinement of $\Pi^{\hat{z}^R}$.*

$$\bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi'^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi'). \tag{C.6}$$

Proof. Recall that taking a refinement increases the partition log-likelihood. The first inequality is due to the fact that Π'^R is a refinement of the partition $\Pi^{\hat{z}^R}$. The second inequality follows from the fact that Π' is a refinement of Π'^R. □

Hence Theorem 4.4 can be proved as below: The conditions in Lemma C.5 are satisfied by the highest dimensional asymptotic setting assumption. By Lemma C.3, C.5, C.6, we have:

$$o_p(M) = \bar{L}_P(\tilde{z}) - \bar{L}_P^R(\hat{z}^R) \geq \bar{L}_P(\tilde{z}) - \bar{L}_P^*(\Pi') = \frac{N_e(\hat{z}^R)}{N}\Omega(M). \text{ Hence } \frac{N_e(\hat{z}^R)}{N} = o_p(1).$$

Reseeding

Section 4.4 describe a reseeding technique that ensures the pseudo-likelihood implementation of the RMLE returns an estimated partition with the desired number of non-empty sets. This section compares the implementation with reseeding (`reseed`) to the implementation without reseeding (`no.reseed`). Overall, `reseed` never attains a smaller likelihood score and often attains a larger likelihood scores. Moreover, `reseed` is more stable over different initializations.

In the following simulation, $K = 30$, $n = 600$, $\theta_{ii} = 8/20$ for all i, and $\theta_{ij} = 10/580$ for all $i \neq j$. So, in expectation, each node connects to 8 nodes in the same block and 10 nodes in other blocks. For each simulated adjacency matrix A, both `reseed` and `no.reseed` are both initialized 50 times with spectral clustering; due to the k-means step in spectral clustering, not all initializations are equivalent. The simulations in Section 4.4 reseeds whenever a block contains either zero nodes or a single node. In the simulation that follows, blocks are reseeded whenever they have fewer than five nodes. This is to demonstrate the robustness of this block culling step.

There are 175 adjacency matrices A simulated from this model. For the ith simulated adjacency matrix, let $\hat{z}^i_{reseed}$ be the partition that attains the largest likelihood over all 50 initializations of `reseed`. Similarly, let $\hat{z}^i_{no.reseed}$ be the same partition for `no.reseed`. Over the 175 simulated adjacency matrices,

$$L^R(A; \hat{z}^i_{reseed}) > L^R(A; \hat{z}^i_{no.reseed})$$

on 22% of the simulations. In the remaining simulations, they find the same maximum. Never does $\hat{z}^i_{no.reseed}$ attain a larger likelihood score. Moreover, on each initialization, the `reseed` was much more likely to find the maximum (72 % compared to 14 %).

Appendix D

Appendix for Chapter 5

Proof of Lemma 5.2

Proof. First notice that under the DC-SBM, the population adjacency matrix has the following form,

$$\mathscr{A} = \Theta_A Z_A P Z_A^T \Theta_A.$$

Simply algebra reveals that the population graph laplacian has similar simply presentation:

$$\mathscr{L}_{\mathscr{A}} = \mathscr{D}_{\mathscr{A}}^{-1/2} \mathscr{A} \mathscr{D}_{\mathscr{A}}^{-1/2} = \Theta_A^{1/2} Z_A D_P^{-1/2} P D_P^{-1/2} Z_A^T \Theta_A^{1/2},$$

where $D_P = diag(P Z_A^T \Theta_A 1)$. Define matrix $C \in \mathcal{R}^{K \times K}$,

$$C = (Z_A^T \Theta_A Z_A)^{1/2} D_P^{-1/2} P D_P^{-1/2} (Z_A^T \Theta_A Z_A)^{1/2}.$$

By assumption, P is positive definite, hence C is also positive definite given D_P and $(Z_A^T \Theta_A Z_A)^{1/2}$ are both diagonal and positive matrices. Spectral decomposition of C gives $C = U \bar{\Lambda} U^T$, where $U \in \mathcal{R}^{K \times K}$ contains eigenvectors of C in its columns and $\bar{\Lambda} = diag(\bar{\lambda}_1, ..., \bar{\lambda}_K), \bar{\lambda}_1 \geq \bar{\lambda}_2, ..., \geq \bar{\lambda}_K > 0$ Define $\tilde{\mathscr{X}} := \Theta_A^{1/2} Z_A (Z_A^T \Theta_A Z_A)^{-1/2} U$. It is easy to check that $\tilde{\mathscr{X}}^T \tilde{\mathscr{X}} = I$ and $\tilde{\mathscr{X}} \bar{\Lambda} \tilde{\mathscr{X}}^T = \mathscr{L}_{\mathscr{A}}$. Hence, $\tilde{\mathscr{X}}$ contains top K eigenvectors of $\mathscr{L}_{\mathscr{A}}$ and $\bar{\Lambda}$ contains top K positive eigenvalues of $\mathscr{L}_{\mathscr{A}}$.

Recall that $\mathscr{X} = \mathscr{D}_{\mathscr{A}}^{-1/2} \tilde{\mathscr{X}}$ and $[\mathscr{D}_{\mathscr{A}}]_{ii} = [\Theta_A]_{ii} [P Z_A^T \Theta_A 1]_{[Z_A]_i}$, we have

$$\mathscr{X} = \mathscr{D}_{\mathscr{A}}^{-1/2} \tilde{\mathscr{X}} = Z_A D_P^{-1/2} (Z_A^T \Theta_A Z_A)^{-1/2} U.$$

Notice the fact that $Z_A^T \Theta_A 1 = Z_A^T \Theta_A Z_A 1$. Also recall the definition of V, $V =$

$diag[(Z_A^T\Theta_A Z_A)P(Z_A^T\Theta_A Z_A)\mathbf{1}]$. Hence we have

$$\mathscr{X} = Z_A diag(PZ_A^T\Theta_A Z_A 1)^{-1/2}(Z_A^T\Theta_A Z_A)^{-1/2}U = Z_A V^{-1/2}U.$$

$\mathscr{X}^*$ projects each row of $\mathscr{X}$ on the unit sphere, it is straight forward to show that $\mathscr{Y}^* = Z_A U$.

□

Proof of Lemma 5.3

Proof. By definition,

$$\begin{aligned}\mathscr{Y} &= (\mathscr{D}_{\mathscr{B}} + \tau I)^{-1}\mathscr{B}\mathscr{X}\bar{\Lambda}^{-1} \\ &= (\mathscr{D}_{\mathscr{B}} + \tau I)^{-1}\mathscr{D}_{\mathscr{B}}\mathscr{D}_{\mathscr{B}}^{-1}\mathscr{B}\mathscr{X}\bar{\Lambda}^{-1} \\ &= (\mathscr{D}_{\mathscr{B}} + \tau I)^{-1}\mathscr{D}_{\mathscr{B}}Z_B D_P^{-1}PZ_A^T\Theta_A\mathscr{X}\bar{\Lambda}^{-1}\end{aligned}$$

$\mathscr{X}$ is also eigenvectors of the random walk graph laplacian,

$$\mathscr{D}_{\mathscr{A}}^{-1}\mathscr{A}\mathscr{X} = \mathscr{X}\bar{\Lambda}.$$

Plugging in the definition of $\mathscr{A}$ and $\mathscr{X}$,

$$\begin{aligned}\mathscr{D}_{\mathscr{A}}^{-1}\mathscr{A}\mathscr{X} &= Z_A D_P^{-1}PZ_A^T\Theta_A\mathscr{X} \\ \Rightarrow Z_A D_P^{-1}PZ_A^T\Theta_A\mathscr{X}\bar{\Lambda}^{-1} &= Z_A V^{-1/2}U.\end{aligned}$$

