Submitted to the Annals of Statistics

arXiv:1007.1684v2 [stat.ML] 30 Dec 2010

SPECTRAL CLUSTERING AND THE HIGH-DIMENSIONAL STOCHASTIC BLOCKMODEL

By Karl Rohe, Sourav Chatterjee and Bin Yu

University of California Berkeley

Networks or graphs can easily represent a diverse set of data sources that are characterized by interacting units or actors. Social networks, representing people who communicate with each other, are one example. Communities or clusters of highly connected actors form an essential feature in the structure of several empirical networks. Spectral clustering is a popular and computationally feasible method to discover these communities.

The Stochastic Blockmodel (Holland, Laskey and Leinhardt, 1983) is a social network model with well defined communities; each node is a member of one community. For a network generated from the Stochastic Blockmodel, we bound the number of nodes "misclustered" by spectral clustering. The asymptotic results in this paper are the first clustering results that allow the number of clusters in the model to grow with the number of nodes, hence the name high-dimensional.

In order to study spectral clustering under the Stochastic Blockmodel, we first show that under the more general latent space model, the eigenvectors of the normalized graph Laplacian asymptotically converge to the eigenvectors of a "population" normalized graph Laplacian. Aside from the implication for spectral clustering, this provides insight into a graph visualization technique. Our method of studying the eigenvectors of random matrices is original.

1. Introduction. Researchers in many fields and businesses in several industries have exploited the recent advances in information technology to produce an explosion of data on complex systems. Several of the complex systems have interacting units or actors that networks or graphs can easily represent, providing a range of disciplines with a suite of potential questions on how to produce knowledge from network data. Understanding the

*The authors are grateful to Michael Mahoney and Benjamin Olding for their stimulating discussions. Also, thank you to Jinzhu Jia and Reza Khodabin for your helpful comments and suggestions on this paper. Karl Rohe is partially supported by a NSF VIGRE Graduate Fellowship. Sourav Chatterjee is supported by NSF grant DMS-0707054 and a Sloan Research Fellowship. Bin Yu is partially supported by NSF grant SES-0835531 (CDI), NSF grant DMS-0907632, NSF grant DMS-0907632, and a grant from MSRA.

AMS 2000 subject classifications: Primary 62H30, 62H25; secondary 60B20

Keywords and phrases: Spectral clustering, latent space model, Stochastic Blockmodel, clustering, convergence of eigenvectors, principal components analysis

system of relationships between people can aid both epidemiologists and sociologists. In biology, the predator-prey pursuits in a natural environment can be represented by a food web, helping researchers better understand an ecosystem. The chemical reactions between metabolites and enzymes in an organism can be portrayed in a metabolic network, providing biochemists with a tool to study metabolism. Networks or graphs conveniently describe these relationships, necessitating the development of statistically sound methodologies for exploring, modeling, and interpreting networks.

Communities or clusters of highly connected actors form an essential feature in the structure of several empirical networks. The identification of these clusters helps answer vital questions in a variety of fields. In the communication network of terrorists, a cluster could be a terrorist cell; web pages that provide hyperlinks to each other form a community that might host discussions of a similar topic; and a community or cluster in a social network likely shares a similar interest.

Searching for clusters is algorithmically difficult because it is computationally intractable to search over all possible clusterings. Even on a relatively small graph, one with 100 nodes, the number of different partitions exceeds some estimates of the number of atoms in the universe by twenty orders of magnitude (Champion, 1998). For several different applications, physicists, computer scientists, and statisticians have produced numerous algorithms to overcome these computational challenges. Often these algorithms aim to discover clusters which are approximately the "best" clusters as measured by some empirical objective function (see Fortunato (2009) or Fjällström (1998) for comprehensive reviews of these algorithms from the physics or the engineering perspective respectively).

Clustering algorithms generally come from two sources: from fitting procedures for various statistical models that have well defined communities and, more commonly, from heuristics or insights on what network communities should look like. This division is analogous to the difference in multivariate data analysis between parametric clustering algorithms, such as an EM algorithm fitting a mixture of gaussians model, and nonparametric clustering algorithms such as k-means, which are instead motivated by optimizing an objective function. Handcock, Raftery and Tantrum (2007); Nowicki and Snijders (2001); Snijders and Nowicki (1997) and Airoldi *et al.* (2008) all attempt to cluster the nodes of a network by fitting various network models that have well defined communities. In contrast, the Girvan-Newman algorithm (Girvan and Newman, 2002) and spectral clustering are two algorithms in a large class of algorithms motivated by insights and heuristics on communities in networks.

Newman and Girvan (2004) motivate their algorithm by observing, "If two communities are joined by only a few inter-community edges, then all paths through the network from vertices in one community to vertices in the other must pass along one of those few edges." The Girvan–Newman algorithm searches for these few edges and removes them, resulting in a graph with multiple connected components (connected components are clusters of nodes such that there are no connections between the clusters). The Girvan–Newman algorithm then returns these connected components as the clusters. Like the Girvan–Newman algorithm, spectral clustering is a "nonparametric" algorithm motivated by the following insights and heuristics: spectral clustering is a convex relaxation of the Normalized Cut optimization problem (Shi and Malik, 2000), it can identify the connected components in a graph (if there are any) (Donath and Hoffman, 1973; Fiedler, 1973), and it has an intimate connection with electrical network theory and random walks on graphs (Klein and Randić, 1993; Meilă and Shi, 2001).

1.1. *Spectral clustering.* Spectral clustering is both popular and computationally feasible (von Luxburg, 2007). The algorithm has been rediscovered and reapplied in numerous different fields since the initial work of Donath and Hoffman (1973) and Fiedler (1973). Computer scientists have found many applications for variations of spectral clustering, such as load balancing and parallel computations (Hendrickson and Leland, 1995; Van Driessche and Roose, 1995), partitioning circuits for very-large-scale integration design (Hagen and Kahng, 1992) and sparse matrix partitioning (Pothen, Simon and Liou, 1990). Detailed histories of spectral clustering can be found in Spielman and Teng (2007) and von Luxburg, Belkin and Bousquet (2008).

The algorithm is defined in terms of a graph G, represented by a vertex set and an edge set. The vertex set $\{v_1, \ldots, v_n\}$ contains vertices or nodes. These are the actors in the systems discussed above. We will refer to node v_i as node i. We will only consider unweighted and undirected edges. So, the edge set contains a pair (i, j) if there is an edge, or relationship, between nodes i and j. The edge set can be represented by the adjacency matrix $W \in \{0, 1\}^{n \times n}$:

$$W_{ji} = W_{ij} = \begin{cases} 1 & \text{if } (i,j) \text{ is in the edge set} \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Define L and diagonal matrix D both elements of $\mathcal{R}^{n \times n}$ in the following way,

$$\begin{aligned} D_{ii} &= \textstyle\sum_k W_{ik} \\ L &= D^{-1/2} W D^{-1/2}. \end{aligned} \tag{1.2}$$

Some readers may be more familiar defining L as $I - D^{-1/2}WD^{-1/2}$. For spectral clustering, the difference is immaterial because both definitions have the same eigenvectors.

The spectral clustering algorithm addressed in this paper is defined as follows:

Spectral clustering for k many clusters
Input: Adjacency matrix $W \in \{0,1\}^{n\times n}$.

1. Find the eigenvectors $X_1, \ldots, X_k \in \mathcal{R}^n$ corresponding to the k eigenvalues of L that are largest in absolute value. L is symmetric, so choose these eigenvectors to be orthogonal. Form the matrix $X = [X_1, \ldots, X_k] \in \mathcal{R}^{n\times k}$ by putting the eigenvectors into the columns.

2. Treating each of the n rows in X as a point in $\mathcal{R}^k$, run k-means with k clusters. This creates k non-overlapping sets $A_1, \ldots, A_k$ whose union is $1, \ldots, n$.

Output: $A_1, \ldots, A_k$. This means that node i is assigned to cluster g if the ith row of X is assigned to A_g in step 2.

Traditionally, spectral clustering takes the eigenvectors of L corresponding to the largest k eigenvalues. The algorithm above takes the largest k eigenvalues *by absolute value*. The reason for this is explained in Section 3.

Recently, spectral clustering has also been applied in cases where the graph G and its adjacency matrix W are not given, but instead inferred from a measure of pairwise similarity $k(\cdot,\cdot)$ between data points $X_1, \ldots, X_n$ in a metric space. The similarity matrix $K \in \mathcal{R}^{n\times n}$, whose i,jth element is $K_{ij} = k(X_i, X_j)$, takes the place of the adjacency matrix W in the above definition of L, D, and the spectral clustering algorithm. For image segmentation, Shi and Malik (2000) suggested spectral clustering on an inferred network where the nodes are the pixels and the edges are determined by some measure of pixel similarity. In this way, spectral clustering has many similarities with the nonlinear dimension reduction or manifold learning techniques such as Diffusion maps and Laplacian eigenmaps (Belkin and Niyogi, 2003; Coifman *et al.*, 2005).

The normalized graph Laplacian L is an essential part of spectral clustering, Diffusion maps, and Laplacian eigenmaps. As such, its properties have

been well studied under the model that the data points are randomly sampled from a probability distribution, whose support may be a manifold, and the Laplacian is built from the inferred graph based on some measure of similarity between data points. Belkin (2003); Belkin and Niyogi (2008); Bousquet, Chapelle and Hein (2004); Giné and Koltchinskii (2006); Hein (2006); Hein, Audibert and von Luxburg (2005); Lafon (2004); von Luxburg, Belkin and Bousquet (2008) have all shown various forms of asymptotic convergence for this graph Laplacian. Although all of their results are encouraging, their results do not apply to the random network models we study in this paper.

1.2. *Statistical estimation.* Stochastic models are useful because they force us to think clearly about the randomness in the data in a precise and possibly familiar way. Many random network models have been proposed (Barabási and Albert, 1999; Erdös and Rényi, 1959; Frank and Strauss, 1986; Goldenberg *et al.*, 2009; Hoff, Raftery and Handcock, 2002; Holland, Laskey and Leinhardt, 1983; Holland and Leinhardt, 1981; Van Duijn, Snijders and Zijlstra, 2004; Watts and Strogatz, 1998). Some of these models, such as the Stochastic Blockmodel, have well defined communities. The Stochastic Blockmodel is characterized by the fact that each node belongs to one of multiple blocks and the probability of a relationship between two nodes depends only on the block memberships of the two nodes. If the probability of an edge between two nodes in the same block is larger than the probability of an edge between two nodes in different blocks, then the blocks produce communities in the random networks generated from the model.

Just as statisticians have studied when least-squares regression can estimate the "true" regression model, it is natural and important for us to study 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 would provide a rigorous understanding of the behavior of these algorithms, suggest which algorithm to choose in practice, and aid the corroboration of algorithmic output.

This paper studies the performance of spectral clustering, a nonparametric method, on a parametric task of estimating the blocks in the Stochastic Blockmodel. It connects the first strain of clustering research based on stochastic models to the second strain based on heuristics and insights on network clusters. The Stochastic Block Model allows for some first steps in understanding the behavior of spectral clustering and provides a benchmark to measure its performance. However, because this model does not really account for the complexities observed in several empirical networks, good performance on the Stochastic Blockmodel should only be considered

a necessary requirement for a good clustering algorithm.

Researchers have explored the performance of other clustering algorithms under the Stochastic Blockmodel. Snijders and Nowicki (1997) showed the consistency under the two block Stochastic Blockmodel of a clustering routine that clusters the nodes based on their degree distributions. Although this clustering is very easy to compute it is not clear that the estimators would behave well for larger graphs given the extensive literature on the long tail of the degree distribution (Albert and Barabási, 2002). Later, Condon and Karp (1999) provided an algorithm and proved that it is consistent under the Stochastic Blockmodel, or what they call the planted ℓ-partition model. Their algorithm runs in linear time. However, it always estimates clusters that contain an equal number of nodes. More recently, Bickel and Chen (2009) proved that under the Stochastic Blockmodel, the maximizers of the Newman–Girvan modularity (Newman and Girvan, 2004) and what they call the likelihood modularity are asymptotically consistent estimators of block partitions. These modularities are objective functions that have no clear relationship to the Girvan–Newman algorithm. Finding the maximum of the modularities is NP hard (Brandes *et al.*, 2007). It is important to note that all aforementioned clustering results involving the Stochastic Blockmodel are asymptotic in the number of nodes, with a fixed number of blocks.

The work of Leskovec *et al.* (2008) shows that in a diverse set of large empirical networks (tens of thousands to millions of nodes), the size of the "best" clusters is not very large, around 100 nodes. Modern applications of clustering require an asymptotic regime that allows these sorts of clusters. Under the asymptotic regime cited in the previous paragraph, the size of the clusters grows linearly with the number of nodes. It would be more appropriate to allow the number of communities to grow with the number of nodes. This restricts the blocks from becoming too large, following the empirical observations of Leskovec *et al.* (2008).

This paper provides the first asymptotic clustering results that allow the number of blocks in the Stochastic Blockmodel to grow with the number of nodes. Similar to the asymptotic results on regression techniques that allow the number of predictors to grow with the number of nodes, allowing the number of blocks to grow makes the problem one of high-dimensional learning. Following our initial technical report, Choi, Wolfe and Airoldi (2010) also studied community detection under the Stochastic Blockmodel with a growing number of blocks. They used a likelihood-based approach, which is computationally difficult to implement. However, they are able to greatly weaken the assumptions of this paper.

