

# Covariate Assisted Spectral Clustering

Norbert Binkiewicz, Joshua T. Vogelstein, and Karl Rohe\*

## Abstract

Biological and social systems consist of myriad interacting units. The interactions can be intuitively represented in the form of a graph or network. Measurements of these graphs can reveal the underlying structure of these interactions, which provides insight into the systems that generated the graphs. Moreover, in applications such as neuroconnectomics, social networks, and genomics, graph data is accompanied by contextualizing measures on each node. We leverage these node covariates to help uncover latent communities in a graph, using a modification of spectral clustering. Statistical guarantees are provided under a joint mixture model that we call the Node Contextualized Stochastic Blockmodel, including a bound on the mis-clustering rate. For most simulated conditions, covariate assisted spectral clustering yields superior results relative to both regularized spectral clustering without node covariates and an adaptation of canonical correlation analysis. We apply covariate assisted spectral clustering to large brain graphs derived from diffusion MRI data, using the node locations or neurological region membership as covariates. In both cases, covariate assisted spectral clustering yields clusters that are easier to interpret neurologically.

*Keywords:* Network, Node Attribute, Laplacian, Brain Graph, Stochastic Blockmodel

---

\*Norbert Binkiewicz is a PhD candidate at the Department of Statistics, University of Wisconsin-Madison (E-mail: [norbert@stat.wisc.edu](mailto:norbert@stat.wisc.edu)), Joshua T. Vogelstein is an Assistant Professor at the Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University (E-mail: [jovo@jhu.edu](mailto:jovo@jhu.edu)) and Karl Rohe is an Assistant Professor at the Department of Statistics, University of Wisconsin-Madison (E-mail: [karlrohe@stat.wisc.edu](mailto:karlrohe@stat.wisc.edu)). This research was supported by NIH grant 5T32HL083806-08 and NSF grant DMS-1309998. The authors would like to thank Tai Qin and Zoe Russek for helpful comments and suggestions.

# 1 INTRODUCTION

Modern experimental techniques in areas such as genomics and brain imaging generate vast amounts of structured data. The data contain valuable information about the relationships of genes or brain regions. Studying these relationships is essential for solving challenging scientific problems, but few computationally feasible statistical techniques incorporate both the structure and diversity of these data.

A common approach to understanding the behavior of a complex biological or social system is to first discover blocks of highly interconnected units, also known as communities or clusters, that serve or contribute to a common function. These can be genes that are involved in a common pathway or areas in the brain with a common neurological function. Typically, we only observe the pairwise relationships between the units. These can be naturally represented in the form of a graph or network. Analyzing networks has become an important part of social and biological sciences. Examples of such networks include gene regulatory networks, friendship networks, and brain graphs. If we can discover the underlying block structure of such graphs, we can gain insight from the common characteristics or functions of the units within a block.

Extant research has extensively studied the algorithmic and theoretical aspects of finding node clusters within a graph. This includes Bayesian, maximum likelihood, and spectral approaches. Unlike model based methods, spectral clustering is a relaxation of a cost minimization problem and has shown to be effective in various settings (Ng et al. 2001; Von Luxburg 2007). Modifications of spectral clustering, such as regularized spectral clustering, are accurate even for sparse networks (Chaudhuri et al. 2012; Amini et al. 2013; Qin and Rohe 2013). On the other hand, certain Bayesian methods offer additional flexibility in how nodes are assigned to blocks, allowing for a single node to belong to multiple blocks or a mixture of blocks (Nowicki and Snijders 2001; Airoldi et al. 2008). Maximum likelihood

approaches can enhance interpretability by embedding nodes in a latent social space and providing methods for quantifying statistical uncertainty (Hoff et al. 2002; Handcock et al. 2007; Amini et al. 2013). Ultimately, for large graphs, spectral clustering is one of very few computationally feasible methods that has an algorithmic guarantee for finding the globally optimal partition.

The diverse structured data generated by modern technologies often contain additional measurements that can be represented as graph node attributes or covariates. For example, these could be the personal profile information in a friendship network or the spatial location of a brain region in a brain graph. There are two potential advantages of utilizing node covariates in graph clustering. First, if the covariates and the graph have common latent structure, then the node covariates provide additional information to help estimate this structure. Even if the covariates and the graph do not share exactly the same structure, some similarity is sufficient for the covariates to assist in the discovery of the graph structure. Second, by using node covariates in the clustering procedure, we enhance the relative homogeneity of covariates within a cluster and filter out partitions that fail to align with the important covariates. This allows for easy contextualization of the clusters in terms of the member nodes' covariates, providing a natural way to interpret the clusters.

Methods that utilize both node covariates and the graph to cluster the nodes have previously been introduced, but many of these methods rely on ad hoc or heuristic approaches and none provide any theoretical guarantees for statistical estimation. Most existing methods can be broadly classified into Bayesian approaches, spectral techniques, and heuristic algorithms. Many Bayesian models focus on categorical node covariates and are often computationally expensive (Chang et al. 2010; Balasubramanyan and Cohen 2011). A recent Bayesian model proposed by Yang et al. (2013) can discover multi-block membership of nodes with binary node covariates. This method has linear update time in the network size, but does not guarantee linear time convergence. Heuristic algorithms use various approaches, including

(a) embedding the network in a vector space, at which point more traditional methods can be applied to the vector data (Gibert et al. 2012) or (b) using the covariates to augment the graph and applying other graph clustering methods that tune the relative weights of node-to-node and node-to-covariate edges (Zhou et al. 2009). A spectral approach, which is commonly used to incorporate node covariates, directly alters the edge weights based on the similarity of the corresponding nodes’ covariates, and then uses traditional spectral clustering on the weighted graph (Neville et al. 2003; Gunnemann et al. 2013).

This work introduces a spectral approach called covariate assisted spectral clustering (CASC). This approach adds the covariance matrix of the node covariates to the regularized graph Laplacian, boosting the signal in the top eigenvectors of the sum, which is then used for spectral clustering. A tuning parameter is employed to adjust the relative weight of the covariates and the graph; Section 2.4 proposes a way to choose this tuning parameter. A similar framework can be used to jointly cluster multiple graphs (Eynard et al. 2012). Variants of CASC have previously been introduced. Both approaches were derived by first considering an optimization problem to minimize the weighted sum of the k-means objective function and a graph cut objective function. Then, a solution to the spectral relaxation of the original problem was obtained. Wang et al. (2009) decided against using an additive method similar to CASC because setting the method’s tuning parameter is a nonconvex problem. They chose to investigate a method which uses the product of the generalized inverse of the graph Laplacian and the covariate matrix instead. Shiga et al. (2007) recognized the advantage of having a tuning parameter to balance the contribution of the graph and the covariates, but they did not use the Stochastic Blockmodel to study their method. The full utility and flexibility of these types of methods have not yet been presented, and neither paper derives any statistical results about the performance of such methods. In contrast, we were initially motivated to develop CASC by its intuitive interpretation and propensity for theoretical analysis.

Very few of the clustering methods that employ both node covariates and the graph offer any theoretical results and, to our knowledge, this paper gives the first statistical guarantee for these types of approaches. We define the Node Contextualized Stochastic Blockmodel (NC-SBM), which combines the Stochastic Block model with a block mixture model for node covariates (Definition 3.1). Under this model, a bound on the mis-clustering rate of CASC is established (Theorem 3.6). A general lower bound is also derived, demonstrating the conditions under which an algorithm using both the node covariates and the graph can give more accurate clusters than any algorithm using only the node covariates or the graph (Theorem 3.7).

For comparison, an alternative method based on an adaptation of classical canonical correlation analysis is introduced (Hotelling 1936). It uses the product of the regularized graph Laplacian and the covariate matrix as the input to the spectral clustering algorithm. Simulations indicate that canonical correlation performs worse than CASC under the NC-SBM with Bernoulli covariates. However, canonical correlation analysis clustering is computationally faster than CASC and does not require any tuning. In contrast, CASC depends on a single tuning parameter, which interpolates between spectral clustering with only the graph and only the covariates. This parameter can be set without prior knowledge by using an objective function (e.g. the within cluster sum of squares). Some intuitive results for determining what range of tuning parameter values should be considered are provided in the description of the optimization procedure in Section 2.4. Alternatively, the tuning parameter can be set using prior knowledge or to ensure the clusters achieve some desired quality, such as spatial cohesion. As an illustrative example, Section 5 studies diffusion MRI derived brain graphs with CASC using two different sets of node covariates. The first analysis uses spatial location. This produces clusters that are more spatially coherent than those obtained using regularized spectral clustering alone, making them easier to interpret neurologically. The second analysis uses neurological region membership, which yields partitions that closely

align with neurological regions while allowing for patient-wise variability based on brain graph connectivity.

## 2 SPECTRAL CLUSTERING WITH NODE COVARIATES

### 2.1 Notation

Let  $G(E, V)$  be a graph where  $V$  is the set of vertices or nodes and  $E$  is the set of edges, which represent relationships between the nodes. Let  $N$  be the number of nodes. Index the nodes in  $V = \{1, \dots, N\}$ , then  $E$  contains a pair  $(i, j)$  if there is an edge between nodes  $i$  and  $j$ . The graph's edge set can be represented as the adjacency matrix  $\mathbf{A} \in \{0, 1\}^{N \times N}$ , where  $A_{ij} = A_{ji} = 1$  if  $(i, j) \in E$  and  $A_{ij} = A_{ji} = 0$  otherwise. We restrict ourselves to studying undirected and unweighted graphs, although with small modifications most of our results apply to directed and weighted graphs as well.

Define the regularized graph Laplacian as

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

where  $\mathbf{D}_\tau = \mathbf{D} + \tau \mathbf{I}$  and  $\mathbf{D}$  is a diagonal matrix with  $D_{ii} = \sum_j A_{ij}$ . The regularization parameter  $\tau$  is treated as a constant, and is included to improve spectral clustering performance on sparse graphs (Chaudhuri et al. 2012). Throughout, the tuning parameter will be set to  $\tau = N^{-1} \sum_i D_{ii}$ , or the average node degree (Qin and Rohe 2013).

For the graph  $G(E, V)$ , let each node in the set  $V$  have an associated bounded covariate vector  $\mathbf{X}_i \in [-J, J]^R$ , and let  $\mathbf{X} \in [-J, J]^{N \times R}$  be the covariate matrix where each row corresponds to a node covariate vector. Let  $\|\cdot\|$  denote the spectral norm and  $\|\cdot\|_F$  denote the Frobenius norm. For a sequence  $\{a_N\}$  and  $\{b_N\}$ ,  $a_N = \Theta(b_N)$  if and only if  $b_N = O(a_N)$  and  $a_N = O(b_N)$ .

## 2.2 Spectral Clustering Algorithm

The spectral clustering algorithm has been employed to cluster graph nodes using various functions of the adjacency matrix. For instance, applying the algorithm to  $\mathbf{L}_\tau$  corresponds to regularized spectral clustering, where the value of the regularization parameter is set prior to running the algorithm. All of the methods we consider will employ this algorithm, but will use a different input matrix (e.g.  $\mathbf{L}_\tau$ ,  $\tilde{\mathbf{L}}$ ,  $\mathbf{L}^{CCA}$  as defined later).

### Spectral Clustering

1. Given an input matrix  $\mathbf{W}$  and number of clusters  $K$ , find the eigenvectors  $\mathbf{U}_1, \dots, \mathbf{U}_K \in \mathbb{R}^N$  corresponding to the  $K$  largest eigenvalues of  $\mathbf{W}$ .
2. Use the eigenvectors as columns to form the matrix  $\mathbf{U} = [\mathbf{U}_1, \dots, \mathbf{U}_K] \in \mathbb{R}^{N \times K}$ .
3. Consider each row of  $\mathbf{U}$  and treat it as a point in  $\mathbb{R}^K$ . Run k-means with  $K$  clusters on these points.
4. If the  $i$ th row of  $\mathbf{U}$  falls in the  $k$ th cluster, assign node  $i$  to cluster  $k$ .

## 2.3 Combining Graph Structure and Node Covariates

To take advantage of available graph and node covariate data in graph clustering, it is necessary to employ methods that incorporate both of these data types. As discussed in the introduction, spectral clustering has many advantages over other graph clustering methods. Hence, we propose two methods that use the spectral clustering framework and utilize both the graph structure and the node covariates.