This indicates that $D_P^{-1}PZ_A^T\Theta_A\mathscr{X}\bar{\Lambda}^{-1} = V^{-1/2}U$. Hence $Z_B D_P^{-1}PZ_A^T\Theta_A\mathscr{X}\bar{\Lambda}^{-1} = Z_B V^{-1/2}U$, and

$$\mathscr{Y} = (\mathscr{D}_{\mathscr{B}} + \tau I)^{-1}\mathscr{D}_{\mathscr{B}}Z_B V^{-1/2}U.$$

$\mathscr{Y}^*$ projects each row of $\mathscr{Y}$ on the unit sphere, it is straight forward to show that $\mathscr{Y}^* = Z_B U$. □

Proof of Theorem 5.5

Proof.

$$\|F - \mathscr{F}\|_F^2 = \|X^* - \mathscr{X}^*\|_F^2 + \|Y^* - \mathscr{Y}^*\|_F^2.$$

The first part bounds $\|X^* - \mathscr{X}^*\|_F^2$:

By applying an improved version of Davis Kahn theorem from **?**) and Lemma 5.4, we have,

$$\|\tilde{X} - \tilde{\mathscr{X}}\|_F \leq \frac{2\sqrt{2K}}{\bar{\lambda}_K}\|L - \mathscr{L}_{\mathscr{A}}\| \leq 4\sqrt{6}\sqrt{\frac{K\ln(4N_A/\epsilon)}{\bar{\lambda}_K^2\delta_A}}, \tag{D.1}$$

with probability greater than $1-\epsilon$.

Lemma D.1. *For two non-zero vectors v_1, v_2 of the same dimension, we have*

$$\|\frac{v_1}{\|v_1\|_2} - \frac{v_2}{\|v_2\|_2}\|_2 \leq 2\frac{\|v_1 - v_2\|_2}{max(\|v_1\|_2, \|v_2\|_2)}.$$

Let $V_{max} = max_i V_{ii}$, which is the maximum expected volume of K clusters within A. Recall that $\|X_i\| = V_{z_i}^{-1/2}$, by Lemma D.1 and some algebra,

$$\|X^* - \mathscr{X}^*\|_F \leq 2\sqrt{\frac{V_{max}}{\delta_A}}\|\tilde{X} - \tilde{\mathscr{X}}\|_F. \tag{D.2}$$

Combining this with equation (D.1) gives:

$$\|X^* - \mathscr{X}^*\|_F \leq 8\sqrt{6}\frac{\sqrt{KV_{max}\ln(4N_A/\epsilon)}}{\delta_A\bar{\lambda}_K}, \tag{D.3}$$

with probability greater than $1-\epsilon$.

Lemma D.2. *For any $0 < \epsilon < 1$, if $\delta_A \geq 3\ln(4N_A/\epsilon)$, with probability at least $1-\epsilon$,*

$$\|D_A^{-1/2} - \mathscr{D}_{\mathscr{A}}^{-1/2}\| \leq \frac{2\sqrt{3\ln(4N_A/\epsilon)}}{\delta_A}.$$

By triangle inequality,

$$\begin{aligned}
\|X - \mathscr{X}\|_F &= \|D_A^{-1/2}\tilde{X} - \mathscr{D}_{\mathscr{A}}^{-1/2}\tilde{\mathscr{X}}\|_F && \text{(D.4)}\\
&\leq \|(D_A^{-1/2} - \mathscr{D}_{\mathscr{A}}^{-1/2})\tilde{X}\|_F + \|\mathscr{D}_{\mathscr{A}}^{-1/2}(\tilde{X} - \tilde{\mathscr{X}})\|_F && \text{(D.5)}\\
&\leq \sqrt{K}\|(D_A^{-1/2} - \mathscr{D}_{\mathscr{A}}^{-1/2})\| + \delta_A^{-1/2}\|\tilde{X} - \tilde{\mathscr{X}}\|_F && \text{(D.6)}\\
&\leq 8\sqrt{6}\frac{\sqrt{K\ln(4N_A/\epsilon)}}{\delta_A\bar{\lambda}_K} && \text{(D.7)}
\end{aligned}$$

Next part bounds $\|Y^* - \mathscr{Y}^*\|_F^2$ with high probability.

For brevity, Let $D_i(\mathscr{D}_i)$ denote the ii'th element of $D_B(\mathscr{D}_{\mathscr{B}})$. For any $1 \leq i \leq N_B$, By triangle inequality,

$$\|Y_i - \mathscr{Y}_i\| = \|(D_i+\tau)^{-1}B_i X\Lambda^{-1} - (\mathscr{D}_i+\tau)^{-1}\mathscr{B}_i\mathscr{X}\bar{\Lambda}^{-1}\| \tag{D.8}$$

$$\leq \|(D_i+\tau)^{-1}B_i\|\|X\Lambda^{-1} - \mathscr{X}\bar{\Lambda}^{-1}\|_F + \|(D_i+\tau)^{-1}B_i\mathscr{X}\bar{\Lambda}^{-1} - (\mathscr{D}_i+\tau)^{-1}\mathscr{B}_i\mathscr{X}\bar{\Lambda}^{-1}\| \tag{D.9}$$

For the first part:

$$\begin{aligned}
\|(D_i+\tau)^{-1}B_i\|\|X\Lambda^{-1} - \mathscr{X}\bar{\Lambda}^{-1}\|_F &\leq \frac{\sqrt{D_i}}{D_i+\tau}\|X\Lambda^{-1} - \mathscr{X}\bar{\Lambda}^{-1}\|_F \\
&\leq \frac{1}{\sqrt{D_i+\tau}}\big(\|X\Lambda^{-1} - \mathscr{X}\Lambda^{-1}\|_F + \|\mathscr{X}\Lambda^{-1} - \mathscr{X}\bar{\Lambda}^{-1}\|_F\big) \\
&\leq \frac{1}{\sqrt{D_i+\tau}}\big(\frac{1}{\lambda_K}\|X - \mathscr{X}\| + \frac{1}{\sqrt{\delta_A}}\sqrt{K}\|\Lambda^{-1} - \bar{\Lambda}^{-1}\|\big) \\
&\leq \frac{1}{\sqrt{D_i+\tau}}\big(\frac{2}{\bar{\lambda}_K}\|X - \mathscr{X}\| + \frac{4\sqrt{3K\ln(4N_A/\epsilon)}}{\bar{\lambda}_K^2\delta_A}\big).
\end{aligned}$$

The first inequality follows because $\|B_i\| = \sqrt{D_i}$.

Following equation equation D.4 and the next lemma, we have

$$\|(D_i+\tau)^{-1}B_i\|\|X\Lambda^{-1} - \mathscr{X}\bar{\Lambda}^{-1}\|_F \leq C\frac{\sqrt{K\ln(4N_A/\epsilon)}}{\sqrt{\delta_B+\tau}\delta_A\bar{\lambda}_K^2}, \tag{D.10}$$

for all $i \in V_B$ with probability at least $1-\epsilon$.