The Stochastic Blockmodel is an example of the more general latent space model (Hoff, Raftery and Handcock, 2002) of a random network. Under the latent space model, there are latent i.i.d. vectors $z_1, \ldots, z_n$; one for each node. The probability that an edge appears between any two nodes i and j depends only on z_i and z_j and is independent of all other edges and unobserved vectors. The results of Aldous and Hoover show that this model characterizes the distribution of all infinite random graphs with exchangeable nodes (Kallenberg, 2005). The graphs with n nodes generated from a latent space model can be viewed as a subgraph of an infinite graph. In order to study spectral clustering under the Stochastic Blockmodel, we first show that under the more general latent space model, as the number of nodes grows, the eigenvectors of L, the normalized graph Laplacian, converge to eigenvectors of the "population" normalized graph Laplacian that is constructed with a similarity matrix $\mathbb{E}(W|z_1, \ldots, z_n)$ (whose i, jth element is the probability of an edge between node i and j) taking the place of the adjacency matrix W in Equation (1.2). In many ways, $\mathbb{E}(W|z_1, \ldots, z_n)$ is similar to the similarity matrix K discussed above, only this time the vectors $(z_1, \ldots, z_n)$ and their similarity matrix $\mathbb{E}(W|z_1, \ldots, z_n)$ are unobserved.

The convergence of the eigenvectors has implications beyond spectral clustering. Graph visualization is an important tool for social network analysts looking for structure in networks and the eigenvectors of the graph Laplacian are an essential piece of one visualization technique (Koren, 2005). Exploratory graph visualization allows researchers to find structure in the network; this structure could be communities or something more complicated (Freeman, 2000; Liotta, 2004; Wasserman and Faust, 1994). In terms of the latent space model, if $z_1, \ldots, z_n$ form clusters or have some other structure in the latent space, then we might recover this structure from the observed graph using graph visualization. Although there are several visualization techniques, there is very little theoretical understanding of how these techniques perform under stochastic models of structured networks. Because the eigenvectors of the normalized graph Laplacian converge to "population" eigenvectors, this provides support for a visualization technique similar to the one proposed in Koren (2005).

The rest of the paper is organized as follows. The next subsection of the introduction give some preliminary definitions. Following the introduction, there are four main sections; Section 2 studies the latent space model, Section 3 studies the Stochastic Blockmodel as a special case, Section 4 presents some simulation results, and Section 5 investigates the plausibility of a key assumption in five empirical social networks. Section 2 covers the eigenvectors of L under the latent space model. The main technical result is Theorem

2.1 in Section 2.2, which shows that, as the number of nodes grows, the normalized graph Laplacian multiplied by itself converges in Frobenius norm to a symmetric version of the population graph Laplacian multiplied by itself. The Davis-Kahan Theorem then implies that the eigenvectors of these matrices are close in an appropriate sense. Lemma 2.1 specifies how the eigenvectors of a matrix multiplied by itself are closely related to the eigenvectors of the original matrix. Theorem 2.2 combines Theorem 2.1 with the Davis-Kahan Theorem and Lemma 2.1 to show that the eigenvectors of the normalized graph Laplacian converge to the population eigenvectors. Section 3 applies these results to the high-dimensional Stochastic Blockmodel. Lemma 3.1 shows that the population version of spectral clustering can correctly identify the blocks in the Stochastic Blockmodel. Theorem 3.1 extends this result to the sample version of spectral clustering. It uses Theorem 2.2 to bound the number of nodes that spectral clustering "misclusters." This section concludes with two examples. Section 4 presents three simulations that investigate how the asymptotic results apply to finite samples. These simulations suggest an area for future research. The main theorems in this paper require a strong assumption on the degree distribution. Section 5 investigates the plausibility of this assumption with five empirical online social networks. The discussion in Section 6 concludes the paper.

1.3. *Preliminaries.* The latent space model proposed by Hoff, Raftery and Handcock (2002) is a class of a probabilistic model for W.

DEFINITION 1. *For i.i.d. random vectors $z_1, \ldots, z_n \in \mathcal{R}^k$ and random adjacency matrix $W \in \{0,1\}^{n \times n}$, let $\mathbb{P}(W_{ij}|z_i, z_j)$ be the probability mass function of W_{ij} conditioned on z_i and z_j. If a probability distribution on W has the conditional independence relationships*

$$\mathbb{P}(W|z_1, \ldots, z_n) = \prod_{i<j} \mathbb{P}(W_{ij}|z_i, z_j)$$

and $\mathbb{P}(W_{ii} = 0) = 1$ for all i, then it is called an undirected ***latent space model****.*

This model is often simplified to assume $\mathbb{P}(W_{ij}|z_i, z_j) = \mathbb{P}\left(W_{ij}|dist(z_i, z_j)\right)$ where $dist(\cdot, \cdot)$ is some distance function. This allows the "homophily by attributes" interpretation that edges are more likely to appear between nodes whose latent vectors are closer in the latent space.

Define $Z \in \mathcal{R}^{n \times k}$ such that its ith row is z_i for all $i \in V$. **Throughout this paper we assume Z is fixed and unknown**. Because $\mathbb{P}(W_{ij} =$

$1|Z) = \mathbb{E}(W_{ij}|Z)$, the model is then completely parametrized by the matrix

$$\mathscr{W} = \mathbb{E}(W|Z) \in \mathcal{R}^{n\times n},$$

where $\mathscr{W}$ depends on Z, but this is dropped for notational convenience.

The Stochastic Blockmodel, introduced by Holland, Laskey and Leinhardt (1983), is a specific latent space model with well defined communities. We use the following definition of the undirected Stochastic Blockmodel:

DEFINITION 2. *The **Stochastic Blockmodel** is a latent space model with*

$$\mathscr{W} = ZBZ^T,$$

where $Z \in \{0,1\}^{n\times k}$ has exactly one 1 in each row and at least one 1 in each column and $B \in [0,1]^{k\times k}$ is full rank and symmetric.

We refer to $\mathscr{W}$, the matrix which completely parametrizes the latent space model, as the population version of W. Define population versions of L and D both in $\mathcal{R}^{n\times n}$ as

$$\begin{aligned} \mathscr{D}_{ii} &= \textstyle\sum_k \mathscr{W}_{ik} \\ \mathscr{L} &= \mathscr{D}^{-1/2}\mathscr{W}\mathscr{D}^{-1/2} \end{aligned} \tag{1.3}$$

where $\mathscr{D}$ is a diagonal matrix, similar to before.

The results in this paper are asymptotic in the number of nodes n. When it is appropriate, the matrices above are given a superscript of n to emphasize this dependence. Other times, this superscript is discarded for notational convenience.

2. Consistency under the Latent Space Model. We will show that the empirical eigenvectors of $L^{(n)}$ converge in the appropriate sense to the population eigenvectors of $\mathscr{L}^{(n)}$. If $L^{(n)}$ converged to $\mathscr{L}^{(n)}$ in Frobenius norm, then the Davis-Kahan Theorem would give the desired result. However, these matrices do not converge. This is illustrated in an example below. Instead, we give a novel result showing that under certain conditions $L^{(n)}L^{(n)}$ converges to $\mathscr{L}^{(n)}\mathscr{L}^{(n)}$ in Frobenius norm. This implies that the eigenvectors of $L^{(n)}L^{(n)}$ converge to the eigenvectors of $\mathscr{L}^{(n)}\mathscr{L}^{(n)}$. The following lemma shows that these eigenvectors can be chosen to imply the eigenvectors of $L^{(n)}$ converge to the eigenvectors of $\mathscr{L}^{(n)}$.

LEMMA 2.1. *When $M \in \mathcal{R}^{n\times n}$ is a symmetric real matrix,*

1. *λ^2 is an eigenvalue of MM if and only if λ or $-\lambda$ is an eigenvalue of M.*

2. *If* $Mv = \lambda v$, *then* $MMv = \lambda^2 v$.
3. *Conversely, if* $MMv = \lambda^2 v$, *then* v *can be written as a linear combination of eigenvectors of* M *whose eigenvalues are* λ *or* $-\lambda$.

A proof of Lemma 2.1 can be found in Appendix A.

Example : To see how squaring a matrix helps convergence, let the matrix $W \in \mathcal{R}^{n\times n}$ have i.i.d. Bernoulli(1/2) entries. Because the diagonal elements in D grow like n, the matrix W/n behaves similarly to $D^{-1/2}WD^{-1/2}$. Without squaring the matrix, the Frobenius distance from the matrix to its expectation is

$$\|W/n - \mathbb{E}(W)/n\|_F = \frac{1}{n}\sqrt{\sum_{i,j}(W_{ij} - \mathbb{E}(W_{ij}))^2} = 1/2.$$

Notice that, for $i \neq j$,

$$[WW]_{ij} = \sum_k W_{ik}W_{kj} \sim \text{Binomial}(n, 1/4)$$

and $[WW]_{ii} \sim \text{Binomial}(n, 1/2)$. So, for any i, j, $[WW]_{ij} - \mathbb{E}[WW]_{ij} = o(n^{1/2}\log n)$. Thus, the Frobenius distance from the squared matrix to its expectation is

$$\|WW/n^2 - \mathbb{E}(WW)/n^2\|_F = \frac{1}{n^2}\sqrt{\sum_{i,j}\left([WW]_{ij} - \mathbb{E}[WW]_{ij}\right)^2} = o\left(\frac{\log n}{n^{1/2}}\right).$$

When the elements of W are i.i.d. Bernoulli(1/2), $(W/n)^2$ converges in Frobenius norm and W/n does not. The next theorem addresses the convergence of $L^{(n)}L^{(n)}$.

Define

$$\tau_n = \min_{i=1,\ldots,n} \mathscr{D}_{ii}^{(n)}/n. \tag{2.1}$$

Recall that $\mathscr{D}_{ii}^{(n)}$ is the expected degree for node i. So, τ_n is the minimum expected degree, divided by the maximum possible degree. It measures how quickly the number of edges accumulates.

THEOREM 2.1. *Define the sequence of random matrices* $W^{(n)} \in \{0,1\}^{n\times n}$ *to be from a sequence of latent space models with population matrices* $\mathscr{W}^{(n)} \in [0,1]^{n\times n}$. *With* $W^{(n)}$, *define the observed graph Laplacian* $L^{(n)}$ *as in* (1.2).

Let $\mathscr{L}^{(n)}$ be the population version of $L^{(n)}$ as defined in Equation (1.3). *Define τ_n as in Equation* (2.1).

If there exists $N > 0$, such that $\tau_n^2 \log n > 2$ for all $n > N$, then

$$\|L^{(n)}L^{(n)} - \mathscr{L}^{(n)}\mathscr{L}^{(n)}\|_F = o\left(\frac{\log n}{\tau_n^2\ n^{1/2}}\right) \quad a.s.$$

Appendix A contains a non-asymptotic bound on $\|L^{(n)}L^{(n)} - \mathscr{L}^{(n)}\mathscr{L}^{(n)}\|_F$ as well as the proof of Theorem 2.1. The main condition in this theorem is the lower bound on τ_n. This sufficient condition is used to produce Gaussian tail bounds for each of the D_{ii} and other similar quantities.

For any symmetric matrix M, define $\lambda(M)$ to be the eigenvalues of M and for any interval $S \subset \mathcal{R}$, define

$$\lambda_S(M) = \{\lambda(M) \cap S\}.$$

Further, define $\bar{\lambda}_1^{(n)} \geq \cdots \geq \bar{\lambda}_n^{(n)}$ to be the elements of $\lambda(\mathscr{L}^{(n)}\mathscr{L}^{(n)})$ and $\lambda_1^{(n)} \geq \cdots \geq \lambda_n^{(n)}$ to be the elements of $\lambda(L^{(n)}L^{(n)})$. The eigenvalues of $L^{(n)}L^{(n)}$ converge in the following sense,

$$\begin{aligned} \max_i |\lambda_i^{(n)} - \bar{\lambda}_i^{(n)}| &\leq \|L^{(n)}L^{(n)} - \mathscr{L}^{(n)}\mathscr{L}^{(n)}\|_F \\ &= o\left(\frac{\log n}{\tau_n^2\ n^{1/2}}\right) \quad a.s. \end{aligned} \tag{2.2}$$

This follows from Theorem 2.1, Weyl's inequality (Bhatia, 1987), and the fact that the Frobenius norm is an upper bound of the spectral norm.

This shows that under certain conditions on τ_n, the eigenvalues of $L^{(n)}L^{(n)}$ converge to the eigenvalues of $\mathscr{L}^{(n)}\mathscr{L}^{(n)}$. In order to study spectral clustering, it is now necessary to show that the eigenvectors also converge. The Davis-Kahan Theorem provides a bound for this.