Covariate assisted spectral clustering (CASC) uses the leading eigenvectors of

$$\tilde{\mathbf{L}}(h) = \mathbf{L}_\tau + h\mathbf{X}\mathbf{X}^T,$$

where  $h \in [0, \infty)$  is a tuning parameter. When there is little chance for confusion,  $\tilde{\mathbf{L}}$  will be used for notational convenience. When using  $\{0, 1\}$ -Bernoulli covariates, the covariate term can be interpreted as adding to each element  $(i, j)$  a value proportional to the number of covariates equal to one for both  $i$  and  $j$ . In practice, the covariate matrix  $\mathbf{X}$  should be parameterized as in linear regression; specifically, categorical covariates should be re-expressed with dummy variables. For continuous covariates, it can be beneficial to center and scale the columns of  $\mathbf{X}$  before performing the analysis.

To run CASC on the large brain graphs in Section 5, the top  $K$  eigenvectors of  $\tilde{\mathbf{L}}$  are computed using the implicitly restarted Lanczos bidiagonalization algorithm (Baglama and Reichel 2006). At each iteration, it only needs to compute the product  $\tilde{\mathbf{L}}\mathbf{v}$ , where  $\mathbf{v}$  is an arbitrary vector. For computational efficiency, the product is calculated as  $\mathbf{L}_\tau\mathbf{v} + h\mathbf{X}(\mathbf{X}^T\mathbf{v})$ . This takes advantage of the sparsity of  $\mathbf{L}_\tau$  and the low rank structure of  $\mathbf{X}\mathbf{X}^T$ . Ignoring log terms and any special structure in  $\mathbf{X}$ , it takes  $O((|E| + NR)K)$  operations to compute the required top  $K$  eigenvectors of  $\tilde{\mathbf{L}}$ , where  $R$  is the number of columns in  $\mathbf{X}$ . The graph clusters are obtained by iteratively employing the spectral clustering algorithm on  $\tilde{\mathbf{L}}(h)$  while varying the tuning parameter  $h$  until an optimal value is obtained. The details of this procedure are described in the next section.

As an alternative, we propose a modification of classical canonical correlation analysis (Hotelling 1936) whose similarity matrix is the product of the regularized graph Laplacian and the covariate matrix,

$$\mathbf{L}^{CCA} = \mathbf{L}_\tau \mathbf{X}.$$

The spectral clustering algorithm is employed on  $\mathbf{L}^{CCA}$  to obtain node clusters when  $R \geq K$ . This approach inherently provides a dimensionality reduction in the common case where  $R \ll N$ . If  $R \ll N^{-1} \sum_i D_{ii}$ , then spectral clustering with  $\mathbf{L}^{CCA}$  has a faster running time than CASC.



## 2.4 Setting the Tuning Parameter

In order to perform spectral clustering with  $\tilde{\mathbf{L}}(h)$ , it is necessary to determine a specific value for the tuning parameter,  $h$ . The tuning procedure presented here presumes that both the graph and the covariates contain some block information, as demonstrated by the simulations in Section 4. In practice, an initial test can be used to determine if the graph and the covariates contain common block information. The tuning parameter should be chosen to achieve a balance between  $\mathbf{L}_\tau$  and  $\mathbf{X}$  such that the information in both is captured in the leading eigenspace of  $\tilde{\mathbf{L}}$ . For large values of  $h$ , the leading eigenspace of  $\tilde{\mathbf{L}}$  is approximately the leading eigenspace of  $\mathbf{X}\mathbf{X}^T$ . For small values of  $h$ , it is approximately the leading eigenspace of  $\mathbf{L}_\tau$ .

Interestingly, there exists a finite range of  $h$  where the leading eigenspace of  $\tilde{\mathbf{L}}(h)$  is not a continuous function of  $h$ ; outside of this range, the leading eigenspace is always continuous in  $h$ . In simulations, the clustering results are exceedingly stable in the continuous range of  $h$ . Hence, only the values of  $h$  inside a finite interval need to be considered. This section gives an interval  $h \in [h_{min}, h_{max}]$  that is computed with only the eigenvalues of  $\mathbf{L}_\tau$  and  $\mathbf{X}\mathbf{X}^T$ . Within this interval,  $h$  is chosen to minimize an objective function.

To find the initial range  $[h_{min}, h_{max}]$ , define a static vector  $\mathbf{v} \in \mathbb{R}^N$  as a vector that satisfies one of the following properties. For  $\epsilon \geq 0$ ,

- (a)  $\mathbf{v}^T \mathbf{L}_\tau \mathbf{v} \geq \lambda_K(\mathbf{L}_\tau)$  and  $\mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} \leq \epsilon$
- (b)  $\mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} \geq \lambda_K(\mathbf{X}\mathbf{X}^T)$  and  $\mathbf{v}^T \mathbf{L}_\tau \mathbf{v} \leq \epsilon$ .

These are vectors for which  $\mathbf{X}\mathbf{X}^T$  and  $\mathbf{L}_\tau$  are highly differentiated; perhaps there is a cluster in the graph that does not appear in the covariates, or vice versa. These static vectors produce discontinuities in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$ .

For example, let  $\mathbf{v}_*$  be an eigenvector of  $\mathbf{L}_\tau$  and a static vector of type (a), then as  $h$  changes, it will remain a slightly perturbed eigenvector of  $\tilde{\mathbf{L}}(h)$ . When  $\mathbf{v}_*^T \tilde{\mathbf{L}}(h_*) \mathbf{v}_*$  is close

to  $\lambda_K(\tilde{\mathbf{L}}(h_*))$ , then in some neighborhood of  $h_*$ , the slightly perturbed version of  $\mathbf{v}_*$  will transition into the leading eigenspace of  $\tilde{\mathbf{L}}$ . This transition corresponds to a discontinuity in the leading eigenspace.

Appendix A.1 uses the concept of static vectors with  $\epsilon = 0$  to find a limited range of  $h$  for possible discontinuities. Let  $\lambda_i(\mathbf{M})$  be the  $i$ th eigenvalue of matrix  $\mathbf{M}$ . The range of  $h$  values for which discontinuities can occur is given by the interval  $[h_{min}, h_{max}]$ , where

$$h_{min} = \frac{\lambda_K(\mathbf{L}_\tau) - \lambda_{K+1}(\mathbf{L}_\tau)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \text{ and}$$

$$h_{max} = \frac{\lambda_1(\mathbf{L}_\tau)}{\lambda_R(\mathbf{X}\mathbf{X}^T)\mathbb{1}(R \leq K) + (\lambda_K(\mathbf{X}\mathbf{X}^T) - \lambda_{K+1}(\mathbf{X}\mathbf{X}^T))\mathbb{1}(R > K)}.$$

The tuning parameter  $h \in [h_{min}, h_{max}]$  is chosen to be the value which minimizes the k-means objective function, the within cluster sum of squares,

$$O(h) = \sum_{i=1}^K \sum_{\mathbf{u}_j \in F_i} \|\mathbf{u}_j(h) - \mathbf{C}_i(h)\|^2,$$

where  $\mathbf{u}_j$  is the  $j$ th row of  $\mathbf{U}$ ,  $\mathbf{C}_i$  is the centroid of the  $i$ th cluster from k-means, and  $F_i$  is the set of points in the  $i$ th cluster. Hence, the tuning parameter is given by  $h = \operatorname{argmin}_{h \in [h_{min}, h_{max}]} O(h)$ .

The tuning procedure can be enhanced by identifying the  $h$  values at which there is a discontinuous transition in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$  and more closely examining the subintervals defined by these values. This approach is more complicated and only improves clustering performance in limited regimes. Thus, this paper uses the simpler procedure presented here. The enhanced procedure is provided in Appendix A.2.

### 3 MODEL DESCRIPTION AND THEORETICAL RESULTS

To illustrate what covariate assisted spectral clustering (CASC) estimates, this section proposes a statistical model for a network with node covariates and shows that CASC is a weakly consistent estimator of certain parameters in the proposed model.

### 3.1 Node Covariate Stochastic Blockmodel

To derive statistical guarantees for CASC, we assume a joint mixture model for the the graph and the covariates. Under this model, each node belongs to one of  $K$  blocks and each edge in the graph corresponds to an independent Bernoulli random variable. The probability of an edge between any two nodes depends only on the block membership of those nodes (Holland et al. 1983). In addition, each node is associated with  $R$  independent covariates with bounded support. The expectation of these covariates depends only on the block membership.

**Definition 3.1.** (*Node Covariate Stochastic Blockmodel*) Consider a set of nodes  $\{1, 2, \dots, N\}$ . Let  $\mathbf{Z} \in \{0, 1\}^{N \times K}$  assign the  $N$  nodes to the  $K$  blocks; there is exactly one 1 in each row and  $Z_{ij} = 1$  if node  $i$  belongs to block  $j$ . Let  $\mathbf{B} \in [0, 1]^{K \times K}$  be positive definite, full rank, and symmetric, where  $B_{ij}$  is the probability of an edge between a node in block  $i$  and block  $j$ . Conditional on  $\mathbf{Z}$ , the elements of the adjacency matrix are independent Bernoulli random variables. The population adjacency matrix  $\mathbf{A} = E(\mathbf{A}|\mathbf{Z})$  fully identifies the distribution of  $\mathbf{A}$  and

$$\mathbf{A} = \mathbf{Z}\mathbf{B}\mathbf{Z}^T.$$

Let each element of the covariate matrix be bounded by  $J > 0$ . So,  $\mathbf{X} \in [-J, J]^{N \times R}$ . Let  $\mathbf{M} \in [0, 1]^{K \times R}$  be the covariate expectation matrix, where  $M_{i,j}$  is the expectation of the  $j$ th covariate when it is associated with a node in the  $i$ th block. Conditional on  $\mathbf{Z}$ , the elements of  $\mathbf{X}$  are independent and the population covariate matrix,  $\mathbf{X} = E(\mathbf{X}|\mathbf{Z})$ , is given by

$$\mathbf{X} = \mathbf{Z}\mathbf{M}. \tag{1}$$

Under the Node Covariate Stochastic Blockmodel (NC-SBM), CASC seeks to estimate the block membership matrix  $\mathbf{Z}$ .

### 3.2 CASC is Statistically Consistent Under the NC-SBM

The proof of consistency for CASC under the NC-SBM requires three results. First, Lemma 3.2 expresses the eigendecomposition of the population version of the covariate assisted Laplacian,

$$\tilde{\mathcal{L}}(h) = (\mathcal{D} + \tau \mathbf{I})^{-1/2} \mathcal{A}(\mathcal{D} + \tau \mathbf{I})^{-1/2} + h \mathbf{X} \mathbf{X}^T,$$

in terms of  $\mathbf{Z}$ . Second, Theorem 3.3 bounds the spectral norm of the difference between  $\tilde{\mathbf{L}}$  and  $\tilde{\mathcal{L}}$ . Then, the Davis-Kahan Theorem (Davis and Kahan 1970) bounds the difference between the sample and population eigenvectors in Frobenius norm. Finally, Theorem 3.6 combines these results to establish a bound on the mis-clustering rate of CASC. The argument of the proof largely follows Qin and Rohe (2013).

**Lemma 3.2.** (*Equivalence of eigenvectors and block membership*) Under the NC-SBM with  $R$  node covariates and  $K$  blocks,  $\tilde{\mathcal{L}}$  has  $K$  positive eigenvalues and the remaining  $N - K$  eigenvalues are zero. Let  $\mathbf{U} \in \mathbb{R}^{N \times K}$  contain the  $K$  largest eigenvectors of  $\tilde{\mathcal{L}}$  as its columns. Then, there exists an orthogonal matrix  $\mathbf{V} \in \mathbb{R}^{K \times K}$ , such that  $\mathbf{U} = \mathbf{Z}(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{V}$ . Furthermore,

$$\mathbf{Z}_i(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{V} = \mathbf{Z}_j(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{V} \iff \mathbf{Z}_i = \mathbf{Z}_j,$$

where  $\mathbf{Z}_i$  is the  $i$ th row of the block membership matrix.