Lemma D.3. *Let $X_1, ..., X_n$ be independent $0/1$ random variables and $\tau \geq 0$, and $X = \sum X_i$. If $\mathrm{E}X + \tau \geq \frac{32}{9}\ln(1/\epsilon)$, then with probability at lease $1-\epsilon$,*

$$X + \tau \geq \frac{1}{4}(\mathrm{E}X + \tau).$$

For the second part:

$$\begin{aligned}
&\|(D_i+\tau)^{-1}B_i\mathscr{X}\bar{\Lambda}^{-1} - (\mathscr{D}_i+\tau)^{-1}\mathscr{B}_i\mathscr{X}\bar{\Lambda}^{-1}\| \leq \\
&\frac{1}{\bar{\lambda}_K}\|(D_i+\tau)^{-1}B_i\mathscr{X} - (\mathscr{D}_i+\tau)^{-1}B_i\mathscr{X}\| + \frac{1}{\bar{\lambda}_K(\mathscr{D}_i+\tau)}\|B_i\mathscr{X} - \mathscr{B}_i\mathscr{X}\|
\end{aligned}$$

$$\begin{aligned}
\|(D_i+\tau)^{-1}B_i\mathscr{X} - (\mathscr{D}_i+\tau)^{-1}B_i\mathscr{X}\| &\leq \big|\frac{1}{D_i+\tau} - \frac{1}{\mathscr{D}_i+\tau}\big|\|B_i\|\|\mathscr{X}\| \\
&\leq \big|\frac{1}{D_i+\tau} - \frac{1}{\mathscr{D}_i+\tau}\big|\sqrt{D_i}\|\mathscr{X}\| \\
&\leq \frac{|D_i-\mathscr{D}_i|}{(D_i+\tau)(\mathscr{D}_i+\tau)}\sqrt{D_i}\|\mathscr{X}\| \\
&\leq \frac{|D_i-\mathscr{D}_i|}{\sqrt{D_i+\tau}(\mathscr{D}_i+\tau)}\|\mathscr{X}\| \\
&\leq \frac{|D_i-\mathscr{D}_i|}{\mathscr{D}_i+\tau}\frac{2\sqrt{K}}{\sqrt{\delta_B+\tau}\sqrt{\delta_A}} \\
&\leq \frac{2\sqrt{3}}{\sqrt{\delta_A}(\delta_B+\tau)}\sqrt{K\ln(4N_B/\epsilon)}.
\end{aligned}$$

Hence,

$$\frac{1}{\bar{\lambda}_K}\|(D_i+\tau)^{-1}B_i\mathscr{X} - (\mathscr{D}_i+\tau)^{-1}B_i\mathscr{X}\| \leq \frac{2\sqrt{3}}{\bar{\lambda}_K\sqrt{\delta_A}(\delta_B+\tau)}\sqrt{K\ln(4N_B/\epsilon)} \quad \text{(D.11)}$$

holds with probability greater than $1-\epsilon$ for all $i \in V_B$.

To bound the last part $\frac{1}{\bar{\lambda}_K(\mathscr{D}_i+\tau)}\|B_i\mathscr{X} - \mathscr{B}_i\mathscr{X}\|$, we use the following lemma from Chaudhuri et al. (2012):

Lemma D.4. *Let $X_1, ..., X_n$ be independent 0/1 random variables, and $X = \sum \alpha_i X_i$. Let $\|\alpha\|^2 = \sum \alpha_i^2$. With probability at lease $1-2\epsilon$,*

$$|X - \mathrm{E}[X]| \leq \sqrt{\frac{\|\alpha\|^2\ln(1/\epsilon)}{2}}.$$

The proof of lemma D.4 comes directly from Hoeffding's inequality.

Note that $\|B_i\mathscr{X} - \mathscr{B}_i\mathscr{X}\|^2 = \sum_{j=1}^K \|B_i\mathscr{X}_{\cdot j} - \mathscr{B}_i\mathscr{X}_{\cdot j}\|^2$, by lemma D.4,

$$\|B_i\mathscr{X}_{\cdot j} - \mathscr{B}_i\mathscr{X}_{\cdot j}\|^2 \leq \frac{\|\mathscr{X}_{\cdot j}\|^2\ln(1/\epsilon)}{2}.$$

Recall that $\mathscr{X}_{\cdot j} = \mathscr{D}_{\mathscr{A}}^{-1/2}\tilde{\mathscr{X}}_{\cdot j}$ and $\|\tilde{\mathscr{X}}_{\cdot j}\| = 1$ for all j by definition of eigenvector. Simple algebra gives $\|\mathscr{X}_{\cdot j}\|^2 \leq \frac{1}{\delta_A}$. Hence,

$$\|B_i\mathscr{X} - \mathscr{B}_i\mathscr{X}\|^2 \leq \frac{K\ln(N_B/\epsilon)}{2\delta_A},$$

with probability at least $1 - \frac{2\epsilon}{N_B}$.

Applying union bound over $i \in V_B$ gives

$$\frac{1}{\bar{\lambda}_K(\mathscr{D}_i + \tau)} \|B_i \mathscr{X} - \mathscr{B}_i \mathscr{X}\| \leq \frac{1}{2\bar{\lambda}_K \sqrt{\delta_A}(\delta_B + \tau)} \sqrt{K \ln(4N_B/\epsilon)} \tag{D.12}$$

with probability at least $1-2\epsilon$ for all $i \in V_B$. Combining equation (D.10),equation (D.11) and equation D.12, finally we have

$$\|Y_i - \mathscr{Y}_i\| \leq \frac{c_1}{2} \max\Big\{ \frac{\sqrt{K \ln N_A/\epsilon}}{\delta_A \sqrt{\delta_B + \tau} \bar{\lambda}_K^2}, \frac{\sqrt{K \ln N_B/\epsilon}}{\sqrt{\delta_A}(\delta_B + \tau)\bar{\lambda}_K} \Big\}, \qquad \forall i \in V_B. \tag{D.13}$$

Recall that $\mathscr{Y}_i = \frac{\mathscr{D}_i}{\mathscr{D}_i + \tau}[Z_B]_i V^{-1/2} U$ and hence $\|\mathscr{Y}_i\| \geq \frac{\mathscr{D}_i}{\mathscr{D}_i + \tau} V_{max}^{-1/2} \geq \frac{\delta_B}{\delta_B + \tau} V_{max}^{-1/2}$. Applying lemma D.1 gives

$$\|Y_i^* - \mathscr{Y}_i^*\| \leq c_1 \max\Big\{ \frac{\sqrt{(\delta_B + \tau) K V_{max} \ln N_A/\epsilon}}{\delta_A \delta_B \bar{\lambda}_K^2}, \frac{\sqrt{K V_{max} \ln N_B/\epsilon}}{\sqrt{\delta_A} \delta_B \bar{\lambda}_K} \Big\}, \qquad \forall i \in V_B,$$

with probability at least $1 - \epsilon$. Combining all rows in $Y - \mathscr{Y}$ reveals the desired result. □

Proof of Theorem 5.7

Proof. Recall that the set of misclustered nodes is defined as:

$$\mathscr{M} = \{i : \exists j \neq i, s.t. \|C_i - \mathcal{C}_i\|_2 > \|C_i - \mathcal{C}_j\|_2\}.$$

Note that Lemma 5.2 and Lemma 5.3 implies that the population centroid corresponding to i'th row of $\mathscr{F}$,

$$\mathcal{C}_i = Z_i U.$$

Since all population centroids are of unit length and are orthogonal to each other, a simple calculation gives a sufficient condition for one observed centroid to be closest to the population centroid:

$$\|C_i - \mathcal{C}_i\|_2 < 1/\sqrt{2} \Rightarrow \|C_i - \mathcal{C}_i\|_2 < \|C_i - \mathcal{C}_j\|_2 \quad \forall Z_j \neq Z_i.$$

Define the following set of nodes that do not satisfy the sufficient condition,

$$\mathscr{U} = \{i : \|C_i - \mathcal{C}_i\|_2 \geq 1/\sqrt{2}\}.$$

The mis-clustered nodes $\mathscr{M} \in \mathscr{U}$.