PROPOSITION 2.1. ***(Davis-Kahan)*** *Let $S \subset \mathcal{R}$ be an interval. Denote $\mathcal{X}$ as an orthonormal matrix whose column space is equal to the eigenspace of $\mathscr{L}\mathscr{L}$ corresponding to the eigenvalues in $\lambda_S(\mathscr{L}\mathscr{L})$ (more formally, the column space of $\mathcal{X}$ is the image of the spectral projection of $\mathscr{L}\mathscr{L}$ induced by $\lambda_S(\mathscr{L}\mathscr{L})$). Denote by X the analogous quantity for LL. Define the distance between S and the spectrum of $\mathscr{L}\mathscr{L}$ outside of S as*

$$\delta = \min\{|\ell - s|; \ell \text{ eigenvalue of } \mathscr{L}\mathscr{L},\ \ell \notin S,\ s \in S\}.$$

If $\mathcal{X}$ and X are of the same dimension, then there is an orthonormal matrix O, that depends on $\mathcal{X}$ and X, such that

$$\frac{1}{2}\|X - \mathcal{X}O\|_F^2 \leq \frac{\|LL - \mathscr{L}\mathscr{L}\|_F^2}{\delta^2}$$

The original Davis-Kahan Theorem bounds the "canonical angle," also known as the "principal angle," between the column spaces of $\mathcal{X}$ and X. Appendix B explains how this can be converted into the bound stated above. To understand why the orthonormal matrix O is included, imagine the situation that $L = \mathscr{L}$. In this case X is not necessarily equal to $\mathscr{X}$. At a minimum, the columns of X could be a permuted version of those in $\mathscr{X}$. If there are any eigenvalues with multiplicity greater than one, these problems could be slightly more involved. The matrix O removes these inconveniences and related inconveniences.

The bound in the Davis-Kahan Theorem is sensitive to the value δ. This reflects that when there are eigenvalues of $\mathscr{L}\mathscr{L}$ close to S, but not inside of S, then a small perturbation can move these eigenvalues inside of S and drastically alter the eigenvectors. The next theorem combines the previous results to show that the eigenvectors of $L^{(n)}$ converge to the eigenvectors of $\mathscr{L}^{(n)}$. Because it is asymptotic in the number of nodes, it is important to allow S and δ to depend on n. For a sequence of open intervals $S_n \subset \mathcal{R}$, define

$$\delta_n = \inf\{|\ell - s|;\ \ell \in \lambda(\mathscr{L}^{(n)}\mathscr{L}^{(n)}),\ \ell \notin S_n,\ s \in S_n\} \tag{2.3}$$

$$\delta'_n = \inf\{|\ell - s|;\ \ell \in \lambda_{S_n}(\mathscr{L}^{(n)}\mathscr{L}^{(n)}),\ s \notin S_n\} \tag{2.4}$$

$$S'_n = \{\ell; \ell^2 \in S_n\}. \tag{2.5}$$

The quantity δ'_n is added to measure how well S_n insulates the eigenvalues of interest. If δ'_n is too small, then some important empirical eigenvalues might fall outside of S_n. By restricting the rate at which δ_n and δ'_n converge to zero, the next theorem ensures the dimensions of X and $\mathscr{X}$ agree for a large enough n. This is required in order to use the Davis-Kahan Theorem.

THEOREM 2.2. *Define $W^{(n)} \in \{0,1\}^{n\times n}$ to be a sequence of growing random adjacency matrices from the latent space model with population matrices $\mathscr{W}^{(n)}$. With $W^{(n)}$, define the observed graph Laplacian $L^{(n)}$ as in* (1.2)*. Let $\mathscr{L}^{(n)}$ be the population version of $L^{(n)}$ as defined in Equation* (1.3)*. Define τ_n as in Equation* (2.1)*. With a sequence of open intervals $S_n \subset \mathcal{R}$, define δ_n, δ'_n, and S'_n as in Equations* (2.3)*,* (2.4)*, and* (2.5)*.*

Let $k_n = |\lambda_{S'_n}(L^{(n)})|$, the size of the set $\lambda_{S'_n}(L^{(n)})$. Define the matrix $X_n \in \mathcal{R}^{n\times k_n}$ such that its orthonormal columns are the eigenvectors of symmetric matrix $L^{(n)}$ corresponding to all the eigenvalues contained in $\lambda_{S'_n}(L^{(n)})$. For $\mathscr{K}_n = |\lambda_{S'_n}(\mathscr{L}^{(n)})|$, define $\mathscr{X}_n \in \mathcal{R}^{n\times\mathscr{K}_n}$ to be the analogous matrix for symmetric matrix $\mathscr{L}^{(n)}$ with eigenvalues in $\lambda_{S'_n}(\mathscr{L}^{(n)})$.

Assume that $n^{-1/2}(\log n)^2 = O(\min\{\delta_n, \delta'_n\})$. Also assume that there exists positive integer N such that for all $n > N$, it follows that $\tau_n^2 > 2/\log n$.

Eventually, $k_n = \mathscr{K}_n$. *Afterwards, for some sequence of orthonormal rotations* O_n,

$$\|X_n - \mathscr{X}_n O_n\|_F = o\left(\frac{\log n}{\delta_n \tau_n^2 \, n^{1/2}}\right) \quad a.s.$$

A proof of Theorem 2.2 is in Appendix C. There are two key assumptions in Theorem 2.2:

(1) $n^{-1/2}(\log n)^2 = O(\min\{\delta_n, \delta_n'\})$
(2) $\tau_n^2 > 2/\log n$.

The first assumption ensures that the "eigengap," the gap between the eigenvalues of interest and the rest of the eigenvalues, does not converge to zero too quickly. The theorem is most interesting when S includes only the leading eigenvalues. This is because the eigenvectors with the largest eigenvalues have the potential to reveal clusters or other structures in the network. When these leading eigenvalues are well separated from the smaller eigenvalues, the eigengap is large. The second assumption ensures that the expected degree of each node grows sufficiently fast. If τ_n is constant, then the expected degree of each node grows linearly. The assumption $\tau_n^2 > 2/\log n$ is almost as restrictive.

The usefulness of Theorem 2.2 depends on how well the eigenvectors of $\mathscr{L}^{(n)}$ represent the characteristics of interest in the network. For example, under the Stochastic Blockmodel with B full rank, if S_n is chosen so that S_n' contains all nonzero eigenvalues of $\mathscr{L}^{(n)}$, then the block structure can be determined from the columns of $\mathcal{X}_n$. It can be shown that nodes i and j are in the same block if and only if the ith row of $\mathcal{X}_n$ equals the jth row. The next section examines how spectral clustering exploits this structure, using X_n to estimate the block structure in the Stochastic Blockmodel.

3. The Stochastic Blockmodel. The work of Leskovec *et al.* (2008) shows that the sizes of the best clusters are not very large in a diverse set of empirical networks, suggesting that the appropriate asymptotic framework should allow for the number of communities to grow with the number of nodes. This section shows that, under suitable conditions, spectral clustering can correctly partition most of the nodes in the Stochastic Blockmodel, even when the number of blocks grows with the number of nodes.

The Stochastic Blockmodel, introduced by Holland, Laskey and Leinhardt (1983), is a specific latent space model. Because it has well defined communities in the model, community detection can be framed as a problem of statistical estimation. The important assumption of this model is that of

stochastic equivalence within the blocks; if two nodes i and j are in the same block, rows i and j of $\mathscr{W}$ are equal.

Recall in the definition of the undirected Stochastic Blockmodel,

$$\mathscr{W} = ZBZ^T,$$

where $Z \in \{0,1\}^{n\times k}$ is fixed and has exactly one 1 in each row and at least one 1 in each column and $B \in [0,1]^{k\times k}$ is full rank and symmetric. In this definition there are k blocks and n nodes. If the i,gth element of Z equals one ($Z_{ig} = 1$) then node i is in block g. As before, z_i for $i = 1,\ldots,n$ denotes the ith row of Z. The matrix $B \in [0,1]^{k\times k}$ contains the probability of edges within and between blocks. Some researchers have allowed for Z to be random, we have decided to focus instead on the randomness of W conditioned on Z. The aim of a clustering algorithm is to estimate Z (up to a permutation of the columns) from W.

This section bounds the number of "misclustered" nodes. Because a permutation of the columns of Z is unidentifiable in the Stochastic Blockmodel, it is not obvious what a "misclustered" node is. Before giving our definition of "misclustered," some preliminaries are needed to explain why it is a reasonable definition. The next paragraphs examine the behavior of spectral clustering applied to the population graph Laplacian $\mathscr{L}$. Then, this is compared to spectral clustering applied to the observed graph Laplacian L. This motivates our definition of "misclustered."

Recall that the spectral clustering algorithm applied to L,

(1) finds the eigenvectors, $X \in R^{n\times k}$,
(2) treats each row of the matrix X as a point in $\mathcal{R}^k$, and
(3) runs k-means on these points.

k-means is an objective function. Applied to the points $\{x_1,\ldots,x_n\} \subset R^k$ it is (Steinhaus, 1956),

$$\min_{\{m_1,\ldots,m_k\}\subset\mathcal{R}^k} \sum_i \min_g \|x_i - m_g\|_2^2. \tag{3.1}$$

The analysis in this paper addresses the true optimum of (3.1). (In practice, this optimization problem can suffer from local optima.) The vectors $m_1^*,\ldots,m_k^*$ that optimize the k-means function are referred to as the *centroids* of the k clusters.

This next lemma shows that spectral clustering applied to the population Laplacian, $\mathscr{L}$, can discover the block structure in the matrix Z. This lemma is essential to defining "misclustered."

LEMMA 3.1. *Under the Stochastic Blockmodel with k blocks,*

$$\mathscr{W} = ZBZ^T \in R^{n\times n} \text{ for } B \in R^{k\times k} \text{ and } Z \in \{0,1\}^{n\times k},$$

define $\mathscr{L}$ as in (1.3)*. There exists a matrix $\mu \in R^{k\times k}$ such that the columns of $Z\mu$ are the eigenvectors of $\mathscr{L}$ corresponding to the nonzero eigenvalues values. Further,*

$$z_i\mu = z_j\mu \;\Leftrightarrow\; z_i = z_j, \tag{3.2}$$

where z_i is the ith row of Z.

A proof of Lemma 3.1 is in Appendix D.

Equivalence statement (3.2) implies that under the k block Stochastic Blockmodel there are k unique rows in the eigenvectors $Z\mu$ of $\mathscr{L}$. This has important consequences for the spectral clustering algorithm. The spectral clustering algorithm applied to $\mathscr{L}$ will run k-means on the rows of $Z\mu$. Because there are only k unique points, each of these points will be a centroid of one of the resulting clusters. Further, if $z_i\mu = z_j\mu$, then i and j will be assigned to the same cluster. With equivalence statement (3.2), this implies that spectral clustering applied to the matrix $\mathscr{L}$ can perfectly identify the block memberships in Z. Obviously, $\mathscr{L}$ is not observed. In practice, spectral clustering is applied to L. Let $X \in R^{n\times k}$ be a matrix whose orthonormal columns are the eigenvectors corresponding to the largest k eigenvalues (in absolute value) of L.

DEFINITION 3. *Spectral clustering applies the k-means algorithm to the rows of X, i.e. each row is a point in R^k. Each row is assigned to one cluster and each of these clusters has a centroid. Define $c_1, \ldots, c_n \in R^k$ such that c_i is the the centroid corresponding to the ith row of X.*

Recall that $z_i\mu$ is the centroid corresponding to node i from the population analysis. If the observed centroid c_i is closer to the population centroid $z_i\mu$ than it is to any other population centroid $z_j\mu$ for $z_j \neq z_i$, then it appears that node i is correctly clustered. This definition is appealing because it removes some of the cluster identifiability problem. However, the eigenvectors add one additional source of undentifiability. Let $O \in R^{k\times k}$ be the orthonormal rotation from Theorem 2.2. Consider node i to be correctly clustered if, c_i is closer to $z_i\mu O$ than it is to any other (rotated) population centroid $z_j\mu O$ for $z_j \neq z_i$. The slight complication with O stems from the fact that the vectors $c_1, \ldots, c_n$ are constructed from the eigenvectors in X

and Theorem 2.2 shows these eigenvectors converge to the *rotated* population eigenvectors: $\mathscr{X}O = Z\mu O$.

Define P to be the population of the largest block in Z.

$$P = \max_{j=1,\ldots,k} (Z^T Z)_{jj} \tag{3.3}$$

The following provides a sufficient condition for a node to be correctly clustered.

LEMMA 3.2. *For the orthonormal matrix $O \in \mathcal{R}^{k\times k}$ from Theorem 2.2,*

$$\|c_i - z_i \mu O\|_2 < 1/\sqrt{2P} \implies \tag{3.4}$$

$$\|c_i - z_i \mu O\|_2 < \|c_i - z_j \mu O\|_2 \quad \text{for any } \; z_j \neq z_i. \tag{3.5}$$

A proof of Lemma 3.2 is in Appendix D.

Line (3.5) is the previously motivated definition of correctly clustered. Thus, Lemma 3.2 shows that the inequality in line (3.4) is a sufficient condition for node i to be correctly clustered.