A proof of Lemma 3.2 is contained in Appendix A.3. The lemma implies that the rows of the population eigenvectors are equal if and only if the corresponding nodes belong to the same block. Hence, to derive a bound on the mis-clustering rate, we will need a bound on the difference between the population eigenvectors and the sample eigenvectors. In order to establish this bound, the following theorem bounds the spectral norm of the difference between  $\tilde{\mathbf{L}}$  and  $\tilde{\mathcal{L}}$ .

**Theorem 3.3.** (*Concentration inequality*) Let  $d = \min \mathcal{D}_{ii}$ ,  $\mathcal{X}_{ik}^{(p)} = E(X_{ik}^p)$ ,

$$\delta = \frac{1}{d + \tau} + 8h^2 \sum_k \left( \sum_i \mathcal{X}_{ik}^{(2)} \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + \mathcal{X}_{ik}^{(4)} \right), \text{ and } S = \max \left( \frac{1}{d + \tau}, 3hNJ^2 \right).$$

For any  $\epsilon > 0$ , if

$$(i) \ d + \tau > 3 \log(4N/\epsilon)$$

$$(ii) \ \delta/S^2 > 3 \log(4N/\epsilon),$$

then with probability at least  $1 - \epsilon$ ,

$$\|\tilde{\mathbf{L}} - \tilde{\mathcal{L}}\| \leq 5\sqrt{3\delta \log(4N/\epsilon)}.$$

A proof of Theorem 3.3 is contained in Appendix A.4. Now we use the result of Theorem 3.3 and the Davis-Kahan Theorem to bound the difference between the sample and population eigenvectors.

**Theorem 3.4.** (*Empirical and population eigenvector bound*) Let  $\lambda_K$  be the  $K$ th largest eigenvalue of  $\tilde{\mathbf{L}}$  and  $\mathbf{O}$  be a rotation matrix. Let the columns of  $\mathbf{U}$  and  $\mathbf{U}$  contain the top  $K$  eigenvectors of  $\tilde{\mathbf{L}}$  and  $\tilde{\mathcal{L}}$ , respectively. Given assumptions (i) and (ii) in Theorem 3.3 and (iii)  $\sqrt{3\delta \log(4N/\epsilon)} \leq \lambda_K/10$ , then with probability at least  $1 - \epsilon$ ,

$$\|\mathbf{U} - \mathbf{U}\mathbf{O}\|_F \leq \frac{40\sqrt{3K\delta \log(4N/\epsilon)}}{\lambda_K}.$$

A proof of Theorem 3.4 is contained in Appendix A.5.

The next theorem bounds the proportion of mis-clustered nodes. In order to define mis-clustering, recall that the spectral clustering algorithm uses k-means to cluster the rows of  $\mathbf{U}$ . Let  $\mathbf{C}_i$  and  $\mathcal{C}_i$  be the cluster centroid of the  $i$ th node generated using k-means on  $\mathbf{U}$  and  $\mathbf{U}$ , respectively. A node  $i$  is correctly clustered if  $\mathbf{C}_i$  is closer to  $\mathcal{C}_i$  than  $\mathcal{C}_j$  for  $\forall j$  such that  $\mathbf{Z}_j \neq \mathbf{Z}_i$ . In order to avoid identifiability problems and since clustering only requires the estimation of the correct subspace, the formal definition is augmented with a rotation matrix  $\mathbf{O}$ . The following definition formalizes this intuition.

**Definition 3.5.** (*Set of mis-clustered nodes*) Let  $\mathbf{O}$  be a rotation matrix that minimizes  $\|\mathbf{U}\mathbf{O}^T - \mathbf{U}\|_F$ . Define the set of mis-clustered nodes as

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

Using the definition of mis-clustering and the result from Theorem 3.4, the next theorem bounds the mis-clustering rate,  $|\mathcal{M}|/N$ .

**Theorem 3.6.** (*Mis-clustering rate bound*) Under assumptions (i) and (ii) in Theorem 3.3 and (iii) in Theorem 3.4, with probability at least  $1 - \epsilon$ ,

$$\frac{|\mathcal{M}|}{N} \leq \frac{c_0 K \delta \log(4N/\epsilon)}{N \lambda_K^2}.$$

A proof of Theorem 3.6 is contained in Appendix A.6.

**Remark 1.** (Choice of  $h$ ) It is instructive to compare the value of  $h$  suggested by the results in Theorem 3.6 with the possible values of  $h$  based on the optimization procedure in Section 2.4. The value of  $h$  suggested by Theorem 3.6 is the value that minimizes the upper bound. Notice that the bound depends on two terms,  $\delta$  and  $\lambda_K$ , and these are constrained by two assumptions in the theorem (ii)  $\delta/S^2 > 3 \log(4N/\epsilon)$  and (iii)  $\sqrt{3\delta \log(4N/\epsilon)} \leq \lambda_K/10$ . A more detailed analysis, contained in Appendix A.7, shows that under some simplifying assumptions (a) the bound has a minimum when  $h = \Theta((N \log N)^{-1})$  and (b) the value for which both conditions are satisfied is  $h = O((N \log N)^{-1})$ . Hence, the key factor in obtaining the best clustering result is ensuring that  $h$  is sufficiently small to keep  $h\mathbf{X}\mathbf{X}^T$  from overwhelming the signal in  $\mathbf{L}_\tau$ . This is sensible since in the sparse setting  $\mathbf{L}_\tau$  has  $O(N \log N)$  non-zero data points, while  $\mathbf{X}$  has  $O(N)$  non-zero data points.

Computing  $h_{min}$  and  $h_{max}$  with  $\tilde{\mathbf{L}}$ , instead of  $\tilde{\mathbf{L}}$ , for convenience, gives  $h_{min} = \Theta(N^{-1})$  and  $h_{max} = \Theta(N^{-1})$ . This suggests that the routine in Section 2.4 will yield  $h = \Theta(N^{-1})$ , which differs from the value suggested by theory. Alternatively, if we allow the number of covariates to grow with the number of nodes such that  $R = \Theta(\log N)$ , then the sparse

graph and the covariates will both have  $\Theta(N \log N)$  non-zero elements. In this case, the empirically determined tuning parameter value and the value suggested by theory are both  $h = \Theta((N \log N)^{-1})$ .

### 3.3 General Lower Bound

The next theorem gives a lower bound for clustering a graph with node covariates. This bound uses Fano's inequality and is similar to that shown in Chaudhuri et al. (2012) for a graph without node attributes. We restrict ourselves to a NC-SBM with  $K = 2$  blocks, but allow for an arbitrary number of covariates  $R$ .

**Theorem 3.7.** (*Covariate assisted clustering lower bound*) *Consider the Stochastic Block-model with  $K = 2$  blocks and  $R$  independent covariates as in Definition 3.1 with  $\mathbf{B}$  such that  $B_{1,1} \geq B_{2,2}$ . Let the KL-divergence of the covariates be  $\Gamma = \sum^R KL(\gamma_i, \gamma'_i)$ , where  $\gamma_i$  is the distribution of the  $i$ th covariate, and  $\Delta = B_{1,1} - B_{1,2}$ . For a fixed  $B_{1,1}$  and  $N \geq 8$ , in order to correctly recover the block assignments with probability at least  $1 - \epsilon$ ,  $\Delta$  must satisfy*

$$\Delta \geq \frac{B_{1,1}(1 - B_{1,1})}{\left(\frac{2}{3N} \left(\frac{\log 2}{2}(1 - \epsilon) - \Gamma - \frac{\log 2}{N}\right)\right)^{-1/2} + (1 - B_{1,1})}.$$

**Remark 2.** (Lower bound interpretation) Suppose

$$B_{1,1} - B_{1,2} < \frac{B_{1,1}(1 - B_{1,1})}{\left(\frac{2}{3N} \left(\frac{\log 2}{2}(1 - \epsilon) - \frac{\log 2}{N}\right)\right)^{-1/2} + (1 - B_{1,1})} \text{ and} \quad (2)$$

$$\Gamma < \frac{\log 2}{2}(1 - \epsilon) - \frac{\log 2}{N}, \quad (3)$$

then only an algorithm that uses both the graph and node covariates can yield correct blocks with high probability. Condition (2) specifies when the graph is insufficient and condition (3) specifies when the covariates are insufficient to individually recover the block membership with high probability.

**Remark 3.** (Bound comparison) The upper bound for CASC in Theorem 3.6 can be compared to the general lower bound. Simplifying the general lower bound gives the condition  $\Delta \geq \Theta(N^{-1/2})$  for perfect clustering with probability  $1 - \epsilon$ . According to Theorem 3.6, CASC achieves perfect clustering with probability  $1 - \epsilon$  when  $\sqrt{c_0 K \delta \log(4N/\epsilon)} < \lambda_K$ . This condition can be rewritten in terms of  $\Delta$  under the assumptions of Theorem 3.7. As discussed in Remark 1, the theorem’s conditions are satisfied when  $h = O((N \log N)^{-1})$ . For  $h = \Theta(N^{-3/2})$ , the smallest bound is obtained, yielding  $\Delta > \Theta(\sqrt{(\log N)/N})$ . In this case, the lower and upper bound differ by a factor of  $\sqrt{\log N}$ . A more detailed analysis is given in Appendix A.9. The covariates do not change the asymptotics of either the lower bound or the upper bound when  $R$  is constant. If  $R = \Theta(\log N)$ , then both bounds can reach zero depending on certain constants. In other words, the graph becomes unnecessary to achieve a high probability of perfect clustering under these conditions.

## 4 SIMULATIONS

In these simulations, consider a NC-SBM with  $K = 2$  blocks and  $R = 2$  node Bernoulli covariates. Define the block probabilities for the graph and the covariates as

$$\mathbf{B} = \begin{bmatrix} p & q \\ q & p \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} m_1 & m_2 \\ m_2 & m_1 \end{bmatrix} \quad (4)$$

where  $p > q$  and  $m_1 > m_2$ . This implies that the probability of an edge within a block is  $p$ , which is greater than  $q$ , the probability of an edge between two blocks. In the first block, the probability of the first covariate being one is  $m_1$  and the probability of the second covariate being one is  $m_2$ . The opposite is true in block two.

These simulations compare four different methods. The first two are canonical correlation analysis clustering (CCA) and covariate assisted spectral clustering (CASC), both of which utilize the node edges as well as the node covariates to cluster the graph. The other two methods utilize either the node edges or the node covariates. For the node edges, regularized



spectral clustering (RSC) is used; for the node covariates, spectral clustering on the covariate matrix (SC-X) is used.

#### 4.1 Simulations with Varying Graph Signal

The first set of simulations investigates the effect of varying the block signal in the graph on the mis-clustering rate. This is done in two different ways. First, the difference in the within and between block probabilities,  $p - q$ , is varied. Second, the number of nodes in the graph,  $N$ , is varied with  $p - q$  held constant.

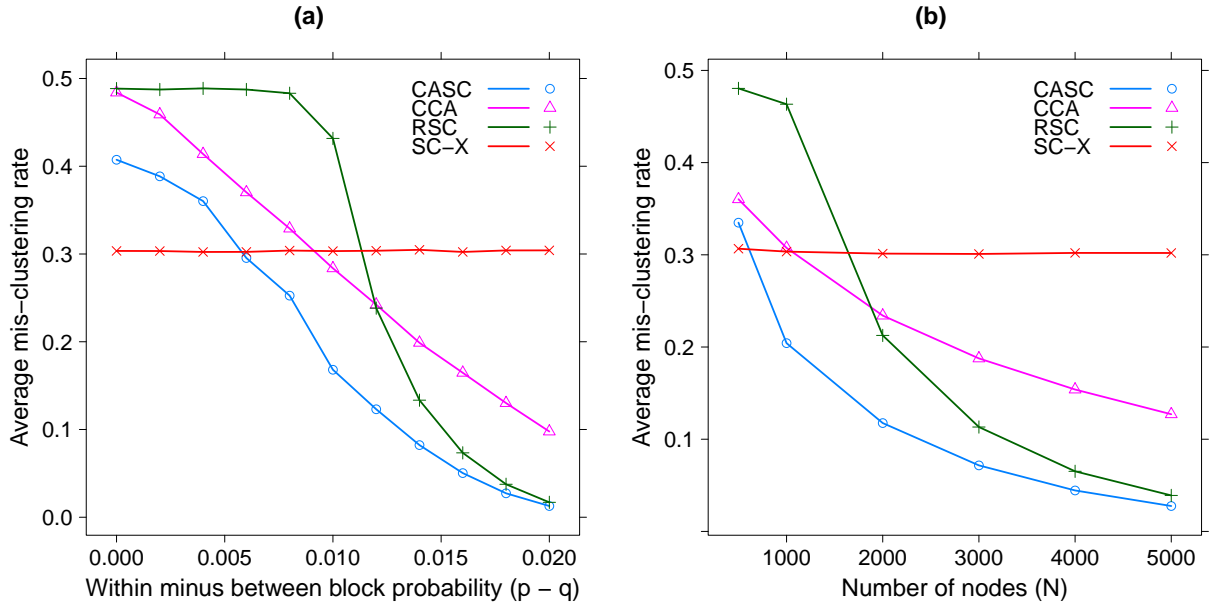


Figure 1. Average mis-clustering rate of four different clustering methods versus (a) the difference in within and between block probabilities,  $p - q$ , with  $N = 1000$  and (b) the number of nodes,  $N$ , with  $p - q = .009$ . The mutual fixed parameters are  $p = .03$ ,  $m_1 = .5$ , and  $m_2 = .1$ .