Define $Q \in \mathcal{R}^{N\times K}$, where the i'th row of Q is C_i, the observed centroid of node i from k-means. By definition of k-means, we have

$$\|F - Q\|_2 \leq \|F - \mathscr{F}\|_2.$$

By triangle inequality,

$$\|Q - ZU\|_2 = \|Q - \mathscr{F}\|_2 \leq \|X^*_\tau - Q\|_2 + \|F - \mathscr{F}\|_2 \leq 2\|F - \mathscr{F}\|_2.$$

The misclustering rate is bounded as follows,

$$\begin{aligned}
\frac{|\mathscr{M}|}{N} &\leq \frac{|\mathscr{U}|}{N} = \frac{1}{N}\sum_{i\in\mathscr{U}} 1 \\
&\leq \frac{2}{N}\sum_{i\in\mathscr{U}} \|C_i - Z_iU\|_2^2 \\
&\leq \frac{2}{N}\|Q - ZU\|_F^2 \\
&\leq \frac{8}{N}(\|X^* - \mathscr{X}^*\|_F^2 + \|Y^* - \mathscr{Y}^*\|_F^2)
\end{aligned}$$

Finally, applying the result from Theorem 5.5 and the assumption that $\bar{\lambda}_K$ is bounded below by positive constant, we have

$$\frac{|\mathscr{M}|}{N} = O_p\bigg(\frac{N_A}{N} \bullet \frac{KV_{max}\ln N_A}{N_A\delta_A^2} + \frac{N_B}{N} \bullet \max\Big\{\frac{(\delta_B+\tau)KV_{max}\ln N_A}{\delta_A^2\delta_B^2}, \frac{KV_{max}\ln N_B}{\delta_A\delta_B^2}\Big\}\bigg).$$

□

Appendix E

Appendix for Chapter 6

Proof of Theorem 6.4

Proof. Recall that the transitivity ratio of A is

$$trans(A) = \frac{\text{number of closed triplets in } A}{\text{number of connected triples of vertices in } A}.$$

Both the numerator and the denominator of the transitivity ratio have other formulations that suggest how they can be computed.

$$\begin{aligned} \text{number of closed triplets in } A &= 6 \times \text{Number of triangles in } A \\ &= \text{trace}(AAA) \end{aligned}$$

$$\begin{aligned} \text{number of connected triples of vertices in } A &= 2 \times \text{Number of 2-stars in } A \\ &= 2\sum_j \binom{d_j}{2} \\ &= \sum_j d_j^2 - d_j. \end{aligned}$$

For ease of notation, define $X_n = trace(AAA)$ and $Y_n = \sum_i d_i^2 - d_i$. So, $trans(A) = X_n/Y_n$. To show that transitivity converges to zero, use

$$P\left(\frac{X_n}{Y_n} > \epsilon\right) \leq \frac{E(X_n/Y_n)}{\epsilon}$$

and the following Lemma.

Lemma E.1. *If $\lambda_n = o(n)$, then there exists a sequence f_n such that $E(X_n) = o(f_n)$ and*

$$P(Y_n \geq f_n) \to 1.$$

Using Lemma E.1 and fact that $X_n/Y_n \leq 1$ a.s.,

$$\begin{aligned} E\frac{X_n}{Y_n} &\leq E\left(\frac{X_n}{f_n}1\{Y_n > f_n\} + 1\{Y_n < f_n\}\right) \\ &\leq \frac{E(X_n)}{f_n} + P(Y_n < f_n) \\ &\to 0. \end{aligned}$$

Now, to prove Lemma E.1. For ease of notation, define $d = \sum_i d_i$. From Bickel, Levina, Chen, define $\hat{\rho} = \frac{d}{2n(n-1)}$. They show that $\hat{\rho}/\rho_n \xrightarrow{P} 1$, where $\rho_n = P(A_{12} = 1)$. So, this converges to zero:

$$P\left(\frac{d}{2n(n-1)\rho_n} < 1/2\right) = P\left(d < n(n-1)\rho_n\right)$$

Define $M_n = n(n-1)\rho_n$. Then, $P\left(d > M_n\right) \to 1$. Define $f_n = M_n^2/n - M_n$. Notice that

$$min_{\sum_i d_i = m} \sum_i d_i^2 - d_i \geq n((m/n)^2 - m/n) = m^2/n - m.$$

Putting these pieces together,

$$\begin{aligned} P(Y_n \geq f_n) &= \int_\Omega 1\{Y_n \geq f_n\} dP \\ &\geq \int_{d > M_n} 1\{\sum_i d_i^2 - d \geq f_n\} dP \\ &= \sum_{m > M_n} \int_{d=m} 1\{\sum_i d_i^2 - d \geq f_n\} dP \\ &\geq \sum_{m > M_n} \int_{d=m} 1\{m^2/n - m \geq f_n\} dP \\ &= \sum_{m > M_n} \int_{d=m} dP \\ &= P(d > M_n) \to 1. \end{aligned}$$

The last piece is to show that $E(X_n) = o(f_n)$. From the definition of f_n and the fact that $\rho_n = \lambda_n/n$,

$$\begin{aligned} f_n &= \frac{(n(n-1)\rho_n)^2}{n} - n(n-1)\rho_n \\ &= n(n-1)^2(\lambda_n/n)^2 - n(n-1)\lambda_n/n \\ &= \lambda_n(n-1)\left(\frac{n-1}{n}\lambda_n - 1\right) \\ &\to \lambda_n^2 n \end{aligned}$$

Define

$$S_{12} = \sum_i A_{i1}A_{2i}$$

as the number of two stars with nodes 1 and 2 as end points. Then, under the assumption that

$$p_{\max} = o\left(\frac{P(A_{13}=1)}{P(A_{13}=1|A_{23}=1)}\right),$$

it follows that

$$\begin{aligned} E(X_n) &= p_{ct}n(n-1)E(S_{12}) \\ &\leq p_{\max}n^2(nE(A_{13}A_{23})) \\ &= p_{\max}n^3E(A_{13}|A_{23}=1)E(A_{23}=1) \\ &= o(\rho_n^2 n^3) \\ &= o(f_n). \end{aligned}$$

□

Proof of Theorem 6.8:

Proof. Let $r = c_0/n$ and let p be fixed.

Number of triangles: Let Δ_n denote the number of triangles. Notice that there are three types of triangles: (1) let Δ^i denote the number of triangles with all nodes in block i; (2) let Δ_{21} denote the number of triangles with 2 nodes in the same block and one node in a separate block; (3) let Δ_{111} denote the number of triangles with

nodes in three separate blocks.

$$E(\Delta_{21}) = K(K-1)\binom{s}{2}spr^2 = K(K-1)\binom{s}{2}spc_0^2/(s^2K^2) \leq p(s-1)/2,$$

$$E(\Delta_{111}) = \binom{K}{3}s^3r^3 \leq K^3s^3c_0^3/(6s^3K^3) = c_0^3/6.$$

By the Markov inequality, $\Delta_{21}/K \xrightarrow{P} 0, \Delta_{111}/K \xrightarrow{P} 0$. Finally, Δ^i are iid. So, by LLN, their average converges in probability to their expectation. Putting these pieces together with Slutsky's theorem, the number of triangles over K is,

$$\frac{1}{K}\Delta_n \xrightarrow{P} E(\Delta^i) = \binom{s}{3}p^3.$$

Number of two stars: Let S_n be the number of two-stars. Define the events $B = \{|S_n - E(S_n)| > t\}$ and $A = \{\text{Maximum Degree} \leq M\}$

$$P(B) = P(BA) + P(BA^c) \leq P(BA) + P(A^c)$$

Apply the bounded difference inequality within the set BA. Define $A_i \in \{0,1\}^{n-i}$ for $i = 1, \ldots, n-1$ as the ith row of the upper triangle of the adjacency matrix A. To bound the bounded difference constant, first notice that $c_1 \geq c_i$ for all i. Moreover, we have