DEFINITION 4. *Define the set of misclustered nodes as the nodes that do not satisfy the sufficient condition* (3.4)*,*

$$\mathscr{M} = \left\{ i : \; \|c_i - z_i \mu O\|_2 \geq 1/\sqrt{2P} \right\}. \tag{3.6}$$

The next theorem bounds the size of the set $\mathscr{M}$

THEOREM 3.1. *Suppose $W \in \mathcal{R}^{n\times n}$ is an adjacency matrix from the Stochastic Blockmodel with k_n blocks. Define the population graph Laplacian, $\mathscr{L}$, as in* (1.3)*. Define $|\bar{\lambda}_1| \geq |\bar{\lambda}_2| \geq \cdots \geq |\bar{\lambda}_{k_n}| > 0$ as the absolute values of the k_n nonzero eigenvalues of $\mathscr{L}$. Define $\mathscr{M}$, the set of misclustered nodes, as in* (3.6)*. Define τ_n as in* (2.1) *and assume there exists N such that for all $n > N$, $\tau_n^2 > 2/\log n$. Define P_n as in* (3.3)*. If $n^{-1/2}(\log n)^2 = O(\lambda_{k_n}^2)$, then the number of misclustered nodes is bounded*

$$|\mathscr{M}| = o\left(\frac{P_n (\log n)^2}{\lambda_{k_n}^4 \tau_n^4 \; n} \right).$$

A proof of Theorem 3.1 is in Appendix D. The two main assumptions of Theorem 3.1 are

(1) $n^{-1/2}(\log n)^2 = O(\lambda_{k_n}^2)$
(2) eventually, $\tau_n^2 \log n > 2$.

They imply the conditions needed to apply Theorem 2.2. The first assumption requires that the smallest nonzero eigenvalue of $\mathscr{L}$ is not too small. Combined with an appropriate choice of S_n, this assumption implies the eigengap assumption in Theorem 2.2. The second assumption is exactly the same as the second assumption in Theorem 2.2. Section 4 investigates the sensitivity of spectral clustering to these two assumptions. Section 5 examines the plausibility of assumption (2) on five empirical online social networks.

In all previous spectral clustering algorithms, it has been suggested that the eigenvectors corresponding to the largest eigenvalues reveal the clusters of interest. The above theorem suggests that before finding the largest eigenvalues, you should first order them by absolute value. This allows for large and negative eigenvalues. In fact, eigenvectors of L corresponding to eigenvalues close to negative one (all eigenvalues of L are in $[-1, 1]$) discover "heterophilic" structure in the network that can be useful for clustering. For example, in the network of dating relationships in a high school, two people of opposite sex are more likely to date than people of the same sex. This pattern creates the two male and female "clusters" that have many fewer edges within than between clusters. In this case, L would likely have an eigenvalue close to negative one. The corresponding eigenvector would reveal these "heterophilic" clusters.

Example: To examine the ability of spectral clustering to discover heterophilic clusters, imagine a Stochastic Blockmodel with two blocks and two nodes in each block. Define

$$B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In this case, there are no connections within blocks and every member is connected to the two members of the opposite block. There is no variability in the matrix W. The rows and columns of L can be reordered so that it is a block matrix. The two block matrices down the diagonal are 2×2 matrices of zeros and all the elements in the off diagonal blocks are equal to $1/2$. There are two nonzero eigenvalues of L. Any constant vector is an eigenvector of L with eigenvalue equal to one. The remaining eigenvalue belongs to any eigenvector that is a constant multiple of $(1, 1, -1, -1)$. In this case, with perfect "heterophilic" structure, the eigenvector that is useful for finding the clusters has eigenvalue negative one.

Heuristically, the reason spectral clustering can discover these heterophilic blocks is related to our method of proof. The i, jth element of WW is the number neighbors that nodes i and j have in common. In both heterophilic

and homophilic cases, if nodes i and j are in the same block, then they should have several neighbors in common. Thus $[WW]_{ij}$ is large. Similarly, $[LL]_{ij}$ is large. This shows that the number of common neighbors is a measure of similarity that is robust to the choice of hetero- or homophilic clusters. Because spectral clustering uses a related measure of similarity, it is able to detect both types of clusters.

In order to clarify the bound on $|\mathscr{M}|$ in Theorem 3.1, a simple example illustrates how λ_{k_n}, τ_n, and P might depend on n.

DEFINITION 5. *The **four parameter Stochastic Blockmodel** is parametrized by k, s, r, and p. There are k blocks each containing s nodes. The probability of a connection between two nodes in two separate blocks is $r \in [0,1]$ and the probability of a connection between two nodes in the same block is $p + r \in [0,1]$.*

Example: In the four parameter Stochastic Blockmodel, there are $n = ks$ nodes. Notice that $P_n = s$ and $\tau_n > r$. Appendix D shows that the smallest nonzero eigenvalue of the population graph Laplacian is equal to

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

Using Theorem 3.1, if $p \neq 0$ and $k = O(n^{1/4}/\log n)$, then

$$|\mathscr{M}| = o(k^3(\log n)^2) \quad a.s. \tag{3.7}$$

Further, the proportion of nodes that are misclustered converges to zero,

$$\frac{|\mathscr{M}|}{n} = o(n^{-1/4}) \quad a.s.$$

This example is particularly surprising after noticing that if $k = n^\alpha$ for $\alpha \in (0, 1/4)$, then the vast majority of edges connect nodes in different blocks. To see this, look at a sequence of models such that $k = n^\alpha$. Note that $s = n^{1-\alpha}$. So, for each node, the expected number of connections to nodes in the same block is $(p+r)n^{1-\alpha}$ and the expected number of connections to nodes in different blocks is $r(n - n^{1-\alpha})$.

$$\frac{\text{Expected number of in block connections}}{\text{Expected number of out of block connections}} = \frac{(p+r)n^{1-\alpha}}{r(n-n^{1-\alpha})} = O(n^{-\alpha})$$

These are not the tight communities that many imagine when considering networks. Instead, a dwindling fraction of each node's edges actually connect

to nodes in the same block. The vast majority of edges connect nodes in different blocks.

A more refined result would allow r to decay with n. However, when r decays, so does the minimum expected degree and the tail bounds used in proving Theorem 2.1 requires the minimum expected degree to grow nearly as fast as n. Allowing r to decay with n is an area for future research.

4. Simulations. Three simulations in this section illustrate how the asymptotic bounds in this paper can be a guide for finite sample results. These simulations emphasize the importance of the eigengap in Theorem 2.2 and suggest that the asymptotic bounds in this paper hold for relatively small networks. The simulations also suggest two shortcomings of the theoretical results in this paper. First, Simulation 1 shows that spectral clustering appears to be consistent in some situations. Unfortunately, the theoretical results in Theorem 3.1 are not sharp enough to prove consistency. Second, Simulation 3 suggests that spectral clustering is still consistent even when the minimum expected node degree grows more slowly than the number of nodes. However, the theorems above require a stronger condition, that the minimum expected degree grows almost linearly with the number of nodes.

All data are simulated from the four parameter Stochastic Blockmodel (Definition 5). In the first simulation, the number of nodes in each block s grows while the number of blocks k and the probabilities p and r remain fixed. In the second simulation, k grows while s, p, and r remain fixed. In the final simulation, s and k remain fixed while r and p shrink such that p/r remains fixed. Because kr/p is fixed, the eigengap is also fixed.

There is one important detail to recreate our simulation results below. The spectral clustering result stated in Theorem 3.1, requires the true optimum of the k-means objective function. This is very difficult to ensure. However, only one step in the proof of Theorem 3.1 requires the true optimum. The optimum of k-means satisfies inequality D.4 in the appendix. In simulations, this inequality can be verified directly. For the simulations below, the k-means algorithm is run several times, all with random initializations, until the bound D.4 is met.

Simulation 1: In this simulation, $k = 5,\ p = .2,\ r = .1$ and the number of members in each group grows from 8 to 215. This implies that n grows from 40 to 1075. Equation (3.7) suggests that the number of misclustered nodes should grow more slowly than $(\log n)^2$. In fact, Figure 1 shows that once there are enough nodes, the number of misclustered nodes converges to zero. The top plot displays the number of misclustered nodes plotted against $\log n$, which initially increases. Then, it falls precipitously.

The lower plot in Figure 1 displays why the number of misclustered nodes falls so precipitously. It plots $\log \|LL - \mathscr{L}\mathscr{L}\|_F$ (dashed bold line) and $\log \|X - \mathscr{X}O\|_F$ (solid bold line) on the vertical axis against $\log n$ on the horizontal axis. Also displayed in this plot is a line with slope -1/2 (solid thin line). Note that the solid bold line starts to run parallel to the solid thin line once $\log n > 4.5$. After this point, the eigenvectors converge, and spectral clustering begins to correctly cluster all of the nodes. The proof of the convergence of the eigenvectors for Theorem 2.2, requires an eigengap condition,

$$n^{-1/2} \log n = O(\min\{\delta_n, \delta_n'\}).$$

Similarly to the example in the previous section, S_n can be chosen in this four parameter model so that $\min\{\delta_n, \delta_n'\} = (k(r/p) + 1)^{-2}$. In this simulation, the eigenvectors begin to converge, and the number of miclustered nodes drops just after the bound $n^{-1/2} < (k(r/p)+1)^{-2}$ is met. Ignoring the $\log n$ factor, this suggests that the eigengap condition in Theorem 2.2 is necessary.

This simulation demonstrates the importance of the relationship between the sample size and the eigengap. In this simulation, there needs to be roughly 50 nodes in each block to separate the informative eigenvectors from the uninformative eigenvectors. Once there are enough nodes, the empirical eigenvectors are close to the population eigenvectors. Then, spectral clustering can estimate the block structure.

The lower plot in Figure 1 also suggests that, ignoring $\log n$ factors, the rates of convergence given in Theorem 2.1 and Theorem 2.2 are sharp. Both LL and the eigenvectors X converge at a rate $O(n^{-1/2})$. This is because the the dashed bold line and the solid bold line (for large enough n) are approximately parallel to the solid thin line.

Simulation 2: In this simulation from the four parameter Stochastic Blockmodel, each block contains 35 nodes, $p = .3$, and $r = .05$. The number of blocks k grows from 2 to 110. Equation (3.7) suggests that under this asymptotic regime, the number of misclustered nodes should grow more slowly than $k^3(\log n)^2$. Figure 2 shows how this theoretical quantity can be an appropriate guide.

Figure 2 plots the log of the number of misclustered nodes (bold line) against $\log k$. For comparison, a line with slope 3 is also plotted (thin line). Because the bold line has a slope approximately equal to the thin line, the number of misclustered nodes is approximate to k^3.

This simulation demonstrates that as the number of blocks grows, the number of misclustered nodes also grows. Although $\|LL - \mathscr{L}\mathscr{L}\|_F$ converges under this asymptotic regime, $\|X - \mathscr{X}O\|_F$ does not because the eigengap shrinks more quickly than the number of nodes can tolerate.

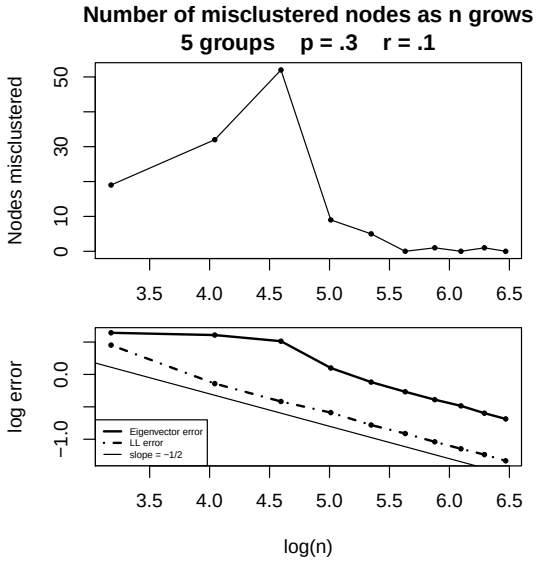


FIG 1. *The top panel in this figure displays the number of misclustered nodes plotted against* $\log n$. *The bottom panel displays both* $\log \|LL - \mathscr{L}\mathscr{L}\|_F$ *and* $\log \|X - \mathscr{X}O\|_F$ *plotted against* $\log n$. *Each dot represents one simulation of the model. In addition, the bottom panel has a line with slope* -1/2. *This figure illustrates two things. First, after a certain threshold (around* $\log n = 4.7$*), the eigenvectors of the graph Laplacian begin to converge and after this point, the number of misclustered nodes converges to zero. Second, the lines representing* $\log \|LL - \mathscr{L}\mathscr{L}\|_F$ *and* $\log \|X - \mathscr{X}O\|_F$ *are approximately parallel to the line with slope* -1/2. *This suggests that they converge around rate* $O(n^{-1/2})$, *similar to the theoretical results in Lemma 2.1 and Theorem 2.2.*

Simulation 3: The theorems in this paper assume that the smallest expected degree grows close to linearly with the number of nodes in the graph. This simulation examines the sensitivity of spectral clustering to this assumption. Recall that the smallest expected degree is equal to $n\tau$.

In this simulation, there are three different designs all from the four parameter Stochastic Blockmodel. Each design has three blocks ($k = 3$). One design contains 50 nodes in each block, another contains 150 in each block, and the last design contains 250 nodes in each block. To investigate how sensitive spectral clustering is to the value of $\tau = p/k + r$, the probabilities p and r must change. However, to isolate the effect of τ from the effect of the eigengap $(k(r/p) + 1)^{-2}$, it is necessary to keep the ratio p/r constant. Fixing $p/r = 2$ ensures that the eigengap is fixed at $4/25$.