The simulation results in Figure 1 show that CASC has strictly better performance than canonical correlation analysis clustering in terms of the mis-clustering rate. In fact, CASC performs better than all the other methods. The only exception is spectral clustering on  $\mathbf{X}$

when the difference in the within and between block probabilities of the graph is very small. In this regime,  $\mathbf{L}_\tau$  is mostly a noise term and should be given much less weight than the optimization procedure described in Section 2.4 specifies.

## 4.2 Simulations with Varying Covariate Signal

The second set of simulations investigates the effect of varying the block signal of the covariates on the mis-clustering rate. First, the difference between the covariate probabilities in blocks one and two,  $m_1 - m_2$ , is varied. Second, the number of covariates is varied while maintaining the probability structure in (4). In other words, the new covariate probability matrix  $\mathbf{M}'$  is constructed by appending the columns of  $\mathbf{M}$  to obtain the desired number of covariates.

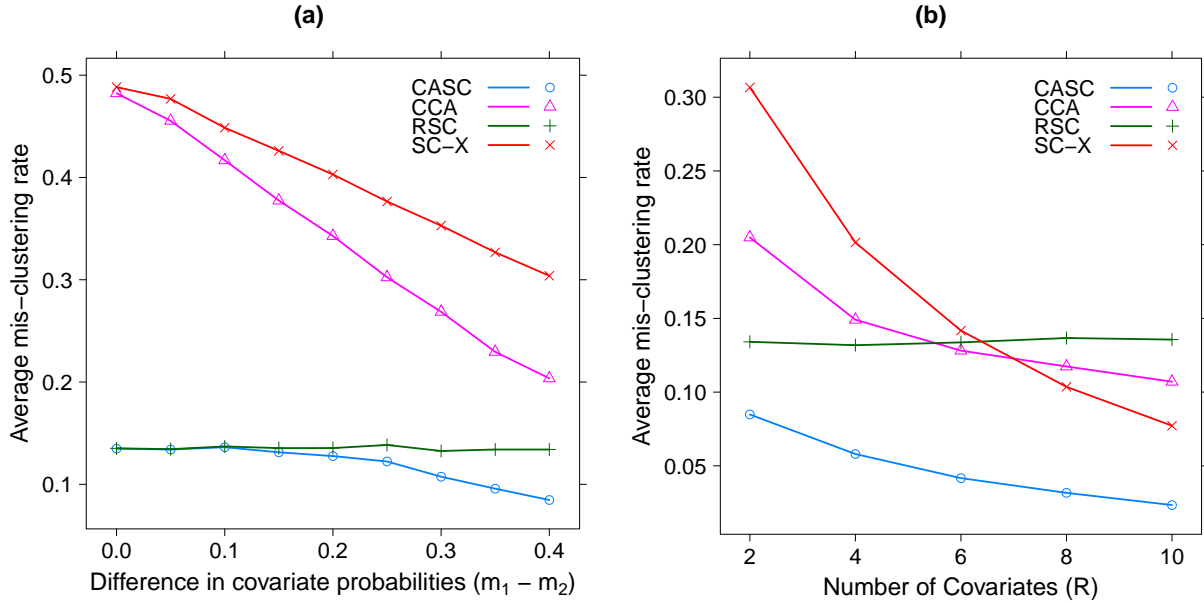


Figure 2. Average mis-clustering rate of four different clustering methods versus (a) the difference in block specific covariate probabilities,  $m_1 - m_2$ , with  $R = 2$  and (b) the number of covariates,  $R$ , with  $m_1 - m_2 = .4$ . The mutual fixed parameters are  $p = .03$ ,  $q = .016$ , and  $m_1 = .5$ .

As shown in Figure 2, CASC tends to have a better mis-clustering rate than the other methods as the covariate probabilities or the number of covariates are altered. Even when the difference in the covariate block probabilities is small and  $\mathbf{X}$  effectively becomes a noise term, the tuning procedure chooses a sufficiently small  $h$  such that CASC has the same performance as regularized spectral clustering.

### 4.3 Simulations Under Model Misspecification

The final set of simulations considers the case where the block membership in the covariates is not necessarily the same as the block membership in the graph. The node Bernoulli covariates no longer satisfy (1) as in Definition 3.1, but

$$\mathbf{X} = \mathbf{Y}\mathbf{M},$$

where  $\mathbf{Y} \in \{0, 1\}^{N \times K}$  is a block membership matrix that differs from  $\mathbf{Z}$ . As such, the underlying clusters in the graph do not align with the clusters in the covariates. This simulation varies the proportion of block assignments in  $\mathbf{Y}$  which agree with the block assignments in  $\mathbf{Z}$  to investigate the robustness of the methods to this form of model misspecification.

The results in Figure 3 show that CASC is robust to covariate block membership model misspecification. The mis-clustering rate shown is computed relative to the block membership of the graph. For this specific case, CASC is able to achieve a lower mis-clustering rate than regularized spectral clustering as long as the proportion of agreement between the block membership of the graph and the covariates is greater than 0.6. Since a two block model is used, the lowest proportion of agreement possible is 0.5 due to identifiability.

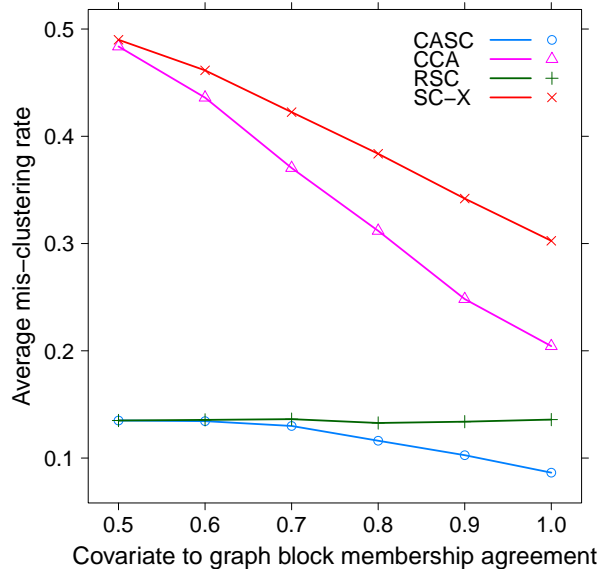


Figure 3. Average mis-clustering rate, relative to the block membership of the graph, of four different clustering methods versus the proportion of block memberships in agreement between  $\mathbf{Z}$ , the graph, and  $\mathbf{Y}$ , the covariates, with model parameters  $p = .03$ ,  $q = .016$ ,  $m_1 = .5$ ,  $m_2 = .1$ , and  $N = 1000$ .

## 5 ILLUSTRATION ON A DIFFUSION MRI NEUROCONNECTOME GRAPH

As an illustrative example, covariate assisted spectral clustering was applied to brain graphs recovered from diffusion MRI (Craddock et al. 2013). Each node in a brain graph corresponds to a voxel in the brain. The edges between nodes are weighted by the number of estimated fibers that pass through both voxels. The center of a voxel is treated as the spatial location of the corresponding node. These spatial locations were centered and used as the first set of covariates in the analysis. The data set used in this analysis contains 42 brain graphs obtained from 21 different individuals. Only the largest connected components of the brain graphs were used, which range in size from 707,000 to 935,000 nodes with a mean density of

744 edges per node. In addition, the brain graphs contain brain atlas labels corresponding to 70 different neurological brain regions, which were treated as a second set of covariates.

Whereas the simulations attempted to demonstrate the effectiveness of covariate assisted spectral clustering in utilizing node covariates to help discover the underlying block structure of the graph, this analysis focuses on the ability of covariate assisted spectral clustering to discover highly connected clusters with relatively homogeneous covariates. The node covariates contextualize the brain clusters and improve their interpretability. Like other clustering methods, covariate assisted spectral clustering is mainly an exploratory tool which may or may not provide answers directly but can often provide insight about relationships within the data. In this example, it is used to examine the relationships between brain graph connectivity, spatial location, and brain atlas labels.

The utility of covariate assisted spectral clustering was explored by partitioning the brain graphs into 100 clusters. Since the brain graphs have heterogeneous node degrees, the rows of the eigenvector matrix were normalized when applying the spectral clustering algorithm to improve the clustering results in this analysis (Qin and Rohe 2013). Figure 4 shows a section of a sample brain graph with nodes plotted at their corresponding spatial locations and colored by cluster membership. For reference, the neurological brain atlas clusters with 70 different regions and an additional category for unlabelled nodes are also plotted. The brain graphs were clustered using three different approaches: regularized spectral clustering (RSC), covariate assisted spectral clustering with spatial location (CASC-X) and with brain atlas membership (CASC-BA).

As shown in Figure 4, using regularized spectral clustering yielded spatially diffuse clusters of densely connected nodes. By adding spatial location using covariate assisted spectral clustering, we obtained densely connected and spatially coherent clusters. Regularized spectral clustering had two clusters of about 80,000 nodes and 4 clusters with less than 1,000 nodes, while the largest cluster from covariate assisted spectral clustering had less than

50,000 nodes and no clusters had less than 1,000 nodes. Both greater spatial coherence and increased uniformity in cluster size demonstrated by covariate assisted spectral clustering are important qualities for interpreting the partition. In addition, the clusters from covariate assisted spectral clustering have a greater similarity with the brain atlas labels than regularized spectral clustering, but this similarity is still not very substantial. This suggests that brain graph connectivity is governed by more than just the neurological regions in the brain atlas.

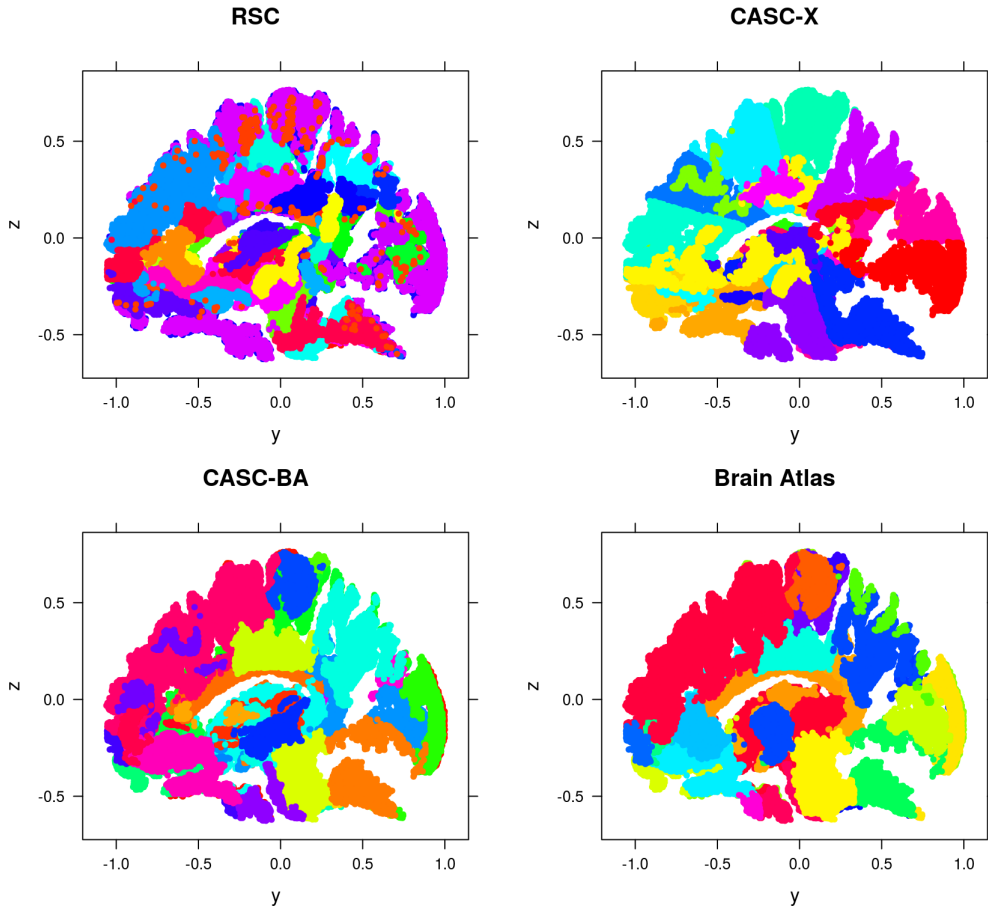


Figure 4. A section of a brain graph with nodes plotted at their spatial location and colored by cluster membership.