$$c_1 \leq 3M^2.$$

This is because node 1 belongs to at most $3\binom{M}{2}$ triplets. By changing the edges of node 1, S_n can increase or decrease by at most $3\binom{M}{2}$. By the bounded difference inequality,

$$P(BA) \leq 2\exp\left(-\frac{t^2}{n9M^4}\right).$$

Choose $M = \log n$ and $t = K\epsilon$ for any $\epsilon > 0$, we have $P(BA) \to 0$. More over, by concentration inequality,

$$P(A^c) = P(\cup_i\{d_i \geq M\}) \leq nP(d_1 \geq M) \to 0$$

Therefore, we have

$$S_n/K \xrightarrow{P} E(S_n)/K.$$

Notice that $E(S_n)/K$ is equal to the expected number of two-stars whose center is in

the first block. So,

$$\begin{aligned} E(S_n)/K &= s\left(\binom{s-1}{2}p^2 + (s-1)(n-s)pr + \binom{n-s}{2}r^2\right) \\ &\to s\left(\binom{s-1}{2}p^2 + (s-1)pc_0 + c_0^2/2\right) \end{aligned}$$

Finally,

$$TranRatio(A) = \frac{3 \times \text{number of triangles}}{\text{number of 2-stars}} \xrightarrow{P} \frac{3E\Delta_n}{ES_n}.$$

□

Proof of Theorem 6.11:

Proof. Define the following events

$$\begin{aligned} B_\alpha &= \{S_* \text{ and } S_*^c \text{ are separated with cutting level } \alpha\}, \\ C_\alpha &= \{S_* \text{ is clustered within one block with cutting level } \alpha\}, \\ D_\alpha &= \{\text{Every pair of nodes in } S_* \text{ have at least } \alpha \text{ common neighbor.}\} \end{aligned}$$

If both events B_α and C_α are satisfied, then for any $i \in S_*$, LocalTrans(A, i, α) recovers block S_* correctly. Events D_α implies that S_* is clustered within one block with cutting level α, that is $D_\alpha \in C_\alpha$. To see this, assume the contrary, then there exists a partition $S_* = S_1 \cup S_2$, such that for any $j \in S_1, k \in S_2, T(j,k) < \alpha$. However, D_α implies that S_* is connected, hence there exists $u \in S_1, v \in S_2$, such that $A(u,v) = 1$. Moreover, D_α also implies that u, v have at least α common neighbors. Hence, $T(u,v) \geq \alpha$. This is a contradiction.

The following lemma leads to the desired results.

Lemma E.2. *Under the conditions above,*

$$\begin{aligned} P(B_1^c) &= O(p_{out}^2 ns(s+\lambda)), \\ P(B_2^c) &= O(p_{out}^3 ns(s+\lambda)^2), \\ P(D_\alpha^c) &\leq \frac{1}{2}s^2 \sum_{k=0}^{\alpha-1} \binom{s-2}{k} (1-p_{in}^2)^{s-2-k}. \end{aligned}$$

By Lemma E.2, we have

$$\begin{aligned} P(\text{correctly clustering } S_* \text{ with cutting level 1}) &\geq P(B_1 \cap C_1) \\ &= 1 - P(B_1^c \cup C_1^c) \\ &\geq 1 - P(B_1^c) + P(C_1^c) \\ &\geq 1 - P(B_1^c) + P(D_1^c) \\ &= 1 - \left(\frac{1}{2}s^2(1-p_{in}^2)^{s-2} + O(p_{out}^2 ns(s+\lambda))\right) \end{aligned}$$

$$\begin{aligned} P(\text{correctly clustering } S_* \text{ with cutting level 2}) &\geq P(B_2 \cap C_2) \\ &= 1 - P(B_2^c \cup C_2^c) \\ &\geq 1 - P(B_2^c) + P(C_2^c) \\ &\geq 1 - P(B_2^c) + P(D_2^c) \\ &= 1 - \left(s^3(1-p_{in}^2)^{s-3} + O(p_{out}^3 ns(s+\lambda)^2)\right) \end{aligned}$$

□

Proof of Lemma E.2:

Proof.

$$\begin{aligned} P(B_1^c) &= P(\text{There exists at least two nodes } i \in S_* \text{ and } j \in S_*^c, \text{ such that } T_{ij} \geq 1) \\ &= P(\bigcup_{i \in S_*} \bigcup_{j \in S_*^c} \bigcup_{k \in S_* \cup S_*^c, k \neq i,j} \{A_{ij}A_{jk}A_{ki} = 1\}) \\ &\leq sn[P(\bigcup_{k \in S_*} \{A_{ij}A_{jk}A_{ki} = 1\}) + P(\bigcup_{k \in S_*^c} \{A_{ij}A_{jk}A_{ki} = 1\})] \\ &\leq sn(sp_{out}^2 + np_{out}^2 \frac{\lambda}{n}) \\ &= O(p_{out}^2 ns(s+\lambda)) \end{aligned}$$

$$
\begin{aligned}
P(B_2^c) &= P(\text{There exists at least two nodes } i \in S_* \text{ and } j \in S_*^c, \text{ such that } T_{ij} \geq 2) \\
&= P(\bigcup_{i \in S_*} \bigcup_{j \in S_*^c} \bigcup_{k<l \in S_* \cup S_*^c, k,l \neq i,j} \{A_{ij}A_{jk}A_{ki}A_{jl}A_{li} = 1\}) \\
&\leq snP(\bigcup_{k,l \in S_*} \cup \bigcup_{k \in S_*, l \in S_*^c} \cup \bigcup_{k,l \in S_*^c} \{A_{ij}A_{jk}A_{ki}A_{jl}A_{li} = 1\}) \\
&\leq sn(\frac{1}{2}s^2 p_{out}^3 + \frac{1}{2}n^2 p_{out}^3 (\frac{\lambda}{n})^2 + nsp_{out}^3 \frac{\lambda}{n}) \\
&= O(p_{out}^3 ns(s+\lambda)^2)
\end{aligned}
$$

$$
\begin{aligned}
P(D_\lambda^c) &= P(\text{There exists at least two nodes } i, j \in S_*, \text{ such that } i \text{ and } j \text{ has less than } \alpha \text{ neighbors}) \\
&= P\left(\bigcup_{i,j \in S_*} \{i \text{ and } j \text{ has less than } \alpha \text{ neighbors}\}\right) \\
&\leq \frac{1}{2}s^2 \sum_{k=0}^{\alpha-1} \binom{s-2}{k} (1-p_{in}^2)^{s-2-k}
\end{aligned}
$$

□

Proof of Theorem 6.13:

Proof. In the proof, we assume that $i, j \in S_*$, $P(A_{ij} = 1) = p_{in}$. The proof can be easily extended to the case where $P(A_{ij} = 1) \geq p_{in}$. First, we prove that $P(\max_{i \in S_*} D_{ii} \geq 2\mathbb{E}D_{11})$ is well bounded with some non-vanishing probability. $\forall i \in S_*$,

$$\mathbb{E}D_{ii} = \mathbb{E}D_{11} = \sum_{j \in S_*, j \neq i} \mathbb{E}A_{ij} + \sum_{j \in S_*^c} \mathbb{E}A_{ij} = (s-1)p_{in} + \lambda.$$

$\forall i \in S_*,\ \forall \epsilon > 0$,

$$P(D_{ii} \geq \mathbb{E}D_{ii} + \epsilon) \leq \exp\left\{-\frac{\epsilon}{2(\mathbb{E}D_{ii} + \epsilon/3)}\right\}.$$