The results for Simulation 3 are displayed in Figure 3. The value τ is on the horizontal axis, and the number of misclustered nodes is on the vertical axis. There are three lines. The thickest line represents the design with 50

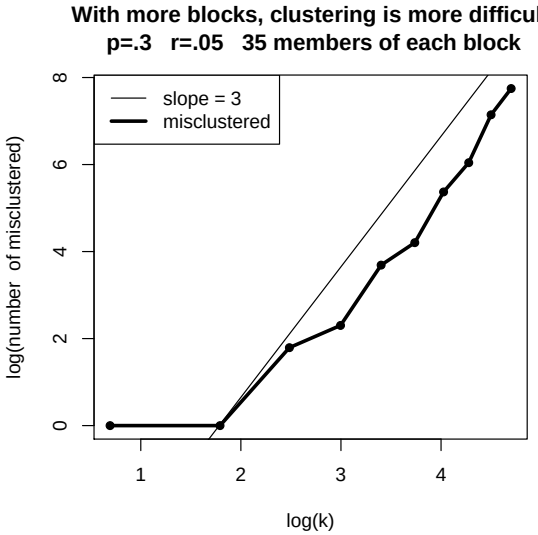


FIG 2. *This figure plots the number of miclustered nodes (thicker line) against* $\log k$. *Each dot represents one simulation from the model. Additionally, there is a line with slope* 3 *(thinner line). Equation 3.7 says that the number of misclustered nodes is* $o(k^3 \log k)$. *Because the thicker line has a slope that is similar to the thinner line, this result appears to be a good approximation.*

nodes in each block. The line of medium thickness represents the design with 150 nodes in each block. The thinnest line represents the design with 250 nodes in each block. All three lines increase as τ approaches zero (reading the figure from right to left). The thickest line starts to increase at $\tau = .20$. The thinnest line starts to increase at $\tau = .07$. The line with medium thickness increases in between these two lines.

Because the thinner lines start to increase at a smaller value of τ, this suggests that as n increase, τ can decrease. As such, spectral clustering should be able to correctly cluster the nodes in a Stochastic Blockmodel graph when the minimum expected degree does not grow linearly with the number of nodes in the graph.

Lemma 2.1, Theorem 2.2, and Theorem 3.1 all require the minimum expected degree to grow at the same rate as n (ignoring $\log n$ terms). Although the strict assumption is inappropriate for large networks, this simulation demonstrates (1) that spectral clustering works for smaller networks and (2) that the asymptotic theory presented earlier in the paper can be a guide to smaller networks. In these networks, it is not as unreasonable that each node would be connected to a significant proportion of the other nodes.

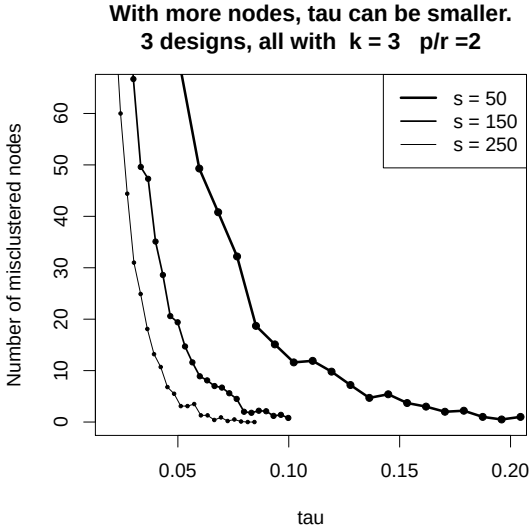


FIG 3. *This figure displays the number of misclustered nodes from three different models plotted against* $\tau = \min_i E(D_{ii})/n$. *The first model has* 50 *nodes in each block (thickest line), the second model has* 150 *nodes in each block (line with medium thickness), the third model has* 250 *nodes in each block (thinnest line). Each dot represents the average of ten simulations from the model. In each of these models,* p *and* r *decrease such that* p/r *is always equal to* 2. *This ensures that* τ *goes to zero, while the eigengap remains constant. Each of the three models is sensitive to small values of* τ. *However, the larger models can tolerate a smaller value of* τ. *This suggests that as* n *increases,* τ *should be allowed to decrease. The theorems in this paper do not allow for that possibility.*

5. Empirical edge density. In several networks there is a natural or canonical notion of what an edge represents. In an online social network, friendship is the canonical notion of an edge. With this canonical notion, the edges in most empirical networks are not dense enough to suggest the asymptotic framework assumed in Lemma 2.1, Theorem 2.2, and Theorem 3.1.

Although it is an area of future research to weaken the strong assumption on the expected node degrees, there are potentially other notions of similarity that can replace the canonical notion. Define the canonical edge set E_c to contain (i, j) if nodes i and j are connected with a canonical edge. One possible extension of E_c is

$$E_{ff} = \{(i,j) : \text{if } (i,k) \in E_c \text{ and } (k,j) \in E_c \text{ for some } k\} \tag{5.1}$$

In words, $(i, j) \in E_{ff}$ if i and j are friends of friends.

Table 5 investigates the edge density of five empirical network defined us-

ing both E_c and E_{ff}. These five networks come from the Facebook networks of five universities: California Institute of Technology (Caltech), Princeton University, Georgetown University, University of Oklahoma, University of North Carolina at Chapel Hill (UNC). Traud *et al.* (2008) made these data sets publicly available and investigated the community structure in them.

Let W^c denote the adjacency matrix constructed from E_c. Let W^{ff} denote the adjacency matrix constructed from E_{ff}. Let $deg^c \in \mathcal{R}^n$ and $deg^{ff} \in \mathcal{R}^n$ denote the degree sequences of the nodes with respect to the two edge sets E_c and E_{ff}. That is, $deg_i^{ff} = \sum_j W_{ij}^{ff}$. Similarly for deg^c. Define

$$\overline{deg}^c = \frac{1}{n}\sum_i deg_i^c \tag{5.2}$$

$$\overline{deg}^{ff} = \frac{1}{n}\sum_i deg_i^{ff} \tag{5.3}$$

$$T_c = \frac{100\%}{n}\sum_i \mathbf{1}\{deg_i^c > n/10\} \tag{5.4}$$

$$T_{ff} = \frac{100\%}{n}\sum_i \mathbf{1}\{deg_i^{ff} > n/10\}. \tag{5.5}$$

The first two quantities are equal to the average node degrees. The last two quantities are the percent of nodes connected to more that 10% of the nodes in the network.

Table 5 demonstrates how the edge density increases after replacing E_c with E_{ff}. The statistics T_c and T_{ff}, in the last two lines of the table, can be used to gauge the suitability of the assumption $\tau^2 > 2/\log n$ in the theorems above. Recall that τ is the minimum expected degree divided by n. So, for example, if $T_{ff} = 1$, then it is reasonable to expect that $\tau > 1/10$. Because there are some nodes that have a very small degree, T_c and T_{ff} look at the proportion of nodes that are well connected.

It is an empirical observation that graphs have sparse degrees. This suggests that the assumption $\tau^2 > 2/\log n$ in Lemma 2.1, Theorem 2.2, and Theorem 3.1 is not satisfied in practice. Table 5 demonstrates that by using an alternative notion of adjacency or connected, the network can become much more connected.

6. Discussion. The goal of this paper is to bring statistical rigor to the study of community detection by assessing how well spectral clustering can estimate the clusters in the Stochastic Blockmodel. The Stochastic Blockmodel is easily amenable to the analysis of clustering algorithms because of

TABLE 1

This table describes five basic characteristics of the Facebook social network within five universities. In the table below, $\overline{deg}^{c}$ is the average node degree using the canonical edges of friendship and $\overline{deg}^{ff}$ is the average node degree using the "friends-of-friends" edges as defined with Equation 5.1. The statistics T_c and T_{ff} (defined in Equations 5.4 and 5.5) are equal to the percent of nodes that are connected to more than 10% of the nodes in the graph. The table below shows that the network is much more connected when using edges defined by "friends-of-friends." All numbers are rounded to the nearest integer.

School	Caltech	Princeton	Georgetown	Oklahoma	UNC
n	769	6596	9414	17425	18163
$\overline{deg}^{c}$	43	89	90	102	84
$\overline{deg}^{ff}$	487	2663	3320	5420	5242
T_c	16	0	0	0	0
T_{ff}	94	88	87	81	79

its simplicity and well defined communities. The fact that spectral clustering performs well on the Stochastic Blockmodel is encouraging. However, because the Stochastic Blockmodel fails to represent fundamental features that most empirical networks display, this result should only be considered a first step.

This paper has two main results. The first main result, Theorem 2.2, proves that under the latent space model, the eigenvectors of the empirical normalized graph Laplacian converge to the eigenvectors of the population normalized graph Laplacian–so long as (1) the minimum expected degree grows fast enough and (2) the eigengap that separates the leading eigenvalues from the smaller eigenvalues does not shrink too quickly. This theorem has consequences in addition to those related to spectral clustering.

Visualization is an important tool for social networks analysts (Freeman, 2000; Liotta, 2004; Wasserman and Faust, 1994). However, there is little statistical understanding of these techniques under stochastic models. Two visualization techniques, factor analysis and multidimensional scaling, have variations that utilize the eigenvectors of the graph Laplacian. Similar approaches were suggested for social networks as far back as the 1950's (Bock and Husain, 1952; Breiger, Boorman and Arabie, 1975). Koren (2005) suggests visualizing the graph using the eigenvectors of the unnormalized graph Laplacian. The analogous method for the normalized graph Laplacian would use the i^{th} row of X as the coordinates for the i^{th} node. Theorem 2.2 shows that, under the latent space model, this visualization is not much different than visualizing the graph by instead replacing X with $\mathscr{X}$. If there is structure in the latent space of a latent space model (for example, the $z_1, \ldots, z_n$ form clusters) and this structure is represented in the eigenvectors of the

population normalized graph Laplacian, then plotting the eigenvectors will potentially reveal this structure.

The Stochastic Blockmodel is a specific latent space model that satisfies these conditions. It has well defined clusters or blocks and Lemma 3.1 shows that, under weak assumptions, the eigenvectors of the population normalized graph Laplacian perfectly identify the block structure. Theorem 2.2 suggests that you could discover this clustering structure by using the visualization technique proposed by Koren (2005). The second main result, Theorem 3.1, goes further to suggest just how many nodes you might miscluster by running k-means on those points (this is spectral clustering). Theorem 3.1 proves that if (1) the minimum expected degree grows fast enough and (2) the smallest nonzero eigenvalue of the population normalized graph Laplacian shrinks slowly enough, then the proportion of nodes that are misclustered by spectral clustering vanishes in the asymptote.

The asymptotic framework applied in Theorem 3.1 allows the number of blocks to grow with the number of nodes; this is the first such high-dimensional clustering result. Allowing the number of clusters to grow is reasonable because as Leskovec *et al.* (2008) noted, large networks do not necessarily have large communities. In fact, in a wide range of empirical networks, the tightest communities have a roughly constant size. Allowing the number of blocks to grow with the number of nodes ensures the clusters do not become too large.

There are two main limitations of our results that are highlighted in the simulations in Section 4. First, Theorem 3.1 does not show that spectral clustering is consistent under the Stochastic Blockmodel; it only gives a bound on the number of misclassified nodes. Improving this bound is an area for future research. The second shortcoming is that Lemma 2.1, Theorem 2.2, and Theorem 3.1 all require the minimum expected degree to grow at the same rate as n (ignoring $\log n$ terms). In large empirical networks, the canonical edges are not dense enough to suggest this type of asymptotic framework. Section 5 suggests alternative definitions of edges that might increase the edge density. That said, studying spectral clustering under more realistic degree distributions is an area for future research.

APPENDIX A: PROOF OF THEOREM 2.1

First a proof of Lemma 2.1,

PROOF. By eigendecomposition, $M = \sum_{i=1}^{n} \lambda_i u_i u_i^T$ where $u_1, \ldots, u_n$ are

orthonormal and eigenvectors of M. So,

$$MM = \left(\sum_{i=1}^{n} \lambda_i u_i u_i^T\right)\left(\sum_{i=1}^{n} \lambda_i u_i u_i^T\right) = \sum_{i=1}^{n} \lambda_i^2 u_i u_i^T.$$

Right multiplying by any u_i yields $MMu_i = \lambda^2 u_i$. This proves one direction of part one in the lemma, if λ is an eigenvalue of M, then λ^2 is an eigenvalue of MM. It also proves part two of the lemma, all eigenvectors of M are also eigenvectors of MM.

To see that if λ^2 is an eigenvalue of MM, then λ or $-\lambda$ is an eigenvalue of M, notice that both M and MM have exactly n eigenvalues (counting multiplicities) because both matrices are real and symmetric. So, the previous paragraph specifies n eigenvalues of MM by squaring the eigenvalues of M. Because MM has exactly n eigenvalues, there are no other eigenvalues.

The rest of the proof is devoted to part three of the lemma. Let $MMv = \lambda^2 v$. By eigenvalue decomposition, $M = \sum_i \lambda_i u_i u_i^T$ and because $u_1, \ldots, u_n$ are orthonormal (M is real and symmetric) there exists $\alpha_1, \ldots, \alpha_n$ such that $v = \sum_i \alpha_i u_i$.

$$\begin{aligned}\lambda^2 \sum_i \alpha_i u_i = \lambda^2 v = MMv &= M\left(\sum_i \lambda_i u_i u_i^T v\right) = M\left(\sum_i \lambda_i \alpha_i u_i\right) \\ &= \sum_i \lambda_i \alpha_i M u_i = \sum_i \lambda_i^2 \alpha_i u_i\end{aligned}$$

By the orthogonality of the u_i's, it follows that $\lambda^2\alpha_i = \lambda_i^2\alpha_i$ for all i. So, if $\lambda_i^2 \neq \lambda^2$, then $\alpha_i = 0$. □

For $i = 1, \ldots, n$, define $c_i = \mathscr{D}_{ii}/n$ and $\tau = \min_{i=1,\ldots,n} c_i$.