The relation between the brain atlas and the brain graph was studied further by treating

the brain atlas membership as the node covariates. This allowed the discovery of highly connected regions with relatively homogeneous graph atlas labels. As shown in Figure 4, relative to the brain atlas, some of the clusters are broken up, a few are joined together, and others overlap with multiple brain atlas regions, but the high similarity is clearly visible. Importantly, this approach gives us clusters that are highly aligned with known neurological regions while allowing for individual variability of the partitions based on brain graph connectivity.

Table 1. The adjusted Rand index between different partitions.

	CASC-X	Brain Atlas	CASC-BA	SC-X
RSC	0.095	0.082	0.085	0.092
CASC-X	-	0.169	0.189	0.278
Brain Atlas	-	-	0.838	0.226
CASC-BA	-	-	-	0.227

The adjusted Rand index (ARI) was used to quantify the similarity of the partitions of a brain graph specified by the different clustering methods and the brain atlas in Table 1. Note that the alignment with the partitions based only on spacial location (SC-X) and either covariate assisted spectral clustering with spatial location or the brain atlas is greater than between the two methods. This indicates that both covariate assisted spectral clustering and the brain atlas are spatially coherent yet not highly overlapping. Brain graph connectivity appears to be making the CASC-X clusters have a different spatial configuration than the brain atlas, which can be observed in Figure 4. As expected, CASC-BA has the highest ARI partition similarity with the brain atlas but low similarity with the regularized spectral clustering partitions. If a more balanced partition alignment is desired, the tuning parameter can be adjusted accordingly.

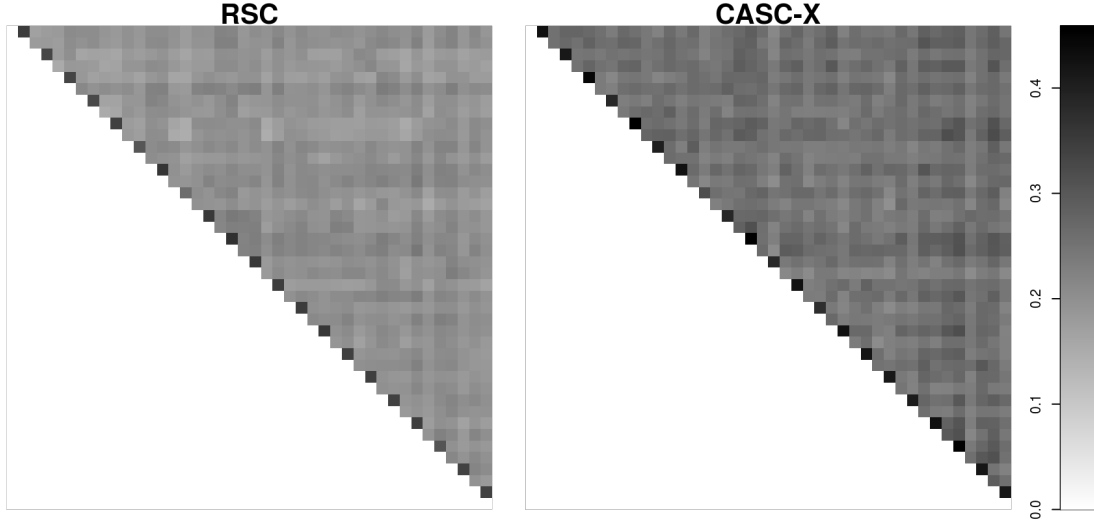


Figure 5. Heat maps comparing the partitions of 42 brain graphs using the adjusted Rand index with adjacent rows corresponding to two scans of the same individual.

The relationship between all 42 brain graphs was analyzed by using ARI to compare partitions between them, as shown in Figure 5. To conduct the comparison, the nodes of each brain graph were matched by spatial location, and any non-matching nodes were ignored. Both regularized spectral clustering and covariate assisted spectral clustering with spatial location were able to clearly distinguish between individuals based on their brain graph partitions, but covariate assisted spectral clustering gave partitions which are more homogeneous both within and between individuals. This increased partition consistency is favorable since a high degree of variation in the clusters between individuals would make them more difficult to interpret.

## 6 DISCUSSION

This paper demonstrates the accuracy, utility, and flexibility of covariate assisted spectral clustering (CASC) as a method for clustering graphs with node covariates. Under the Node Contextualized Stochastic Blockmodel (NC-SBM), incorporating node covariates via CASC



gives better statistical bounds than regularized spectral clustering and lower mis-clustering rates in simulations than other spectral methods. When model assumptions are relaxed, the simulations show that the covariate block structure needs only some overlap with the graph block structure for CASC to improve the clustering results.

Although the NC-SBM is useful for studying graph clustering methods, data often deviates from the model’s assumptions. CASC is not limited to these settings. As the brain graph analysis demonstrates, CASC can be a useful tool for obtaining clusters that satisfy a priori criteria as well. In the brain graph example, this might include spatially coherent clusters or clusters with relatively homogeneous neurological labels. More generally, CASC can be used to find highly connected communities with relatively homogeneous covariates, where the balance between these two objectives is controlled by the tuning parameter and can be set empirically or decided by the analyst. Relatively homogeneous covariates contextualize the clusters making them easier to interpret and allow the analyst to focus on partitions that align with important covariates. Beyond its scientific interest, the brain graph analysis demonstrates the computational efficiency of CASC since the analysis could not have been feasibly conducted with existing methods. Nevertheless, determining an optimal tuning parameter value still presents a computational burden. Using a low rank update algorithm for eigenvector decomposition can potentially further reduce this cost.

This work is meant as a step in the direction of developing a statistical understanding of graphs with node covariates. Further work is needed to better understand the use of CASC for network contextualization. Methods for determining the relative contribution of the graph and the covariates to a graph partition and tests to signify which covariates are informative would be useful tools for such analyses. Ultimately, a thorough examination of the relationship between graph structure and node covariates is essential for a deep understanding of the underlying social or biological systems.

## REFERENCES

- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(1981-2014):3.
- Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.
- Baglama, J. and Reichel, L. (2006). Restarted block lanczos bidiagonalization methods. *Numerical Algorithms*, 43(3):251–272.
- Balasubramanyan, R. and Cohen, W. W. (2011). Block-lda: Jointly modeling entity-annotated text and entity-entity links. In *SDM*, volume 11, pages 450–461. SIAM.
- Chang, J., Blei, D. M., et al. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.
- Chaudhuri, K., Chung, F., and Tsias, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012:1–23.
- Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., and Milham, M. P. (2013). Imaging human connectomes at the macroscale. *Nature Methods*, 10(6):524–539.
- Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.
- Eynard, D., Glashoff, K., Bronstein, M. M., and Bronstein, A. M. (2012). Multimodal diffusion geometry by joint diagonalization of laplacians. *arXiv preprint arXiv:1209.2295*.

- Gibert, J., Valveny, E., and Bunke, H. (2012). Graph embedding in vector spaces by node attribute statistics. *Pattern Recognition*, 45(9):3072–3083.
- Gunnemann, S., Farber, I., Raubach, S., and Seidl, T. (2013). Spectral subspace clustering for graphs with feature vectors. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 231–240. IEEE.
- Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.
- Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3-4):321–377.
- McSherry, F. (2001). Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE.
- Neville, J., Adler, M., and Jensen, D. (2003). Clustering relational data using attribute and link information. In *Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence*, pages 9–15.
- Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering analysis and an algorithm. *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press*, 14:849–856.
- Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455):1077–1087.

- Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128.
- Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915.
- Shiga, M., Takigawa, I., and Mamitsuka, H. (2007). A spectral clustering approach to optimally combining numerical vectors with a modular network. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 647–656. ACM.
- Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.
- Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.
- Wang, F., Ding, C. H., and Li, T. (2009). Integrated kl (k-means-laplacian) clustering: A new clustering approach by combining attribute data and pairwise relations. In *SDM*, pages 38–48. SIAM.
- Yang, J., McAuley, J., and Leskovec, J. (2013). Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1151–1156. IEEE.
- Zhou, Y., Cheng, H., and Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1):718–729.

## APPENDIX: DERIVATIONS AND ALGORITHMIC DETAILS

### A.1 Discontinuous Transitions in the Leading Eigenspace of $\tilde{\mathbf{L}}$

Discontinuous changes in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$  are a major concern when determining an optimal  $h$  value since they have a large effect on the clustering results. They can be studied algebraically by expressing  $\tilde{\mathbf{L}}(h)$  in terms of the eigenvectors of  $\mathbf{L}_\tau$  and  $\mathbf{X}\mathbf{X}^T$ .

Let  $\mathbf{L}_\tau = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$  and  $\mathbf{P}$  be the orthogonal basis of the column space of  $(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{X}\mathbf{X}^T$ , the component of  $\mathbf{X}\mathbf{X}^T$  orthogonal to  $\mathbf{V}$ . Let  $\mathbf{X}\mathbf{X}^T = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$  and  $\tilde{\mathbf{X}}_i = \sqrt{\tilde{\lambda}_i}\tilde{\mathbf{V}}_i$ , so  $\mathbf{X}\mathbf{X}^T = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ . Then,  $\tilde{\mathbf{L}}$  can be written as follows.

$$\begin{aligned}
\tilde{\mathbf{L}} &= \mathbf{L}_\tau + h\mathbf{X}\mathbf{X}^T \\
&= \mathbf{L}_\tau + h\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T \\
&= \begin{bmatrix} \mathbf{V} & \tilde{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & 0 \\ 0 & h\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V} & \tilde{\mathbf{X}} \end{bmatrix}^T \\
&= \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{V}^T\tilde{\mathbf{X}} \\ 0 & \mathbf{P}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\tilde{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & 0 \\ 0 & h\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ \tilde{\mathbf{X}}^T\mathbf{V} & \tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{P} \end{bmatrix} \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix}^T \\
&= \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} + h\mathbf{V}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{V} & h\mathbf{V}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{P} \\ h\mathbf{P}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{V} & h\mathbf{P}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{P} \end{bmatrix} \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix}^T \\
&= \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix} \mathbf{S} \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix}^T \\
&= \left( \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix} \mathbf{V}' \right) \mathbf{\Lambda}' \left( \mathbf{V}'^T \begin{bmatrix} \mathbf{V} & \mathbf{P} \end{bmatrix}^T \right)
\end{aligned}$$

Note that

$$(\mathbf{\Lambda} + h\mathbf{V}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{V})_{ij} = \lambda_i\delta_{ij} + h \sum_k (\mathbf{V}_i^T\tilde{\mathbf{X}}_k)(\tilde{\mathbf{X}}_k^T\mathbf{V}_j)$$

and

$$\mathbf{P}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{V} = (\mathbf{P}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\tilde{\mathbf{X}})[(\tilde{\mathbf{X}}_i^T\mathbf{V}_j)_{ij}].$$

Hence, for any  $j$  such that  $\tilde{\mathbf{X}}_i^T \mathbf{V}_j = 0, \forall i$ , the  $j$ th row and column of  $\mathbf{S}$  will be zero except for the diagonal element. This means that  $\mathbf{U}_j$  will not be rotated by  $\mathbf{V}'$  and will be an eigenvector of  $\tilde{\mathbf{L}}$  for all values of  $h$ . The eigenvalue  $\lambda_j$  will not change either, but its position relative to the other eigenvalues will change with  $h$ . The change in the relative position of  $\lambda_j$  will result in a discontinuous transition in the leading eigenspace of  $\tilde{\mathbf{L}}$  if  $j \geq K$ .