Take $\epsilon = \mathbb{E}D_{11}$, and take union bound for all $i \in S_*$, we have

$$P(\max_{i \in S_*} D_{ii} \geq 2\mathbb{E}D_{11}) \leq s \exp\left\{-\frac{3}{8}\mathbb{E}D_{11}\right\} = s \exp\{-\frac{3}{8}[(s-1)p_{in} + \lambda]\}.$$

Let O denote the set $\{D_{ii} \leq 2\mathbb{E}D_{11}, \forall i \in S_*\}$. Then within the set O, by the same argument from the proof of Theorem 3, we have that with probability at least $\frac{1}{2}s^2(1-p_{in}^2)^{s-2}$, S_* is clustered within one block by LocalTrans(L_τ, i, cut) for any $i \in S_*$ with cut

$$cut = (2\mathbb{E}D_{11} + \tau)^{-3} = (2(s-1)p_{in} + 2\lambda + \tau)^{-3}.$$

Second part proves that $\forall i \in S_*, j \in S_*^c, P(T_{ij} \geq cut)$ is $o(1)$. Notice that for any $j \in S_*^c$, the (j,j)th element of D_τ (denote as D_{jj}^τ) is $d_j^* + \sum_{i \in S_*} A_{ij} + \tau$, so we have

$$D_{jj}^\tau \geq d_j^* + \tau, \quad \forall j \in S_*^c$$

For any $i \in S_*$, we have

$$D_{ii}^\tau \geq \tau, \quad \forall i \in S_*$$

$\forall i \in S_*,\ j \in S_*^c$,

$$\begin{aligned}
P(T_{ij} \geq cut) &\leq \frac{d_j^*}{n} P(T_{ij} \geq\ cut | A_{ij} = 1) \\
&= \frac{d_j^*}{n} P\left(\frac{1}{D_{ii}^\tau D_{jj}^\tau} \sum_{k=1}^n \frac{A_{ik}A_{kj}}{D_{kk}^\tau} \geq cut \right) \\
&\leq \frac{d_j^*}{n} P\left(\frac{1}{\tau(d_j^* + \tau)} \sum_{k=1}^n \frac{A_{ik}A_{kj}}{D_{kk}^\tau} \geq cut \right) \\
&\leq \frac{d_j^*}{n} \left[P\left(\sum_{k \in S_*} A_{ik}A_{kj} \geq \tau^2(d_j^* + \tau)cut/2 \right) + \right. \\
&\qquad \left. P\left(\sum_{k \in S_*^c} \frac{A_{ik}A_{kj}}{d_k + \tau} \geq \tau(d_j^* + \tau)cut/2 \right) \right]
\end{aligned}$$

For the first term, when n large,

$$\begin{aligned}
P(\sum_{k \in S_*} A_{ik}A_{kj} \geq \tau^2(d_j^* + \tau)cut/2) &\leq P(\sum_{k \in S_*, k \neq i} A_{ik}A_{kj} > 0) \\
&\leq 1 - (1 - p_{in}\frac{d_j^*}{n})^{s-1} \\
&\leq p_{in}(s-1)\frac{d_j^*}{n}
\end{aligned}$$

On the other hand, notice that $\{A_{ik}A_{kj}, k \in S_*/\{i\}\}$ are independent random variables with $A_{ik}A_{kj} \sim Ber(p_{in}\frac{d_j^*}{n})$ by the assumption. $E[\sum_{k\in S_*, k\neq i} A_{ik}A_{kj}] = p_{in}(s-1)\frac{d_j^*}{n}$. By concentration inequality, when n is sufficiently large (independent of j), we have

$$\begin{aligned} P(\sum_{k\in S_*} A_{ik}A_{kj} \geq \tau^2(d_j^*+\tau)cut/2) &\leq P\left(\sum_{k\in S_*} A_{ik}A_{kj} \geq p_{in}(s-1)\frac{d_j^*}{n} + \right. \\ &\qquad \left. d_j^*\left(\tau^2 cut/2 - \frac{p_{in}(s-1)}{n}\right)\right) \\ &\leq \exp\left(-\frac{c_1^2(d_j^*)^2}{2c_2 d_j^*/n + 2c_1 d_j^*/3}\right) \\ &= \exp(-c_1 d_j^*) \end{aligned}$$

where $c_1 = \tau^2 cut/3$.

For the second term, without loss of generality, assume that $A_{1j} = A_{2j} = ... = A_{d_j^*,j} = 1, A_{k,j} = 0, \forall k > d_j^*$. Notice that $\{A_{ik}, k = 1, 2, ..., n, k \neq i\}$ are independent random variables with $A_{ik} \sim Ber(\frac{d_k^*}{n})$. Applying concentration inequality on the sequence $\{A_{ik}, k = 1, 2, ..., d_j^*\}$, with $a_k = \frac{1}{d_k^*}$. Define $X = \sum_{k=1}^{d_j^*} a_k A_{ik}$, then $\mathbb{E}X = \frac{d_j^*}{n}$, and

$$v = \sum_{k=1}^{d_j^*} a_k^2 \mathbb{E}A_{ik} = \frac{1}{n}\sum_{k=1}^{d_j^*}\frac{1}{d_j^*} \leq \frac{d_j^*}{n}.$$

When n is sufficiently large, we have,

$$\begin{aligned} P\left(\sum_{k\in S_*^c} \frac{A_{ik}A_{kj}}{d_k+\tau} \geq \tau(d_j^*+\tau)cut/2\right) &\leq P\left(\sum_{k=1}^{d_j^*}\frac{A_{ik}}{d_k} \geq \tau(d_j^*+\tau)cut/2\right) \\ &\leq P\left(\sum_{k=1}^{d_j^*}\frac{A_{ik}}{d_k} \geq \frac{d_j^*}{n} + d_j^*(\tau cut/2 - \frac{1}{n})\right) \\ &\leq \exp\left(-\frac{c^2(d_j^*)^2}{2(v + cd_j^*/3)}\right) \\ &\leq \exp\left(-\frac{c^2(d_j^*)^2}{2(d_j^*/n + cd_j^*/3)}\right) \\ &= \exp(-cd_j^*) \end{aligned}$$

where $c = \tau cut/3$.

On the other hand,

$$\begin{aligned} P\left(\sum_{k\in S_*^c} \frac{A_{ik}A_{kj}}{d_k+\tau} \geq \tau(d_j^*+\tau)cut/2\right) &\leq P\left(\sum_{k=1}^{d_j^*} \frac{A_{ik}}{d_k+\tau} \geq \tau(d_j^*+\tau)cut/2\right) \\ &\leq P\left(\sum_{k=1}^{d_j^*} A_{ik} > 0\right) \\ &\leq 1-\left(1-\frac{d_j^*}{n}\right)^{d_j^*} \\ &\leq \frac{(d_j^*)^2}{n} \end{aligned}$$

To sum up,when n is sufficiently large, we have that $\forall i \in S_*,\ j \in S_*^c$

$$P(T_{ij} \geq cut) \leq \frac{d_j^*}{n} \min\left\{\frac{(d_j^*)^2}{n} + p_{in}(s-1)\frac{d_j^*}{n},\ \ 2e^{-cd_j^*}\right\}.$$

where $cut = (2\mathbb{E}D_{11}+\tau)^{-3} = (2(s-1)p_{in}+2\lambda+\tau)^{-3}$, and $c = \tau cut/3$.

Next we show that, for any $i \in S_*, j \notin S_*$, $P(T_{ij} \geq cut) = O(n^{3\epsilon-2})$, where $\epsilon > \log_n(\frac{1}{c})$.