LEMMA A.1. *If* $n^{1/2}/\log n > 2$,

$$\mathbb{P}\left(\|LL - \mathscr{L}\mathscr{L}\|_F \geq \frac{32\sqrt{2}\log n}{\tau^2\ n^{1/2}}\right) \leq 4n^{2-2\tau^2\log n}.$$

The main complication of the proof of Lemma A.1 is controlling the dependencies between the elements of LL. We do this with an intermediate step that uses the matrix

$$\tilde{L} = \mathscr{D}^{-1/2} W \mathscr{D}^{-1/2}$$

and two sets Γ and Λ. Γ constrains the matrix D, while Λ constrains the matrix $W\mathscr{D}^{-1}W$. These sets will be defined in the proof. To ease the notation, define

$$\mathbb{P}_{\Gamma\Lambda}(B) = \mathbb{P}(B \cap \Gamma \cap \Lambda)$$

where B is some event.

PROOF. This proof shows that under the sets Γ and Λ the probability of the norm exceeding $32\sqrt{2}\log(n)\ \tau^{-2}\ n^{-1/2}$ is exactly zero for large enough n and that the probability of Γ or Λ not happening is exponentially small. To ease notation, define $a = 32\sqrt{2}\log(n)\ \tau^{-2}\ n^{-1/2}$.

The diagonal terms behave differently than the off diagonal terms. So, break them appart.

$$\begin{aligned}
\mathbb{P}(\|LL - \mathscr{L}\mathscr{L}\|_F \geq a) &\leq \mathbb{P}_{\Gamma\Lambda}(\|LL - \mathscr{L}\mathscr{L}\|_F \geq a) + \mathbb{P}\left((\Gamma \cap \Lambda)^c\right) \\
&= \mathbb{P}_{\Gamma\Lambda}\left(\sum_{i,j}[LL - \mathscr{L}\mathscr{L}]_{ij}^2 \geq a^2\right) + \mathbb{P}\left((\Gamma \cap \Lambda)^c\right) \\
&\leq \mathbb{P}_{\Gamma\Lambda}\left(\sum_{i\neq j}[LL - \mathscr{L}\mathscr{L}]_{ij}^2 \geq a^2/2\right) \\
&\quad + \mathbb{P}_{\Gamma\Lambda}\left(\sum_{i}[LL - \mathscr{L}\mathscr{L}]_{ii}^2 \geq a^2/2\right) \\
&\quad + \mathbb{P}\left((\Gamma \cap \Lambda)^c\right)
\end{aligned}$$

First, address the sum over the off diagonal terms.

$$\begin{aligned}
&\mathbb{P}_{\Gamma\Lambda}\left(\sum_{i\neq j}[LL - \mathscr{L}\mathscr{L}]_{ij}^2 \geq a^2/2\right) \\
&\leq \mathbb{P}_{\Gamma\Lambda}\left(\cup_{i\neq j}\{[LL - \mathscr{L}\mathscr{L}]_{ij}^2 \geq \frac{a^2}{2n^2}\}\right) \qquad \text{(A.1)} \\
&\leq \sum_{i\neq j}\mathbb{P}_{\Gamma\Lambda}\left(|LL - \mathscr{L}\mathscr{L}|_{ij} \geq \frac{a}{\sqrt{2}n}\right) \\
&\leq \sum_{i\neq j}\mathbb{P}_{\Gamma\Lambda}\left(|LL - \tilde{L}\tilde{L}|_{ij} + |\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ij} \geq \frac{a}{\sqrt{2}n}\right) \\
&\leq \sum_{i\neq j}\left[\mathbb{P}_{\Gamma\Lambda}\left(|LL - \tilde{L}\tilde{L}|_{ij} \geq \frac{a}{\sqrt{8}n}\right)\right. \\
&\quad \left. + \mathbb{P}_{\Gamma\Lambda}\left(|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ij} \geq \frac{a}{\sqrt{8}n}\right)\right] \qquad \text{(A.2)}
\end{aligned}$$

The sum over the diagonal terms is similar,

$$\begin{aligned}\mathbb{P}_{\Gamma\Lambda}\left(\sum_i [LL - \mathscr{L}\mathscr{L}]_{ii}^2 \geq a^2/2\right) &\leq \sum_i \left[\mathbb{P}_{\Gamma\Lambda}\left(|LL - \tilde{L}\tilde{L}|_{ii} \geq \frac{a}{\sqrt{8n}}\right)\right. \\ &\left. + \mathbb{P}_{\Gamma\Lambda}\left(|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ii} \geq \frac{a}{\sqrt{8n}}\right)\right],\end{aligned}$$

with one key difference. In equation (A.1), the union bound address nearly n^2 terms. This yields the $1/n^2$ term in line (A.1). After taking the square root, each term has a lower bound with a factor of $1/n$. However, because there are only n terms on the diagonal, after taking the square root in the last equation above, the lower bound has a factor of $1/\sqrt{n}$.

To constrain the terms $|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ij}$ for $i = j$ and $i \neq j$, define

$$\Lambda = \bigcap_{i,j} \left\{ \Big| \sum_k (W_{ik}W_{jk} - p_{ijk})/c_k \Big| < n^{1/2} \log n \right\},$$

where

$$p_{ijk} = \begin{cases} p_{ik}p_{jk} & \text{if } i \neq j \\ p_{ik} & \text{if } i = j \end{cases}$$

for $p_{ij} = \mathscr{W}_{ij}$. We now show that for large enough n, and any $i \neq j$,

$$\mathbb{P}_\Lambda\left(|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ij} \geq \frac{a}{\sqrt{8}n}\right) = 0 \tag{A.3}$$

$$\mathbb{P}_\Lambda\left(|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ii} \geq \frac{a}{\sqrt{8n}}\right) = 0. \tag{A.4}$$

To see Equation (A.3), expand the left hand side of the inequality for $i \neq j$,

$$\begin{aligned}|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ij} &= \frac{1}{(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}} \left|\sum_k (W_{ik}W_{jk} - p_{ik}p_{jk})/\mathscr{D}_{kk}\right| \\ &= \frac{1}{n^2\sqrt{c_i c_j}} \left|\sum_k (W_{ik}W_{jk} - p_{ik}p_{jk})/c_k\right|\end{aligned}$$

This is bounded on Λ, yielding

$$|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ij} < \frac{\log n}{\tau n^{3/2}} \leq \frac{32\sqrt{2}\log n}{\sqrt{8}\tau^2\ n^{3/2}} = \frac{a}{\sqrt{8}n}.$$

So, Equation (A.3) holds for $i \neq j$. Equation (A.4) is different because $W_{ik}^2 = W_{ik}$. As a result, the diagonal of $\tilde{L}\tilde{L}$ is a biased estimator of the diagonal of $\mathscr{L}\mathscr{L}$.

$$\begin{aligned}
|\tilde{L}\tilde{L} - \mathscr{L}\mathscr{L}|_{ii} &= |\sum_k \frac{W_{ik}^2 - p_{ik}^2}{\mathscr{D}_{ii}\mathscr{D}_{kk}}| \\
&= |\sum_k \frac{W_{ik} - p_{ik}^2}{\mathscr{D}_{ii}\mathscr{D}_{kk}}| \\
&\leq |\sum_k \frac{W_{ik} - p_{ik}}{\mathscr{D}_{ii}\mathscr{D}_{kk}}| + |\sum_k \frac{p_{ik} - p_{ik}^2}{\mathscr{D}_{ii}\mathscr{D}_{kk}}| \\
&= \frac{1}{c_i n^2}\left(|\sum_k (W_{ik} - p_{ik})/c_k| + |\sum_k (p_{ik} - p_{ik}^2)/c_k|\right) \qquad \text{(A.5)}
\end{aligned}$$

Similarly to the $i \neq j$ case, the first term is bounded by $\log(n)\tau^{-1}n^{-3/2}$ on Λ. The second term is bounded by $\tau^{-2}n^{-1}$:

$$\begin{aligned}
\frac{1}{c_i n^2}\left|\sum_k (p_{ik} - p_{ik}^2)/c_k\right| &\leq \frac{1}{c_i n^2}\left|\sum_k 1/\tau\right| \\
&\leq \frac{1}{\tau^2 n}.
\end{aligned}$$

Substituting the value of a in reveals that on the set Λ, both terms in (A.5) are bounded by $a(2\sqrt{8n})^{-1}$. So, their their sum is bounded by $a(\sqrt{8n})^{-1}$, satisfying Equation (A.4).

This next part addresses the difference between LL and $\tilde{L}\tilde{L}$, showing that for large enough n, any $i \neq j$, and some set Γ,

$$\begin{aligned}
\mathbb{P}_{\Gamma\Lambda}(|LL - \tilde{L}\tilde{L}|_{ij} \geq \frac{a}{\sqrt{8n}}) &= 0 \\
\mathbb{P}_{\Gamma\Lambda}(|LL - \tilde{L}\tilde{L}|_{ii} \geq \frac{a}{\sqrt{8n}}) &= 0
\end{aligned}$$

It is enough to show that for any i and j,

$$\mathbb{P}_{\Gamma\Lambda}(|LL - \tilde{L}\tilde{L}|_{ij} \geq \frac{a}{\sqrt{8n}}) = 0. \qquad \text{(A.6)}$$

For $b(n) = \log(n)n^{-1/2}$, define $u(n) = 1 + b(n), l(n) = 1 - b(n)$. With

these define the following sets,

$$\begin{aligned}
\Gamma &= \bigcap_i \{D_{ii} \in \mathscr{D}_{ii}[l(n), u(n)]\} \\
\Gamma(1) &= \bigcap_i \left\{ \frac{1}{D_{ii}} \in \frac{1}{\mathscr{D}_{ii}} \left[u(n)^{-1}, l(n)^{-1}\right] \right\} \\
\Gamma(2) &= \bigcap_{i,j} \left\{ \frac{1}{(D_{ii}D_{jj})^{1/2}} \in \frac{1}{(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}} \left[u(n)^{-1}, l(n)^{-1}\right] \right\} \\
\Gamma(3) &= \bigcap_{i,j,k} \left\{ \frac{1}{D_{kk}(D_{ii}D_{jj})^{1/2}} \in \frac{\left[u(n)^{-2}, l(n)^{-2}\right]}{\mathscr{D}_{kk}(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}} \right\}
\end{aligned}$$

Notice that $\Gamma \subseteq \Gamma(1) \subseteq \Gamma(2)$ and $\Gamma \subseteq \Gamma(3)$. Define another set,

$$\Gamma(4) = \bigcap_{i,j,k} \left\{ \frac{1}{D_{kk}(D_{ii}D_{jj})^{1/2}} \in \frac{[1 - 16b(n), 1 + 16b(n)]}{\mathscr{D}_{kk}(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}} \right\}.$$

The next steps show that this set contains Γ. It is sufficient to show $\Gamma(3) \subset \Gamma(4)$. This is true because

$$\begin{aligned}
\frac{1}{u(n)^2} = \frac{1}{(1 + b(n))^2} &= \frac{b(n)^{-2}}{(b(n)^{-1} + 1)^2} > \frac{b(n)^{-2} - 1}{(b(n)^{-1} + 1)^2} \\
&= \frac{b(n)^{-1} - 1}{b(n)^{-1} + 1} = 1 - \frac{2}{b(n)^{-1} + 1} > 1 - 16\ b(n).
\end{aligned}$$

The 16 in the last bound is larger than it needs to be so that the upper and lower bounds in $\Gamma(4)$ are symmetric. For the other direction,

$$\begin{aligned}
\frac{1}{l(n)^2} = \frac{1}{(1 - b(n))^2} &= \frac{b(n)^{-2}}{(b(n)^{-1} - 1)^2} = \left(1 + \frac{1}{b(n)^{-1} - 1}\right)^2 \\
&= 1 + \frac{2}{b(n)^{-1} - 1} + \frac{1}{(b(n)^{-1} - 1)^2}.
\end{aligned}$$

We now need to bound the last two elements here. We are assuming, $\sqrt{n}/\log n > 2$. Equivalently, $1 - b(n) > 1/2$. So, we have both of the following:

$$\frac{1}{(b(n)^{-1} - 1)^2} < \frac{2}{b(n)^{-1} - 1} \quad \text{and} \quad \frac{2}{b(n)^{-1} - 1} = \frac{2b(n)}{1 - b(n)} < 8b(n).$$

Putting these together,

$$\frac{1}{l(n)^2} < 1 + 16\ b(n).$$

This shows that $\Gamma \subset \Gamma(4)$. Now, under the set Γ, and thus $\Gamma(4)$,

$$\begin{aligned}
|LL - \tilde{L}\tilde{L}|_{ij} &= \Big|\sum_k \left(\frac{W_{ik}W_{jk}}{D_{kk}(D_{ii}D_{jj})^{1/2}} - \frac{W_{ik}W_{jk}}{\mathscr{D}_{kk}(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}}\right)\Big| \\
&\leq \sum_k \Big|\frac{1}{D_{kk}(D_{ii}D_{jj})^{1/2}} - \frac{1}{\mathscr{D}_{kk}(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}}\Big| \\
&\leq \sum_k \Big|\frac{16\ b(n)}{\mathscr{D}_{kk}(\mathscr{D}_{ii}\mathscr{D}_{jj})^{1/2}}\Big| \\
&\leq \sum_k \frac{16\ b(n)}{\tau^2 n^2} \\
&\leq \frac{16\ b(n)}{\tau^2 n}.
\end{aligned}$$

This is equal to $a(\sqrt{8}n)^{-1}$, showing Equation (A.4) holds for all i and j.