For any  $i$  such that  $\tilde{\mathbf{X}}_i^T \mathbf{V}_j = 0, \forall j$ ,  $\tilde{\mathbf{V}}_i$  is a column in  $\mathbf{P}$  by construction. Row  $i$  in the lower left block of  $\mathbf{S}$  is given by

$$\begin{aligned} \tilde{\mathbf{V}}_i^T (\mathbf{I} - \mathbf{V}\mathbf{V}^T) \tilde{\mathbf{X}} [(\tilde{\mathbf{X}}_i^T \mathbf{V}_j)_{ij}] &= [0, \dots, \sqrt{\tilde{\lambda}_i}, 0, \dots] [(\tilde{\mathbf{X}}_i^T \mathbf{V}_j)_{ij}] \\ &\quad - [0, \dots, 1, 0, \dots] \text{diag}(\sqrt{\tilde{\lambda}_1}, \dots, \sqrt{\tilde{\lambda}_R}) [(\tilde{\mathbf{X}}_i^T \mathbf{V}_j)_{ij}] \\ &= [0, \dots, 0], \end{aligned}$$

and, since  $\mathbf{S}$  is symmetric, this is also column  $i$  in the upper right block of  $\mathbf{S}$ . The lower right block of  $\mathbf{S}$  has row  $i$ , and by symmetry column  $i$ , given by

$$\begin{aligned} \tilde{\mathbf{V}}_i^T (\mathbf{I} - \mathbf{V}\mathbf{V}^T) \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T (\mathbf{I} - \mathbf{V}\mathbf{V}^T) \mathbf{P} &= \tilde{v}_i^T (\tilde{\mathbf{X}} \tilde{\mathbf{X}}^T - \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \mathbf{V}\mathbf{V}^T) \mathbf{P} \\ &= \tilde{\lambda}_i \tilde{\mathbf{V}}_i^T \mathbf{P} \\ &= [0, \dots, \tilde{\lambda}_i, 0, \dots]. \end{aligned}$$

Thus, for any  $i$  such that  $\tilde{\mathbf{X}}_i^T \mathbf{V}_j = 0, \forall j$  the  $i$ th row and column of  $\mathbf{S}$  will be zero except for the diagonal element. This means that  $\tilde{\mathbf{V}}_i$  and  $\tilde{\lambda}_i$  will be an eigenvector and eigenvalue of  $\tilde{\mathbf{L}}$  for all values of  $h$ , but will occupy different relative positions in the eigendecomposition based on the value of  $h$ . The change in the relative position of  $\tilde{\lambda}_i$  will result in a discontinuous transition in the leading eigenspace of  $\tilde{\mathbf{L}}$  if  $i \geq K$ .

Knowing the interval on which such discontinuous transitions are possible can reduce the computational burden of choosing an optimal  $h$ . The values of  $h$  for which transitions occur can be identified as points at which the eigengap equals zero,  $\lambda_K(\tilde{\mathbf{L}}) - \lambda_{K+1}(\tilde{\mathbf{L}}) = 0$ . First, consider the lowest possible value of  $h$  for which such a transition can occur,  $h = \text{argmin}_h \{h :$

$\lambda_K(\tilde{\mathbf{L}}) - \lambda_{K+1}(\tilde{\mathbf{L}}) = 0\}$ . Note that  $\lambda_K(\tilde{\mathbf{L}}) \geq \lambda_K(\mathbf{L}_\tau)$ , where the equality holds when  $\mathbf{V}_K$  is orthogonal to  $\mathbf{X}$  and  $h$  is sufficiently small, and  $\lambda_{K+1}(\tilde{\mathbf{L}}) \leq \lambda_{K+1}(\mathbf{L}_\tau) + h\lambda_1(\mathbf{X}\mathbf{X}^T)$ , where the equality holds when  $\mathbf{V}_{K+1}$  is identical to  $\tilde{\mathbf{V}}_1$ . Hence, the earliest possible transition occurs when

$$\begin{aligned} \lambda_K(\mathbf{L}_\tau) - (\lambda_{K+1}(\mathbf{L}_\tau) + h_{\min}\lambda_1(\mathbf{X}\mathbf{X}^T)) &= 0 \\ h_{\min} &= \frac{\lambda_K(\mathbf{L}_\tau) - \lambda_{K+1}(\mathbf{L}_\tau)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}. \end{aligned}$$

For the highest value of  $h$  for which such a transition is possible, consider  $h^{-1}\tilde{\mathbf{L}}$ . Following the above argument for  $h^{-1}$  with  $\mathbf{X}\mathbf{X}^T$  and  $\mathbf{L}_\tau$  interchanged, a symmetric result is obtained with the additional dependence on the number of covariates,  $R$ . This result yields,

$$h_{\max} = \frac{\lambda_1(\mathbf{L}_\tau)}{\lambda_R(\mathbf{X}\mathbf{X}^T)\mathbb{1}(R \leq K) + (\lambda_K(\mathbf{X}\mathbf{X}^T) - \lambda_{K+1}(\mathbf{X}\mathbf{X}^T))\mathbb{1}(R > K)}.$$

Therefore, discontinuous transitions in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$  can only occur in the interval  $[h_{\min}, h_{\max}]$ .

## A.2 Enhanced Tuning Procedure

The enhanced tuning procedure subdivides the interval  $[h_{\min}, h_{\max}]$  at values where there is a discontinuous transition in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$ . A computationally stable method for identifying the transitions is to check the orthogonality of  $\mathbf{U}_K$ , the  $K$ th eigenvector of  $\tilde{\mathbf{L}}$ , with  $\mathbf{L}_\tau$  and  $\mathbf{X}\mathbf{X}^T$ . This can be done with a grid search on  $t_i \in [h_{\min}, h_{\max}]$ . Let

$$\Phi_1(t_i) = \frac{t_i \mathbf{U}_K^T \mathbf{X} \mathbf{X}^T \mathbf{U}_K}{\lambda_K(\tilde{\mathbf{L}})} \text{ and } \Phi_2(t_i) = \frac{\mathbf{U}_K^T \mathbf{L}_\tau \mathbf{U}_K}{\lambda_K(\tilde{\mathbf{L}})},$$

where  $\lambda_K(\tilde{\mathbf{L}})$  and  $\mathbf{U}_K$  are the  $K$ th eigenvalue and eigenvector of  $\tilde{\mathbf{L}}(t_i)$ . If  $\Phi_1(t_i) < \epsilon$  and  $\Phi_1(t_{i+1}) > \epsilon$ , then a static vector from case (a) is in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$  for  $h \leq t_i$ . If  $\Phi_2(t_i) > \epsilon$  and  $\Phi_2(t_{i+1}) < \epsilon$ , then a static vector from case (b) is in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$  for  $h \geq t_{i+1}$ . Hence, the number of static vectors in the leading eigenspace of  $\tilde{\mathbf{L}}(h)$

is a sum of  $\mathbb{1}(h \leq t_i)$  for each case (a) discontinuity and  $\mathbb{1}(h \geq t_{i+1})$  for each case (b) discontinuity.

The above procedure divides the interval  $[h_{min}, h_{max}]$  into subintervals. The permissible interval for  $h$  is chosen from these subintervals using two criteria. First, if minimum value of the objective function in a given subinterval is greater than the maximum value for any of the other subintervals, then that interval is discarded. Second, from the remaining subintervals, the interval is chosen to minimize the number of static vectors in the leading eigenspace of  $\tilde{\mathbf{L}}$ . This yields the subinterval  $[\tilde{h}_{min}, \tilde{h}_{max}]$ . Eliminating subintervals in the first step preserves static vectors that yield tight clusters. Minimizing the number of static vectors in the second step favors clusters that are present in both the graph and the covariates.

After discovering any discontinuities and adjusting the interval of interest, the tuning parameter  $h \in [\tilde{h}_{min}, \tilde{h}_{max}]$  is chosen to be the value which minimizes the k-means objective function, the within cluster sum of squares,

$$O(h) = \sum_{i=1}^K \sum_{\mathbf{u}_j \in F_i} \|\mathbf{u}_j(h) - \mathbf{C}_i(h)\|^2$$

where  $\mathbf{u}_j$  is the  $j$ th row of  $\mathbf{U}$ ,  $\mathbf{C}_i$  is the centroid of the  $i$ th cluster from k-means, and  $F_i$  is the set of points in the  $i$ th cluster.

### **Tuning Procedure**

1. Compute the top  $K + 1$  eigenvalues of  $\mathbf{L}_\tau$  and  $\mathbf{X}\mathbf{X}^T$  to calculate  $h_{min}$  and  $h_{max}$ .
2. Let  $t_i$  take values on a grid in  $[h_{min}, h_{max}]$ . Compute  $O(t_i)$ ,  $\Phi_1(t_i)$ , and  $\Phi_2(t_i)$ .
3. For each  $t_i$  check for a transition of a static vector into or out of the leading eigenspace of  $\tilde{\mathbf{L}}(t_i)$ . Let  $t_j^*$  be the value of  $t_i$  before the  $j$ th transition.
  - (a) If  $\Phi_1(t_i) < \epsilon$  and  $\Phi_1(t_{i+1}) > \epsilon$ , then set  $\Xi_{1i}(h) = \mathbb{1}(h \leq t_i)$  and  $t_j^* = t_i$ .  
Otherwise, set  $\Xi_{1i}(h) = 0$ .



(b) If  $\Phi_2(t_i) > \epsilon$  and  $\Phi_2(t_{i+1}) < \epsilon$ , then set  $\Xi_{2i}(h) = \mathbb{1}(h \geq t_{i+1})$  and  $t_j^* = t_i$ .

Otherwise, set  $\Xi_{1i}(h) = 0$ .

4. Let  $I_1 = [h_{min}, t_1^*]$ ,  $I_i = (t_i^*, t_{i+1}^*]$ , and  $I_{max} = (t_{max}^*, h_{max}]$ . For each  $i$ , if  $\min_{I_i} O(h) < \max_{I_j} O(h), \forall j$ , then add  $I_i$  to the set of subintervals  $\tilde{I}$ .

5. Compute the number of static vectors in the leading eigenspace of  $\tilde{\mathbf{L}}(t_i)$  as  $\Xi(t_i) = \sum_i (\Xi_{1i}(t_i) + \Xi_{2i}(t_i))$ .

6. Let  $[\tilde{h}_{min}, \tilde{h}_{max}] = \operatorname{argmin}_{I \in \tilde{I}} \Xi(h)$ .

7. Choose  $h = \operatorname{argmin}_{h \in [\tilde{h}_{min}, \tilde{h}_{max}]} O(h)$ .

Based on simulation studies, using  $\epsilon = .05$  in the tuning procedure yields good performance.

This enhanced tuning procedure generates noticeable improvement in the performance of CASC when the signal in the graph is low. In particular, if we run the simulations described in Section 4 using the enhanced tuning procedure, there is a visible improvement relative to the results shown in Figure 1. The results are compared in Figure A.1. The other simulation results are virtually indistinguishable.