Case 1: $d_j^* < n^\epsilon$.

$$\begin{aligned} nP(T_{ij} \geq cut) &\leq \min\left\{\frac{(d_j^*)^3}{n} + p_{in}s\frac{(d_j^*)^2}{n},\ 2d_j^*e^{-cd_j^*}\right\} \\ &\leq \frac{(d_j^*)^3}{n} + p_{in}s\frac{(d_j^*)^2}{n} \\ &\leq \frac{n^{3\epsilon}}{n} + p_{in}s\frac{n^{2\epsilon}}{n} \\ &= O(n^{3\epsilon-1}) \end{aligned}$$

Case 2: $d_j^* \geq n^\epsilon$. Notice that the derivate of $u(x) = 2d_j^*\exp(-cd_j^*)$ is $(1-$

$cx)\exp(-cx)$. So, if $n^\epsilon > 1/c$ then $u(d_j^*) \le u(n^\epsilon) = 2n^\epsilon \exp(-cn^\epsilon)$.

$$\begin{aligned} nP(T_{ij} \ge cut) &\le \min\left\{ \frac{(d_j^*)^3}{n} + p_{in} s \frac{(d_j^*)^2}{n},\ 2d_j^* \exp(-cd_j^*) \right\} \\ &\le 2d_j^* \exp(-cd_j^*) \\ &\le 2n^\epsilon \exp(-cn^\epsilon) \\ &= o(n^{-1}) \end{aligned}$$

So, independent of d_j^*,

$$P(T_{ij} \ge cut) = O(n^{3\epsilon-2})$$

Putting the pieces together,

$$\begin{aligned} P(S_* \text{ and } S_*^c \text{ are not separated }) &\le s \sum_{j \in S_*^c} \frac{d_j^*}{n} \min\left\{ \frac{(d_j^*)^2}{n} + p_{in} s \frac{d_j^*}{n},\ 2e^{-cd_j^*} \right\} \\ &\le snO(n^{3\epsilon-2}) \\ &= O(n^{3\epsilon-1}) \end{aligned}$$

Finally, recall that $O = \{D_{ii} \le 2\mathbb{E}D_{11}, \forall i \in S_*\}$, we have that for any $i \in S_*$,

$$\begin{aligned} P(\{\text{LocalTrans}(L_\tau, i, cut) \text{ returns } S_*\}) &\ge 1 - \frac{1}{2}s^2(1-p_{in}^2)^{s-2} - O(n^{3\epsilon-1}) - P(O^c) \\ &\ge 1 - \Bigg(\frac{1}{2}s^2(1-p_{in}^2)^{s-2} + \\ &\qquad s\exp\{-\frac{3}{8}[(s-1)p_{in} + \lambda)\} + O(n^{3\epsilon-1}) \Bigg). \end{aligned}$$

□

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.

Adamic, Lada A, and Bernardo A Huberman. 2000. Power-law distribution of the world wide web. *Science* 287(5461):2115–2115.

Aldous, David J. 1981. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* 11(4):581–598.

Alon, Noga, Raphael Yuster, and Uri Zwick. 1997. Finding and counting given length cycles. *Algorithmica* 17(3):209–223.

Ames, Brendan PW, and Stephen A Vavasis. 2010. Convex optimization for the planted k-disjoint-clique problem. *arXiv preprint arXiv:1008.2814*.

Amini, Arash A, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. 2012. Pseudo-likelihood methods for community detection in large sparse networks. *arXiv preprint arXiv:1207.2340*.

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.

Andersen, Reid, and Fan Chung. 2007. Detecting sharp drops in pagerank and a simplified local partitioning algorithm. *Theory and Applications of Models of Computation* 1–12.

Andersen, Reid, Fan Chung, and Kevin Lang. 2006. Local graph partitioning using pagerank vectors. In *Foundations of computer science, 2006. focs'06. 47th annual ieee symposium on*, 475–486. IEEE.

Andersen, Reid, and Yuval Peres. 2009. Finding sparse cuts locally using evolving sets. In *Proceedings of the 41st annual acm symposium on theory of computing*, 235–244. ACM.

Anderson, Carolyn J, Stanley Wasserman, and Katherine Faust. 1992. Building stochastic blockmodels. *Social networks* 14(1):137–161.

Banerjee, A., I. Dhillon, J. Ghosh, S. Merugu, and D.S. Modha. 2004. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining*, 509–514. ACM.

Belabbas, Mohamed-Ali, and Patrick J Wolfe. 2009. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences* 106(2):369–374.

Bickel, P., D. Choi, X. Chang, and H. Zhang. 2012. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Arxiv preprint arXiv:1207.0865*.

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.

Bickel, P.J., A. Chen, and E. Levina. 2011. The method of moments and degree distributions for network models. *The Annals of Statistics* 39(5):38–59.

Bisson, G., and F. Hussain. 2008. Chi-sim: A new similarity measure for the co-clustering task. In *Machine learning and applications, 2008. icmla'08. seventh international conference on*, 211–217. IEEE.

Borchers, Hans W. 2012. [r] k-means++. `https://stat.ethz.ch/pipermail/r-help/2012-January/300051.html`.

Celisse, Alain, J-J Daudin, and Laurent Pierre. 2011. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *arXiv preprint arXiv:1105.3288*.

Channarond, Antoine, Jean-Jacques Daudin, and Stéphane Robin. 2011. Classification and estimation in the stochastic block model based on the empirical degrees. *arXiv preprint arXiv:1110.6517*.

Chaudhuri, K., F. Chung, and A. Tsiatas. 2012. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research* 1–23.

Chen, A., A.A. Amini, P.J. Bickel, and E. Levina. 2012a. Fitting community models to large sparse networks. *Arxiv preprint arXiv:1207.2340*.

Chen, Yudong, Sujay Sanghavi, and Huan Xu. 2012b. Clustering sparse graphs. *arXiv preprint arXiv:1210.3335*.

Choi, D.S., P.J. Wolfe, and E.M. Airoldi. 2012. Stochastic blockmodels with a growing number of classes. *Biometrika* 99(2):273–284.

Chung, Fan, and Mary Radcliffe. 2011. On the spectra of general random graphs. *the electronic journal of combinatorics* 18(P215):1.

Chung, Fan Rong K, and Linyuan Lu. 2006. *Complex graphs and networks*. 107, American Mathematical Soc.

Chung, F.R.K. 1997. *Spectral graph theory*. Amer Mathematical Society.

Clauset, Aaron. 2005. Finding local community structure in networks. *Physical Review E* 72(2):026132.

Coja-Oghlan, Amin, and André Lanka. 2009. Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics* 23(4):1682–1714.

Csardi, Gabor, and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal* Complex Systems:1695.

Dasgupta, Anirban, John E Hopcroft, and Frank McSherry. 2004. Spectral analysis of random graphs with skewed degree distributions. In *Foundations of computer science, 2004. proceedings. 45th annual ieee symposium on*, 602–610. IEEE.

Dhillon, I.S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining*, 269–274. ACM.

Diaconis, Persi, and Svante Janson. 2007. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*.

Donath, W.E., and A.J. Hoffman. 1973. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development* 17(5):420–425.

Drineas, Petros, and Michael W Mahoney. 2005. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research* 6:2153–2175.

Dunbar, R.I.M. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* 22(6):469–493.

Fan, J., Y. Fan, and J. Lv. 2008. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* 147(1):186–197.

Fiedler, M. 1973. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal* 23(2):298–305.

Fishkind, Donniell E, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. 2013. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications* 34(1):23–39.

Flynn, C.J., and P.O. Perry. 2012. Consistent biclustering. *Arxiv preprint arXiv:1206.6927*.

Fowlkes, Charless, Serge Belongie, Fan Chung, and Jitendra Malik. 2004. Spectral grouping using the nystrom method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26(2):214–225.