The remaining step is to bound $\mathbb{P}((\Gamma \cap \Lambda)^c)$. Using the union bound this is less than or equal to $\mathbb{P}(\Gamma^c) + \mathbb{P}(\Lambda^c)$.

$$\begin{aligned}
\mathbb{P}(\Gamma^c) &= \mathbb{P}\left(\bigcup_i \{D_{ii} \notin \mathscr{D}_{ii}[1 - b(n), 1 + b(n)]\}\right) \\
&\leq \sum_i \mathbb{P}\left(\{D_{ii} \notin \mathscr{D}_{ii}[1 - b(n), 1 + b(n)]\}\right) \\
&< \sum_i 2\exp\left(-2\left(\frac{\mathscr{D}_{ii}\log n}{\sqrt{n}}\right)^2 \frac{1}{n}\right) \\
&\leq 2n\exp(-2\tau^2(\log n)^2) \\
&= 2n^{1-2\tau^2\log n}
\end{aligned}$$

Where the second to last inequality is by Hoeffding's inequality. The next

inequality is Hoeffding's.

$$\begin{aligned}
\mathbb{P}(\Lambda^c) &= \mathbb{P}\left(\bigcup_{i,j}\left\{\Big|\sum_k (W_{ik}W_{jk} - p_{ijk})/c_k\Big| > n^{1/2}\log n\right\}\right) \\
&= \sum_{i,j}\mathbb{P}\left(\Big|\sum_k (W_{ik}W_{jk} - p_{ijk}/c_k\Big| > n^{1/2}\log n\right) \\
&< \sum_{i,j} 2\exp\left(-2n(\log n)^2/\sum_k 1/c_k^2\right) \\
&\leq \sum_{i,j} 2\exp\left(-2(\log n)^2\tau^2\right) \\
&\leq 2n^2\exp\left(-2(\log n)^2\tau^2\right) \\
&\leq 2n^{2-2\tau^2\log n}.
\end{aligned}$$

Because W is symmetric, the independence of the $W_{ik}W_{jk}$ across k is not obvious. However, because $W_{ii} = W_{jj} = 0$, they are independent across k.

Putting the pieces together,

$$\begin{aligned}
&\mathbb{P}\left(\|LL - \mathscr{L}\mathscr{L}\|_F \geq \frac{32\sqrt{2}\log n}{\tau^2\ n^{1/2}}\right) \\
\leq\ &\mathbb{P}_{\Gamma\Lambda}(\|LL - \mathscr{L}\mathscr{L}\|_F \geq \frac{32\sqrt{2}\log n}{\tau^2\ n^{1/2}}) + \mathbb{P}\left((\Gamma\cap\Lambda)^c\right) \\
<\ &0 + 2n^{1-2\tau^2\log n} + 2n^{2-2\tau^2\log n}. \\
\leq\ &4n^{2-2\tau^2\log n}.
\end{aligned}$$

□

The following proves Theorem 2.1.

PROOF. Adding the n super- and subscripts to Lemma A.1, it states that if $n^{1/2}/\log n > 2$, then

$$\mathbb{P}\left(\|LL - \mathscr{L}\mathscr{L}\|_F \geq \frac{c\log n}{\tau^2\ n^{1/2}}\right) < 4n^{2-2\tau^2\log n}.$$

for $c = 32\sqrt{2}$. By assumption, for all $n > N$, $\tau_n^2\log n > 2$. This implies that $2 - 2\tau_n^2\log n < -2$ for all $n > N$. Rearranging and summing over n, for any

fixed $\epsilon > 0$,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{\|L^{(n)}L^{(n)} - \mathscr{L}^{(n)}\mathscr{L}^{(n)}\|_F}{c\tau_n^{-2}\ \log(n) n^{-1/2}/\epsilon} \geq \epsilon \right) &\leq N + 4 \sum_{n=N+1}^{\infty} n^{2-2\tau_n^2 \log n} \\ &\leq N + 4 \sum_{n=N+1}^{\infty} n^{-2},\end{aligned}$$

which is a summable sequence. By the Borel-Cantelli Theorem,

$$\|L^{(n)}L^{(n)} - \mathscr{L}^{(n)}\mathscr{L}^{(n)}\|_F = o(\tau_n^{-2}\ \log(n)\ n^{-1/2}) \quad a.s.$$

□

APPENDIX B: DAVIS-KAHAN THEOREM

The statement of the theorem below and the preceding explanation come largely from von Luxburg (2007). For a more detailed account of the Davis-Kahan Theorem see Stewart and Sun (1990).

To avoid the issues associated with multiple eigenvalues, this theorem's original statement is instead about the subspace formed by the eigenvectors. For a distance between subspaces, the theorem uses "canonical angles," which are also known as "principal angles." Given two matrices M_1 and M_2 both in $\mathcal{R}^{n\times p}$ with orthonormal columns, the singular values $(\sigma_1, \ldots, \sigma_p)$ of $M_1'M_2$ are the cosines of the principal angles $(\cos\Theta_1, \ldots, \cos\Theta_p)$ between the column space of M_1 and the column space of M_2. Define $\sin\Theta(M_1, M_2)$ to be a diagonal matrix containing the sine of the principal angles of $M_1'M_2$ and define

$$d(M_1, M_2) = \|\sin\Theta(M_1, M_2)\|_F, \tag{B.1}$$

which can be expressed as $(p - \sum_{j=1}^{p} \sigma_j^2)^{1/2}$ by using the identity $\sin^2\theta = 1 - \cos^2\theta$.

PROPOSITION B.1. ***(Davis-Kahan)*** *Let $S \subset \mathcal{R}$ be an interval. Denote $\mathcal{X}$ as an orthonormal matrix whose column space is equal to the eigenspace of $\mathscr{L}\mathscr{L}$ corresponding to the eigenvalues in $\lambda_S(\mathscr{L}\mathscr{L})$ (more formally, the column space of $\mathcal{X}$ is the image of the spectral projection of $\mathscr{L}\mathscr{L}$ induced by $\lambda_S(\mathscr{L}\mathscr{L})$). Denote by X the analogous quantity for LL. Define the distance between S and the spectrum of $\mathscr{L}\mathscr{L}$ outside of S as*

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

Then the distance $d(\mathcal{X}, X) = \|\sin\Theta(\mathcal{X}, X)\|_F$ *between the column spaces of* $\mathcal{X}$ *and* X *is bounded by*

$$d(X, \mathcal{X}) \leq \frac{\|LL - \mathscr{L}\mathscr{L}\|_F}{\delta}.$$

In the theorem $\mathscr{L}\mathscr{L}$ and LL can be replaced by any two symmetric matrices. The rest of this section converts the bound on $d(X, \mathcal{X})$ to a bound on $\|X - \mathcal{X}O\|_F$, where O is some orthonormal rotation. For this, we will make an additional assumption that $\mathcal{X}$ and X have the same dimension. Assume there exists $S \subset \mathcal{R}$ containing k eigenvalues of $\mathscr{L}\mathscr{L}$ and k eigenvalues of LL, but containing no other eigenvalues of either matrix. Because LL and $\mathscr{L}\mathscr{L}$ are symmetric, its eigenvectors can be defined to be orthonormal. Let the columns of $\mathscr{X} \in \mathcal{R}^{n\times k}$ be k orthonormal eigenvectors of $\mathscr{L}\mathscr{L}$ corresponding to the k eigenvalues contained in S. Let the columns of $X \in \mathcal{R}^{n\times k}$ be k orthonormal eigenvectors of LL corresponding to the k eigenvalues contained in S. By singular value decomposition, there exist orthonormal matrices U, V and diagonal matrix Σ such that $\mathscr{X}^T X = U\Sigma V^T$. The singular values, $\sigma_1, \ldots, \sigma_k$, down the diagonal of Σ are the cosines of the principal angles between the columns space of X and the column space of $\mathscr{X}$.

Although the Davis-Kahan Theorem is a statement regarding the principal angles, a few lines of algebra shows that it can be extended to a bound on the Frobenius norm between the matrix X and $\mathscr{X}UV^T$, where the matrix UV^T is an orthonormal rotation.

$$\begin{aligned}
\frac{1}{2}\|X - \mathscr{X}UV^T\|_F^2 &= \frac{1}{2}trace((X - \mathscr{X}UV^T)^T(X - \mathscr{X}UV^T)) \\
&= \frac{1}{2}trace(VU^T\mathscr{X}^T\mathscr{X}UV^T + X^TX - 2VU^T\mathscr{X}^TX) \\
&= \frac{1}{2}\left(k + k - 2trace(VU^T\mathscr{X}^TX)\right) \\
&\leq \frac{\|LL - \mathscr{L}\mathscr{L}\|_F^2}{\delta^2},
\end{aligned}$$

where the last inequality is explained below. It follows from a property of the trace, the fact that the singular values are in $[0, 1]$, the trigonometric identity $\cos^2\theta = 1 - \sin^2\theta$ and the Davis-Kahan Theorem:

$$\begin{aligned}
trace(VU^T\mathscr{X}^TX) &= \sum_{i=1}^{k}\sigma_i \geq \sum_{i=1}^{k}(\cos\Theta_i)^2 = \sum_{i=1}^{k}1 - (\sin\Theta_i)^2 \\
&= k - (d(X, \mathscr{X}))^2 \geq k - \frac{\|LL - \mathscr{L}\mathscr{L}\|_F^2}{\delta^2}.
\end{aligned}$$

This shows that the Davis-Kahan Theorem can instead be thought of as a bounding $\|\mathscr{X}UV^T - X\|_F^2$ instead of $d(\mathscr{X}, X)$. The matrix O in Theorem 2.1 is equal to UV^T. In this way, it is dependent on X and $\mathcal{X}$.

APPENDIX C: PROOF OF THEOREM 2.3

PROOF. By Lemma 2.1, the column vectors of X_n are eigenvectors of $L^{(n)}L^{(n)}$ corresponding to all the eigenvalues in $\lambda_{S_n}(L^{(n)}L^{(n)})$. For the application of the Davis-Kahan Theorem, this means that the column space of X_n is the image of the spectral projection of $L^{(n)}L^{(n)}$ induced by $\lambda_{S_n}(L^{(n)}L^{(n)})$. Similarly, for the column vectors of $\mathscr{X}_n$, the matrix $\mathscr{L}^{(n)}\mathscr{L}^{(n)}$, and the set $\lambda_{S_n}(\mathscr{L}^{(n)}\mathscr{L}^{(n)})$.

Recall that $\bar{\lambda}_1^{(n)} \geq \cdots \geq \bar{\lambda}_n^{(n)}$ are defined to be the eigenvalues of symmetric matrix $\mathscr{L}^{(n)}\mathscr{L}^{(n)}$ and $\lambda_1^{(n)} \geq \cdots \geq \lambda_n^{(n)}$ are defined to be the eigenvalues of symmetric matrix $L^{(n)}L^{(n)}$. From Equation (2.2),

$$\max_i |\lambda_i^{(n)} - \bar{\lambda}_i^{(n)}| = o\left(\frac{\log n}{\tau_n^2 n^{1/2}}\right).$$

By assumption, $\tau_n^2 > 2/\log n$. So,

$$\frac{\log n}{\tau_n^2 n^{1/2}} < \frac{(\log n)^2}{2n^{1/2}} = O(\min\{\delta_n, \delta_n'\}),$$

where the last step follows by assumption. Thus,

$$\max_i |\lambda_i^{(n)} - \bar{\lambda}_i^{(n)}| = o(\min\{\delta_n, \delta_n'\}).$$

This means that, eventually, $\lambda_i^{(n)} \in S_n$ if and only if $\bar{\lambda}_i^{(n)} \in S_n$. Thus the number of elements in $\lambda_{S_n}(\mathscr{L}^{(n)}\mathscr{L}^{(n)})$ is eventually equal to the number of elements in $\lambda_{S_n}(L^{(n)}L^{(n)})$ implying that X_n and $\mathscr{X}_n$ will eventually have the same number of columns, $k_n = \mathscr{K}_n$.