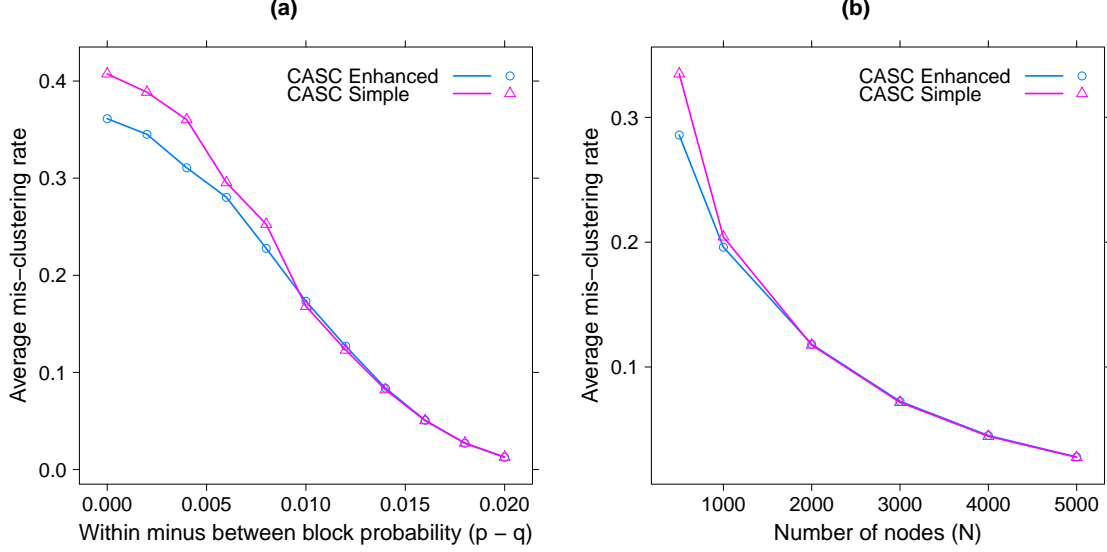


Figure A.1. Average mis-clustering rate of CASC using the simple and enhanced tuning procedure versus (a) the difference in within and between block probabilities,  $p - q$ , with  $N = 1000$  and (b) the number of nodes,  $N$ , with  $p - q = .009$ . The mutual fixed parameters are  $p = .03$ ,  $m_1 = .5$ , and  $m_2 = .1$ .

### A.3 Proof of Lemma 3.2

This proof follows the approach used in Rohe et al. (2011) to establish the equivalence between block membership and population eigenvectors. Note that  $\tilde{\mathcal{L}} = (\mathcal{D} + \tau \mathbf{I})^{-1/2} \mathbf{Z} \mathbf{B} \mathbf{Z}^T (\mathcal{D} + \tau \mathbf{I})^{-1/2} + h \mathbf{Z} \mathbf{M} \mathbf{M}^T \mathbf{Z}^T$ . If we let  $\mathcal{D}_B = \text{diag}(\mathbf{B} \mathbf{Z}^T \mathbf{1}_n + \tau)$ , then  $\tilde{\mathcal{L}} = \mathbf{Z} (\mathcal{D}_B^{-1/2} \mathbf{B} \mathcal{D}_B^{-1/2} + h \mathbf{M} \mathbf{M}^T) \mathbf{Z}^T$ . Recall that  $\mathbf{B}$  is positive definite, symmetric, and full rank by assumption. Let  $\tilde{\mathbf{B}} = \mathcal{D}_B^{-1/2} \mathbf{B} \mathcal{D}_B^{-1/2} + h \mathbf{M} \mathbf{M}^T$ . Assume  $h$  is chosen such that  $\tilde{\mathbf{B}}$  is full rank, which is true  $\forall h$  with the possible exception of a set of values of measure zero. Hence,  $(\mathbf{Z}^T \mathbf{Z})^{1/2} \tilde{\mathbf{B}} (\mathbf{Z}^T \mathbf{Z})^{1/2}$  is symmetric and has real eigenvalues. Note that

$$\det((\mathbf{Z}^T \mathbf{Z})^{1/2} \tilde{\mathbf{B}} (\mathbf{Z}^T \mathbf{Z})^{1/2}) = \det(\mathbf{Z}^T \mathbf{Z})^{1/2} \det(\tilde{\mathbf{B}}) \det(\mathbf{Z}^T \mathbf{Z})^{1/2} > 0,$$

so  $(\mathbf{Z}^T \mathbf{Z})^{1/2} \tilde{\mathbf{B}} (\mathbf{Z}^T \mathbf{Z})^{1/2}$  only has nonzero eigenvalues. By spectral decomposition, let

$$(\mathbf{Z}^T \mathbf{Z})^{1/2} \tilde{\mathbf{B}} (\mathbf{Z}^T \mathbf{Z})^{1/2} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T.$$

Let  $\boldsymbol{\mu} = (\mathbf{Z}^T \mathbf{Z})^{-1/2} \mathbf{V}$ , then

$$\mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1/2} (\mathbf{Z}^T \mathbf{Z})^{1/2} \tilde{\mathbf{B}} (\mathbf{Z}^T \mathbf{Z})^{1/2} (\mathbf{Z}^T \mathbf{Z})^{-1/2} \mathbf{Z} = \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1/2} \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T (\mathbf{Z}^T \mathbf{Z})^{-1/2} \mathbf{Z}$$

$$\mathbf{Z} \tilde{\mathbf{B}} \mathbf{Z}^T = \mathbf{Z} \boldsymbol{\mu} \mathbf{\Lambda} (\mathbf{Z} \boldsymbol{\mu})^T$$

$$\mathbf{Z} \tilde{\mathbf{B}} \mathbf{Z}^T \mathbf{Z} \boldsymbol{\mu} = \mathbf{Z} \boldsymbol{\mu} \mathbf{\Lambda} (\mathbf{Z} \boldsymbol{\mu})^T \mathbf{Z} \boldsymbol{\mu}$$

$$\mathbf{Z} \tilde{\mathbf{B}} \mathbf{Z}^T \mathbf{Z} \boldsymbol{\mu} = \mathbf{Z} \boldsymbol{\mu} \mathbf{\Lambda}.$$

Therefore,  $\mathbf{Z} \boldsymbol{\mu}$  is the matrix of eigenvectors of  $\mathbf{Z} \tilde{\mathbf{B}} \mathbf{Z}^T$ . Also,  $\det(\boldsymbol{\mu}) = \det((\mathbf{Z}^T \mathbf{Z})^{-1/2}) \det(\mathbf{V}) > 0$  so  $\boldsymbol{\mu}^{-1}$  exists and  $\mathbf{Z}_i \boldsymbol{\mu} = \mathbf{Z}_j \boldsymbol{\mu} \iff \mathbf{Z}_i = \mathbf{Z}_j$ .

## A.4 Proof of Theorem 3.3

The spectral norm of the difference between the sample and population covariate assisted Laplacians is bounded by first applying the triangle inequality and bounding the resulting three terms individually.

$$\begin{aligned} \|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}\| &\leq \|(\mathcal{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathcal{D} + \tau \mathbf{I})^{-1/2} + h \mathbf{X} \mathbf{X}^T - E((\mathcal{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathcal{D} + \tau \mathbf{I})^{-1/2} + h \mathbf{X} \mathbf{X}^T)\| \\ &\quad + \|(\mathbf{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathbf{D} + \tau \mathbf{I})^{-1/2} - (\mathcal{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathcal{D} + \tau \mathbf{I})^{-1/2}\| \\ &\quad + \|h(E(\mathbf{X} \mathbf{X}^T) - \boldsymbol{\chi} \boldsymbol{\chi}^T)\| \end{aligned} \tag{5}$$

The second term can be bounded following the proof in the Supplement of Qin and Rohe (2013). Under the assumption that (i)  $d + \tau > 3 \log(4N/\epsilon)$ , where  $d = \min \mathcal{D}_{ii}$ , let  $a = \sqrt{(3 \log(4N/\epsilon))/(d + \tau)}$ , so  $a < 1$ . Then, with probability at least  $1 - \epsilon/2$ ,

$$\|(\mathbf{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathbf{D} + \tau \mathbf{I})^{-1/2} - (\mathcal{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathcal{D} + \tau \mathbf{I})^{-1/2}\| \leq a^2 + 2a.$$

For the first term, use the matrix Bernstein inequality (Tropp 2012). Consider  $\mathbf{T} = (\mathbf{D} + \tau \mathbf{I})^{-1/2} \mathbf{A} (\mathbf{D} + \tau \mathbf{I})^{-1/2} + h \mathbf{X} \mathbf{X}^T$ . This can be expressed as a sum  $\mathbf{T} = \sum_l \mathbf{T}_l$  where

$$\mathbf{T}_l = \begin{cases} (\mathcal{D}_{ii} + \tau)^{-1/2} a_{ij} \mathbf{A}^{ij} (\mathcal{D}_{jj} + \tau)^{-1/2} & \text{for } l = 1, \dots, N^2 \\ h \mathbf{X}_k \mathbf{X}_k^T & \text{for } l = N^2 + 1, \dots, N^2 + R \end{cases},$$

$\mathbf{X}_k$  is the  $k$ th column of  $\mathbf{X}$ , and  $\mathbf{A}^{ij} \in \{0, 1\}^{N \times N}$  has 1 in the  $(i, j)$ th entry and 0 otherwise. Now bound the spectral norm of  $\mathbf{T}_l - \mathcal{T}_l$ , where  $\mathcal{T}_l \equiv E(\mathbf{T}_l)$ , by bounding the two possible expressions for  $\mathbf{T}_l$ .

$$\begin{aligned} \| h \mathbf{X}_k \mathbf{X}_k^T - E(h \mathbf{X}_k \mathbf{X}_k^T) \| &= h \| \mathbf{X}_k \mathbf{X}_k^T - \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \text{diag}(\boldsymbol{\mathcal{X}}_k^{(2)} - \boldsymbol{\mathcal{X}}_k^2) \| \\ &\leq h(\| \mathbf{X}_k \mathbf{X}_k^T \| + \| \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T \| + \max |\boldsymbol{\mathcal{X}}_k^{(2)} - \boldsymbol{\mathcal{X}}_k^2|) \\ &\leq h(NJ^2 + NJ^2 + J^2) \\ &\leq 3hNJ^2 \end{aligned}$$

$$\begin{aligned} \| (\mathcal{D}_{ii} + \tau)^{-1/2} a_{ij} \mathbf{A}^{ij} (\mathcal{D}_{jj} + \tau)^{-1/2} - E((\mathcal{D}_{ii} + \tau)^{-1/2} a_{ij} \mathbf{A}^{ij} (\mathcal{D}_{jj} + \tau)^{-1/2}) \| \\ \leq \| (\mathcal{D}_{ii} + \tau)^{-1/2} (a_{ij} - p_{ij}) (\mathcal{D}_{jj} + \tau)^{-1/2} \mathbf{A}^{ij} \| \\ \leq (\mathcal{D}_{ii} + \tau)^{-1/2} (\mathcal{D}_{jj} + \tau)^{-1/2} \\ \leq \frac{1}{d + \tau} \end{aligned}$$

Hence, this gives the following bound on the spectral norm.

$$\| \mathbf{T}_l - \mathcal{T}_l \| \leq \max \left( \frac{1}{d + \tau}, 3hNJ^2 \right) \equiv S$$

Next, find a bound on the spectral norm of the variance of  $\mathbf{T}$ . Again, first find a bound on the two possible expressions for  $\mathbf{T}_l$ . Let  $\boldsymbol{\mathcal{X}}_k^{(i)}$  be the  $i$ th moment of  $\mathbf{X}_k$ . Start by bounding

the spectral norm of the variance of  $\mathbf{T}_l$  for  $l = N^2 + 1, \dots, N^2 + R$ .