Freitag, D. 2004. Trained named entity recognition using distributional clusters. In *Proceedings of emnlp*, vol. 4, 262–269.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Giesen, Joachim, and Dieter Mitsche. 2005. Reconstructing many partitions using spectral techniques. In *Fundamentals of computation theory*, 433–444. Springer.

Goodman, Leo A. 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review / Revue Internationale de Statistique* 54(3): pp. 243–270.

Guttman, L. 1959. Metricizing rank-ordered or unordered data for a linear factor analysis. *Sankhyā: The Indian Journal of Statistics (1933-1960)* 21(3/4):257–268.

Hartigan, J.A. 1972. Direct clustering of a data matrix. *Journal of the American Statistical Association* 123–129.

Hirschfeld, HO. 1935. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society* 31(04):520–524.

Hoff, P.D. 2009. Multiplicative latent factor models for description and prediction of social networks. *Computational & Mathematical Organization Theory* 15(4):261–272.

Hoff, P.D., A.E. Raftery, and M.S. Handcock. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical Association* 97(460): 1090–1098.

Holland, P., K.B. Laskey, and S. Leinhardt. 1983. Stochastic blockmodels: Some first steps. *Social Networks* 5:109–137.

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

Holmes, S. 2006. Multivariate analysis: The french way. *Festschrift for David Freedman*.

Hoover, Douglas N. 1979. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*.

Jin, Jiashun. 2015. Fast community detection by score. *The Annals of Statistics* 43(1):57–89.

Joseph, Antony, and Bin Yu. 2014. Impact of regularization on spectral clustering. *arXiv preprint arXiv:1312.1733*.

Kallenberg, O. 2005. *Probabilistic symmetries and invariance principles*. Springer Verlag.

Karrer, Brian, and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1):016107.

Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46(5):604–632.

Krzakala, Florent, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. 2013. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences* 110(52):20935–20940.

Kumar, Amit, Yogish Sabharwal, and Sandeep Sen. 2004. A simple linear time (1+ ε)-approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings-annual symposium on foundations of computer science*, 454–462. IEEE.

Lei, Jing, and Alessandro Rinaldo. 2015. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* 43(1):215–237.

Leskovec, J., K.J. Lang, A. Dasgupta, and M.W. Mahoney. 2008. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on world wide web*, 695–704. ACM.

Leskovec, Jure, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6(1):29–123.

Liao, Chung-Shou, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. 2009. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics* 25(12):i253–i258.

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Madeira, S.C., and A.L. Oliveira. 2004. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 1(1):24–45.

Madeira, S.C., M.C. Teixeira, I. Sa-Correia, and A.L. Oliveira. 2010. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 7(1):153–165.

Mahoney, Michael W. 2012a. Randomized algorithms for matrices and data. *Advances in Machine Learning and Data Mining for Astronomy, CRC Press, Taylor & Francis Group, Eds.: Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, Ashok N. Srivastava, p. 647-672* 1:647–672.

Mahoney, Michael W, and Lorenzo Orecchia. 2010. Implementing regularization implicitly via approximate eigenvector computation. *arXiv preprint arXiv:1010.0703*.

Mahoney, M.W. 2012b. Approximate computation and implicit regularization for very large-scale data analysis. *Arxiv preprint arXiv:1203.0786*.

McSherry, F. 2001. Spectral partitioning of random graphs. In *Foundations of computer science, 2001. proceedings. 42nd ieee symposium on*, 529–537. IEEE.

Negahban, S., P. Ravikumar, M.J. Wainwright, and B. Yu. 2010. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Arxiv preprint arXiv:1010.2731*.

Newman, Mark EJ, and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69(2):026113.

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.

Oymak, Samet, and Babak Hassibi. 2011. Finding dense clusters via" low rank+ sparse" decomposition. *arXiv preprint arXiv:1104.5186*.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. The pagerank citation ranking: Bringing order to the web.

Perry, Patrick O, and Michael W Mahoney. 2011. Regularized laplacian estimation and fast eigenvector approximation. *arXiv preprint arXiv:1110.1757*.

Priebe, Carey E, John M Conroy, David J Marchette, and Youngser Park. 2005. Scan statistics on enron graphs. *Computational & Mathematical Organization Theory* 11(3):229–247.

Qin, Tai, and Karl Rohe. 2013. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in neural information processing systems*, 3120–3128.

Ravikumar, P., M.J. Wainwright, G. Raskutti, and B. Yu. 2011. High-dimensional covariance estimation by minimizing ℓ_1 penalized log-determinant divergence. *Electronic Journal of Statistics* 5:935–980.

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

Rohe, Karl, and Tai Qin. 2013. The blessing of transitivity in sparse and stochastic networks. *arXiv preprint arXiv:1307.2302*.

Rohe, Karl, Tai Qin, and Haoyang Fan. 2014. The highest dimensional stochastic blockmodel with a regularized estimator. *Statistica Sinica* 24:1771–1786.

Rohe, Karl, Tai Qin, and Bin Yu. 2015. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*.

Rohe, Karl, and Bin Yu. 2012. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*.

Rohwer, R., and D. Freitag. 2004. Towards full automation of lexicon construction. In *Proceedings of the hlt-naacl workshop on computational lexical semantics*, 9–16. Association for Computational Linguistics.

Rukhin, Andrey, and Carey E Priebe. 2012. On the limiting distribution of a graph scan statistic. *Communications in Statistics-Theory and Methods* 41(7):1151–1170.

Sarkar, Purnamrita, and Peter J Bickel. 2013. Role of normalization in spectral clustering for stochastic blockmodels. *arXiv preprint arXiv:1310.1495*.

Schönemann, Peter H. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31(1):1–10.

Spielman, Daniel A, and Shang-Hua Teng. 2008. A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning. *arXiv preprint arXiv:0809.3232*.

Steinhaus, H. 1956. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci* 1:801–804.

Sussman, Daniel L, Minh Tang, Donniell E Fishkind, and Carey E Priebe. 2012a. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* 107(499):1119–1128.

Sussman, D.L., M. Tang, D.E. Fishkind, and C.E. Priebe. 2012b. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* 107(499):1119–1128.

Tanay, A., R. Sharan, M. Kupiec, and R. Shamir. 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America* 101(9):2981.

Tanay, A., R. Sharan, and R. Shamir. 2005. Biclustering algorithms: A survey. *Handbook of computational molecular biology* 9:26–1.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99(10):6567.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4):395–416.

Wang, H., M. Tang, Y. Park, and C. E. Priebe. 2013. Locality statistics for anomaly detection in time series of graphs. *ArXiv e-prints*. 1306.0267.

Wang, Y.J., and G.Y. Wong. 1987. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* 82(397):8–19.

Wasserman, Stanley, Katherine Faust, and Joseph Galaskiewicz. 1990. Correspondence and canonical analysis of relational data. *Journal of Mathematical Sociology* 15(1):11–64.

Watts, D.J., and S.H. Strogatz. 1998. Collective dynamics of small-worldnetworks. *Nature* 393(6684):440–442.

Williams, Christopher, and Matthias Seeger. 2001. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th annual conference on neural information processing systems*, 682–688. EPFL-CONF-161322.

Wolfe, P, and DS Choi. 2014. Co-clustering separately exchangeable network data. *Annals of Statistics*.

Zhao, Y., E. Levina, and J. Zhu. 2011a. On consistency of community detection in networks. *Arxiv preprint arXiv:1110.3854*.

Zhao, Yunpeng, Elizaveta Levina, and Ji Zhu. 2011b. Community extraction for social networks. *Proceedings of the National Academy of Sciences* 108(18):7321–7326.