Once X_n and $\mathscr{X}_n$ have the same number of columns, define matrices U_n and V_n with singular value decomposition: $\mathscr{X}_n^T X_n = U_n \Sigma_n V_n^T$. Define $O_n = U_n V_n^T$. The result follows from the Davis-Kahan Theorem and Theorem 2.1:

$$\|X_n - \mathscr{X}_n O_n\|_F \leq \frac{2\|L^{(n)}L^{(n)} - \mathscr{L}^{(n)}\mathscr{L}^{(n)}\|_F}{\delta_n} = o\left(\frac{\log n}{\delta_n \tau_n^2\ n^{1/2}}\right) \quad a.s.$$

□

APPENDIX D: STOCHASTIC BLOCKMODEL

Below is a proof of Lemma 3.1

PROOF. First, construct the matrix $B_L \in R^{k\times k}$ such that $\mathscr{L} = ZB_LZ^T$. define $D_B = diag(BZ^T\mathbf{1}_n) \in \mathcal{R}^{k\times k}$ where $\mathbf{1}_n$ is a vector of ones in $\mathcal{R}^n$. For any i, j

$$\mathscr{L}_{ij} = \frac{\mathscr{W}_{ij}}{\sqrt{\mathscr{D}_{ii}\mathscr{D}_{jj}}} = z_i D_B^{-1/2} B D_B^{-1/2} (z_j)^T$$

Define $B_L = D_B^{-1/2} B D_B^{-1/2}$. It follows that $\mathscr{L}_{ij} = (ZB_LZ^T)_{ij}$ and thus $\mathscr{L} = ZB_LZ^T$.

Because B is symmetric, so is B_L and so is $(Z^TZ)^{1/2}B_L(Z^TZ)^{1/2}$. Notice that

$$det((Z^TZ)^{1/2}B_L(Z^TZ)^{1/2}) = det((Z^TZ)^{1/2})det(B_L)det((Z^TZ)^{1/2}) > 0.$$

By eigenvector decomposition, define $V \in R^{k\times k}$ and diagonal matrix $\Lambda \in R^{k\times k}$ such that

$$(Z^TZ)^{1/2}B_L(Z^TZ)^{1/2} = V\Lambda V^T. \tag{D.1}$$

Because the determinant of the left hand side of Equation (D.1) is greater than zero, none of the eigenvalues in Λ are equal to zero. Left multiply Equation (D.1) by $Z(Z^TZ)^{-1/2}$ and right multiply by $(Z^TZ)^{-1/2}Z^T$. This shows

$$ZB_LZ^T = Z\mu\Lambda(Z\mu)^T, \tag{D.2}$$

where $\mu = (Z^TZ)^{-1/2}V$. Notice that $(Z\mu)^T(Z\mu) = I_k$, the $k\times k$ identity matrix. So, right multiplying Equation (D.2) by $Z\mu$ shows that the columns of $Z\mu$ are eigenvectors of $ZB_LZ^T = \mathscr{L}$ with the eigenvalues down the diagonal of Λ. Equation (D.2) shows that these are the only nonzero eigenvalues.

It remains to prove equivalence statement (3.2). Notice

$$det(\mu) = det((Z^TZ)^{-1/2})det(V) > 0.$$

So, μ^{-1} exists and statement (3.2) follows. □

The following is a proof of Lemma 3.2.

PROOF. The following statement is the essential ingredient to prove Lemma 3.2.

$$z_i \neq z_j, \quad \text{then} \quad \|z_i\mu - z_j\mu\|_2 \geq \sqrt{2/P} \tag{D.3}$$

The proof of statement (D.3) requires the following definition,

$$\|\mu\|_m^2 = \min_{x:\|x\|_2=1} \|x\mu\|_2^2.$$

Notice that

$$\|\mu\|_m^2 = \min_{x:\|x\|_2=1} x\mu\mu^T x^T = \min_{x:\|x\|_2=1} x(Z^TZ)^{-1}x^T = 1/P.$$

So,

$$\|z_i\mu - z_j\mu\|_2 = \|(z_i - z_j)\mu\|_2 \geq \sqrt{2}\|\mu\|_m = \sqrt{2/P},$$

Proving statement (D.3). The proof of Lemma 3.2 follows,

$$\|c_iO - z_j\mu\|_2 \geq \|z_i\mu - z_j\mu\|_2 - \|c_iO - z_i\mu\|_2 \geq \sqrt{\frac{2}{P}} - \frac{1}{2}\sqrt{\frac{2}{P}} = \frac{1}{\sqrt{2P}}$$

□

The following is a proof of Theorem 3.1.

PROOF. Define $X \in R^{n\times k}$ to contain the eigenvectors of L corresponding to the largest k eigenvalues and define

$$C = \text{argmin}_{M\in\mathscr{R}(n,k)}\|X - M\|_F^2$$

where $\mathscr{R}(n,k)$ is defined as follows,

$$\mathscr{R}(n,k) = \{M \in R^{n\times k} : \ M \text{ has no more than } k \text{ unique rows}\}.$$

Notice that

$$\min_{M\in\mathscr{R}(n,k)} \|X - M\|_F^2 = \min_{\{m_1,\ldots,m_k\}\subset\mathcal{R}^k} \sum_i \min_g \|x_i - m_g\|_2^2.$$

This shows that the ith row of C is equal to c_i as defined in Definition 3. Because $Z\mu O \in \mathscr{R}(n,k)$, notice that

$$\|X - C\|_2 \leq \|X - Z\mu O\|_2. \tag{D.4}$$

By the triangle inequality and inequality D.4,

$$\|C - Z\mu O\|_2 \leq \|C - X\|_2 + \|X - Z\mu O\|_2 \leq 2\|X - Z\mu O\|_2.$$

In the notation of Theorem 2.2, define $S_n = [\lambda_{k_n}^2/2, 2]$. Then, $\delta = \delta' = \lambda_{k_n}^2/2$. By assumption, $n^{-1/2}(\log n)^2 = O(\lambda_{k_n}^2) = O(\min\{\delta, \delta'\})$. This implies that the results from Theorem 2.2 hold. Putting the pieces together,

$$\begin{aligned} |\mathscr{M}| \leq \sum_{i \in \mathscr{M}} 1 &\leq 2P_n \sum_{i \in \mathscr{M}} \|c_i - z_i \mu O\|_2^2 \\ &\leq 2P_n \|C - Z\mu O\|_F^2 \\ &\leq 8P_n \|X - Z\mu O\|_F^2 \\ &= o\left(\frac{P_n (\log n)^2}{n \lambda_{k_n}^4 \tau_n^4}\right) \quad a.s. \end{aligned}$$

□

In Example 2 in Section 3, it was claimed that

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

The following is a proof of that statement.

Define $B \in \mathcal{R}^{k \times k}$ such that

$$B = pI_k + r\mathbf{1}_k\mathbf{1}_k^T$$

where $I_k \in \mathcal{R}^{k \times k}$ is the identity matrix, $\mathbf{1}_k \in \mathcal{R}^k$ is a vector of ones, $r \in (0, 1)$ and $p \in (0, 1 - r)$. Assume that p and r are fixed and k can grow with n. Let $Z \in \{0, 1\}^{n \times k}$ be such that $Z^T\mathbf{1}_n = s\mathbf{1}_k$. This guarantees that all k groups have equal size s. The Stochastic Blockmodel in Example 2 has the population adjacency matrix, $\mathscr{W} = ZBZ^T$.

Define

$$B_L = \frac{1}{nr + sp}\left(pI_k + r\mathbf{1}_k\mathbf{1}_k^T\right).$$

From the argument in the proof of Lemma 3.1, $\mathscr{L}$ has the same nonzero eigenvalues as $(Z^TZ)^{1/2}B_L(Z^TZ)^{1/2} \in R^{k \times k}$. Let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of $(Z^TZ)^{1/2}B_L(Z^TZ)^{1/2} = (s^{1/2}I_k)B_L(s^{1/2}I_k) = sB_L$. Notice that $\mathbf{1}_k$ is an eigenvector with eigenvalue 1.

$$sB_L\mathbf{1}_k = \frac{s}{nr + sp}\left(pI_k + r\mathbf{1}_k\mathbf{1}_k^T\right)\mathbf{1}_k = \frac{s(p + kr)}{nr + sp}\mathbf{1}_k = \mathbf{1}_k$$

Let $\lambda_1 = 1$. Define

$$\mathcal{U} = \{u : \|u\|_2 = 1,\ u^T\mathbf{1} = 0\}.$$

Notice that for all $u \in \mathcal{U}$,

$$sB_L u = \frac{s}{nr+sp}\left(pI_k + r\mathbf{1}_k\mathbf{1}_k^T\right)u = \frac{sp}{nr+sp}u. \tag{D.5}$$

Equation (D.5) implies that for $i > 1$,

$$\lambda_i = \frac{sp}{nr+sp}.$$

This is also true for $i = k$.

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

This is the smallest nonzero eigenvalue of $\mathscr{L}$.

REFERENCES

Airoldi, E., Blei, D., Fienberg, S. and Xing, E. (2008). Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research* **9** 1981–2014.

Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of modern physics* **74** 47–97.

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509.

Belkin, M. (2003). Problems of learning on manifolds PhD thesis, The University of Chicago.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15** 1373–1396.

Belkin, M. and Niyogi, P. (2008). Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences* **74** 1289–1308.

Bhatia, R. (1987). *Perturbation bounds for matrix eigenvalues.*

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

Bock, R. and Husain, S. (1952). Factors of the Tele. *Sociometry* **15** 206–19.

Bousquet, O., Chapelle, O. and Hein, M. (2004). Measure based regularization. *Advances in Neural Information Processing Systems* **16**.

Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z. and Wagner, D. (2007). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 172–188.

Breiger, R., Boorman, S. and Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology* **12** 328–383.

Champion, M. (1998). How many atoms make up the universe? http://www.madsci.org/posts/archives/oct98/905633072.As.r.html.

Choi, D., Wolfe, P. and Airoldi, E. (2010). Stochastic blockmodels with growing number of classes. *Arxiv preprint arXiv:1011.4644*.

Coifman, R., Lafon, S., Lee, A., Maggioni, M., Nadler, B., Warner, F. and Zucker, S. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **102** 7426–7431.

Condon, A. and Karp, R. (1999). *Algorithms for graph partitioning on the planted partition model*.

Donath, W. and Hoffman, A. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development* **17** 420–425.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen* **6** 156.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal* **23** 298–305.

Fjällström, P. (1998). Algorithms for graph partitioning: A survey. *Computer and Information Science* **3**.

Fortunato, S. (2009). Community detection in graphs. *arXiv* **906**.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81** 832–842.

Freeman, L. (2000). Visualizing social networks. *Journal of social structure* **1** 4.

Giné, E. and Koltchinskii, V. (2006). Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. *Lecture Notes-Monograph Series* 238–259.

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99** 7821.

Goldenberg, A., Zheng, A., Fienberg, S. and Airoldi, E. (2009). A survey of statistical network models. *Foundations and Trends in Machine Learning, to appear*.

Hagen, L. and Kahng, A. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design* **11** 1074–1085.

Handcock, M., Raftery, A. and Tantrum, J. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society-Series A* **170** 301–354.

Hein, M. (2006). Uniform convergence of adaptive graph-based regularization. *Lecture Notes in Computer Science* **4005** 50.

Hein, M., Audibert, J. and von Luxburg, U. (2005). From graphs to manifolds-weak and strong pointwise consistency of graph Laplacians. In *Proceedings of the 18th Conference on Learning Theory (COLT)* 470–485. Springer.

Hendrickson, B. and Leland, R. (1995). An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM Journal on Scientific Computing* **16** 452–469.

Hoff, P., Raftery, A. and Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098.

Holland, P., Laskey, K. and Leinhardt, S. (1983). Stochastic blockmodels: Some first steps. *Social Networks* **5** 109–137.

Holland, P. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76** 33–50.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*. Springer Verlag.

Klein, D. and Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry* **12** 81–95.

Koren, Y. (2005). Drawing graphs by eigenvectors: Theory and practice. *Computers and Mathematics with Applications* **49** 1867–1888.

Lafon, S. (2004). Diffusion maps and geometric harmonics PhD thesis, Yale University.

Leskovec, J., Lang, K., Dasgupta, A. and Mahoney, M. (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.

Liotta, G. (2004). Graph Drawing. *Lecture Notes in Computer Science* **2912**.

Meilă, M. and Shi, J. (2001). A random walks view of spectral segmentation. *AI and Statistics (AISTATS)* **2001**.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E* **69** 26113.

Nowicki, K. and Snijders, T. (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association* **96**.

Pothen, A., Simon, H. and Liou, K. (1990). Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications* **11** 430.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* **22** 888–905.

Snijders, T. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14** 75–100.

Spielman, D. and Teng, S. (2007). Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and its Applications* **421** 284–305.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci* **1** 801–804.

Stewart, G. G. W. and Sun, J. (1990). Matrix perturbation theory.

Traud, A., Kelsic, E., Mucha, P. and Porter, M. (2008). Community structure in online collegiate social networks. *arXiv* **809**.

Van Driessche, R. and Roose, D. (1995). An improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel Computing* **21** 29–48.

Van Duijn, M., Snijders, T. and Zijlstra, B. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica* **58** 234–254.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing* **17** 395–416.

von Luxburg, U., Belkin, M. and Bousquet, O. (2008). Consistency of spectral clustering. *Annals of Statistics* **36** 555.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge Univ Pr.

Watts, D. and Strogatz, S. (1998). Collective dynamics of small-worldnetworks. *Nature* **393** 440–442.

Department of Statistics
University of California
Berkeley, CA 94720, USA
E-mail: karlrohe@stat.berkeley.edu
sourav@stat.berkeley.edu
binyu@stat.berkeley.edu