$$\begin{aligned}
E(\mathbf{X}_k \mathbf{X}_k^T) &= \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \text{diag}(\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)}) \\
E(\mathbf{X}_k \mathbf{X}_k^T) E(\mathbf{X}_k \mathbf{X}_k^T) &= (\boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \text{diag}(\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)})) (\boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \text{diag}(\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)})) \\
&= \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T \text{diag}(\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)}) \\
&\quad - \text{diag}(\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)}) \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T + \text{diag}((\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)})^2) \\
&= \left( \sum_i \mathcal{X}_{ik}^2 \right) \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)}))^T \\
&\quad - (\boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)})) \boldsymbol{\mathcal{X}}_k^T + \text{diag}((\boldsymbol{\mathcal{X}}_k^2 - \boldsymbol{\mathcal{X}}_k^{(2)})^2) \\
\\
E(\mathbf{X}_k \mathbf{X}_k^T \mathbf{X}_k \mathbf{X}_k^T) &= E\left( \left( \sum_i \mathcal{X}_{ik}^2 \right) \mathbf{X}_k \mathbf{X}_k^T \right) \\
&= \begin{cases} \mathcal{X}_{ik} \mathcal{X}_{jk} \sum_{l \neq i, j} \mathcal{X}_{lk}^{(2)} + \mathcal{X}_{ik} \mathcal{X}_{jk}^{(3)} + \mathcal{X}_{jk} \mathcal{X}_{ik}^{(3)} & i \neq j \\ \mathcal{X}_{ik}^{(2)} \sum_{l \neq i} \mathcal{X}_{lk}^{(2)} + \mathcal{X}_{ik}^{(4)} & i = j \end{cases} \\
&= \left( \sum_i \mathcal{X}_{ik}^{(2)} \right) \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T - \boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^{(2)})^T - (\boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^{(2)}) \boldsymbol{\mathcal{X}}_k^T \\
&\quad + \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^{(3)T} + \boldsymbol{\mathcal{X}}_k^{(3)} \boldsymbol{\mathcal{X}}_k^T \\
&\quad + \text{diag}\left( (\boldsymbol{\mathcal{X}}_k^{(2)} - \boldsymbol{\mathcal{X}}_k^2) \left( \sum_i \mathcal{X}_{ik}^{(2)} \right) - \boldsymbol{\mathcal{X}}_k^{(2)2} + 2\boldsymbol{\mathcal{X}}_k^2 \boldsymbol{\mathcal{X}}_k^{(2)} - 2\boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^{(3)} + \boldsymbol{\mathcal{X}}_k^{(4)} \right) \\
\\
\text{Var}(\mathbf{X}_k \mathbf{X}_k^T) &= \boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^T \sum_i (\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2) + \boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k^2 - 2\boldsymbol{\mathcal{X}}_k^{(2)}) + \boldsymbol{\mathcal{X}}_k^{(3)})^T + (\boldsymbol{\mathcal{X}}_k (\boldsymbol{\mathcal{X}}_k^2 - 2\boldsymbol{\mathcal{X}}_k^{(2)}) + \boldsymbol{\mathcal{X}}_k^{(3)}) \boldsymbol{\mathcal{X}}_k^T \\
&\quad + \text{diag}\left( (\boldsymbol{\mathcal{X}}_k^{(2)} - \boldsymbol{\mathcal{X}}_k^2) \left( \sum_i \mathcal{X}_{ik}^{(2)} \right) - \boldsymbol{\mathcal{X}}_k^{(2)2} + 2\boldsymbol{\mathcal{X}}_k^2 \boldsymbol{\mathcal{X}}_k^{(2)} - 2\boldsymbol{\mathcal{X}}_k \boldsymbol{\mathcal{X}}_k^{(3)} + \boldsymbol{\mathcal{X}}_k^{(4)} - (\boldsymbol{\mathcal{X}}_k^{(2)} - \boldsymbol{\mathcal{X}}_k^2)^2 \right)
\end{aligned}$$

$$\begin{aligned}
\left\| \sum_k \text{Var}(\mathbf{X}_k \mathbf{X}_k^T) \right\| &\leq \sum_k \sum_i \left| \mathcal{X}_{ik}^2 \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + 2|\mathcal{X}_{ik}^4 - 2\mathcal{X}_{ik}^2 \mathcal{X}_{ik}^{(2)} + \mathcal{X}_{ik} \mathcal{X}_{ik}^{(3)}| \right| \\
&\quad + \max_i \left| (\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2) \left( \sum_l \mathcal{X}_{lk}^{(2)} \right) - \mathcal{X}_{ik}^{(2)^2} + 2\mathcal{X}_{ik}^2 \mathcal{X}_{ik}^{(2)} - 2\mathcal{X}_{ik} \mathcal{X}_{ik}^{(3)} + \mathcal{X}_{ik}^{(4)} - (\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2)^2 \right| \\
&\leq \sum_k \sum_i \mathcal{X}_{ik}^2 \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + 2(\mathcal{X}_{ik}^2 \mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^4) + 2|\mathcal{X}_{ik}^2 \mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik} \mathcal{X}_{ik}^{(3)}| \\
&\quad + \max_i \left( (\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2) \left( \sum_l \mathcal{X}_{lk}^{(2)} \right) + |2\mathcal{X}_{ik} \mathcal{X}_{ik}^{(3)} - \mathcal{X}_{ik}^{(4)} - \mathcal{X}_{ik}^4| + 2(\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2)^2 \right) \\
&\leq \sum_k \sum_i 3\mathcal{X}_{ik}^2 \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + 2|\mathcal{X}_{ik}^2 \mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik} \mathcal{X}_{ik}^{(3)}| \\
&\quad + \max_i \left( 3(\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2) \left( \sum_l \mathcal{X}_{lk}^{(2)} \right) + |2\mathcal{X}_{ik} \mathcal{X}_{ik}^{(3)} - \mathcal{X}_{ik}^{(4)} - \mathcal{X}_{ik}^4| \right) \\
&\leq 8 \sum_k \left( \sum_i \mathcal{X}_{ik}^{(2)} \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + \mathcal{X}_{ik}^{(4)} \right)
\end{aligned}$$

Finally, bound the spectral norm of the variance of  $\mathbf{T}_l$  for  $l = 1, \dots, N^2$ . Let  $\mathbf{L}^{ij} \in \mathbb{R}^{N \times N}$  have  $L_{ij}$  in the  $(i, j)$ th entry and 0 otherwise.

$$\begin{aligned}
E(\mathbf{L}^{ij})E(\mathbf{L}^{ij^T}) &= E((\mathcal{D}_{ii} + \tau)^{-1/2} a_{ij} \mathbf{A}^{ij} (\mathcal{D}_{jj} + \tau)^{-1/2}) E((\mathcal{D}_{ii} + \tau)^{-1/2} a_{ij} \mathbf{A}^{ij} (\mathcal{D}_{jj} + \tau)^{-1/2}) \\
&= (\mathcal{D}_{ii} + \tau)^{-1} E(a_{ij})^2 (\mathcal{D}_{jj} + \tau)^{-1} \mathbf{A}^{ij} \\
E(\mathbf{L}^{ij} \mathbf{L}^{ij^T}) &= E((\mathcal{D}_{ii} + \tau)^{-1} a_{ij} \mathbf{A}^{ij} a_{ij} \mathbf{A}^{ij} (\mathcal{D}_{jj} + \tau)^{-1}) \\
&= (\mathcal{D}_{ii} + \tau)^{-1} E(a_{ij}) (\mathcal{D}_{jj} + \tau)^{-1} \mathbf{A}^{ii} \\
\text{Var}(\mathbf{L}^{ij}) &= (\mathcal{D}_{ii} + \tau)^{-1} (\mathcal{D}_{jj} + \tau)^{-1} (E(a_{ij}) - E(a_{ij})^2) \mathbf{A}^{ii} \\
\|\text{Var}(\mathbf{L}^{ij})\| &= \left\| \sum_i \sum_j (\mathcal{D}_{ii} + \tau)^{-1} (\mathcal{D}_{jj} + \tau)^{-1} (E(a_{ij}) - E(a_{ij})^2) \mathbf{A}^{ii} \right\| \\
&= \max_i \sum_j (\mathcal{D}_{ii} + \tau)^{-1} (\mathcal{D}_{jj} + \tau)^{-1} (E(a_{ij}) - E(a_{ij})^2) \\
&\leq \max_i \frac{1}{d + \tau} \sum_j (\mathcal{D}_{ii} + \tau)^{-1} E(a_{ij}) \\
&= \frac{1}{d + \tau}
\end{aligned}$$

Combining the two bounds on the spectral norms of the variance terms gives the following:

$$\left\| \sum_l \text{var}(\mathbf{T}_l) \right\| \leq \frac{1}{d + \tau} + 8h^2 \sum_k \left( \sum_i \mathcal{X}_{ik}^{(2)} \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + \mathcal{X}_{ik}^{(4)} \right) \equiv \delta.$$

Let  $b = \sqrt{3\delta \log(4N/\epsilon)}$  and assume (ii)  $\delta/S^2 > 3 \log(4N/\epsilon)$ , then  $b < \delta/S$ . Applying the matrix Bernstein inequality gives,

$$\begin{aligned} P(\|(\mathbf{D} + \tau \mathbf{I})^{-1/2} A(\mathbf{D} + \tau \mathbf{I})^{-1/2} + h \mathbf{X} \mathbf{X}^T - \tilde{\mathbf{L}}\| > b) &\leq 2N \exp\left(-\frac{b^2}{2v^2 + 2Sb/3}\right) \\ &\leq 2N \exp\left(-\frac{3S\delta \log(4N/\epsilon)}{2\delta + 2Sb/3}\right) \\ &\leq 2N \exp\left(-\frac{3\delta \log(4N/\epsilon)}{3\delta}\right) \\ &= \epsilon/2. \end{aligned}$$

Expanding the expression for assumption (ii)  $\delta/S^2 > 3 \log(4N/\epsilon)$  gives

$$\begin{aligned} \min \left( d + \tau + 8\left(\frac{h}{d + \tau}\right)^2 \sum_k \left( \sum_i \mathcal{X}_{ik}^{(2)} \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + \mathcal{X}_{ik}^{(4)} \right), \right. \\ \left. \frac{1}{(d + \tau)h^2 N^2} + \frac{8}{N^2} \sum_k \left( \sum_i \mathcal{X}_{ik}^{(2)} \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + \mathcal{X}_{ik}^{(4)} \right) \right) > 3 \log(4N/\epsilon), \end{aligned}$$

which is less restrictive than assumption (i) when  $h$  is small, but more restrictive for larger  $h$ . They are equal when  $h = 0$ .

The third term can be bounded as

$$\begin{aligned} \|h(E(\mathbf{X} \mathbf{X}^T) - \mathbf{x} \mathbf{x}^T)\| &\leq h \left\| \sum_k \text{diag}(\mathbf{x}_k^{(2)} - \mathbf{x}_k^2) \right\| \\ &\leq h \sum_k \max_i (\mathcal{X}_{ik}^{(2)} - \mathcal{X}_{ik}^2) \\ &\leq h \sum_k \max_i \mathcal{X}_{ik}^{(2)} \\ &\leq h \sum_k \max_i \sqrt{\mathcal{X}_{ik}^{(4)}} \\ &\leq b. \end{aligned}$$

Consequently, joining the results for the three terms in (5), gives the desired bound. With probability at least  $1 - \epsilon$ ,

$$\|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}\| \leq a^2 + 2a + 2b \leq 5b = 5\sqrt{3\delta \log(4N/\epsilon)}.$$

## A.5 Proof of Theorem 3.4

Using Lemma 9 from McSherry (2001), let  $\mathbf{P}_{\tilde{\mathbf{L}}}$  be the projection onto the span of the first  $K$  left singular eigenvectors of  $\tilde{\mathbf{L}}$ . Then,  $\mathbf{P}_{\tilde{\mathbf{L}}}$  is the optimal rank  $K$  approximation to  $\tilde{\mathbf{L}}$  and

$$\|\mathbf{P}_{\tilde{\mathbf{L}}} - \tilde{\mathbf{L}}\|_F^2 \leq 8K \|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}\|^2.$$

Next, apply the Davis-Kahan Theorem to  $\tilde{\mathbf{L}}$  (Davis and Kahan 1970). Let  $W \subset \mathbb{R}$  be an interval and define the distance between  $W$  and the spectrum of  $\tilde{\mathbf{L}}$  outside of  $W$  as

$$\Lambda = \min\{|\lambda - r|; \lambda \text{ eigenvalue of } \tilde{\mathbf{L}}, \lambda \notin W, r \in W\}.$$

Choose  $W = (\lambda_K/2, \infty)$ , where  $\lambda_K$  is the  $K$ th eigenvalue of  $\tilde{\mathbf{L}}$ . Then,  $\Lambda = \lambda_K/2$ . Let  $\omega_K$  be the  $K$ th largest eigenvalue of  $\tilde{\mathbf{L}}$ , then under the assumption that  $\sqrt{3\delta \log(4N/\epsilon)} \leq \lambda_K/10$ ,

$$|\lambda_K - \omega_K| \leq 5\sqrt{3\delta \log(4N/\epsilon)} \leq \lambda_K/2.$$

Hence,  $\omega_K \in W$ , and  $\mathbf{U}$  has the same dimension as  $\mathbf{U}$ . The Davis-Kahan Theorem implies,

$$\begin{aligned} \|\mathbf{U} - \mathbf{U}\mathbf{O}\|_F &\leq \frac{\sqrt{2} \|\mathbf{P}_{\tilde{\mathbf{L}}} \tilde{\mathbf{L}} - \tilde{\mathbf{L}}\|_F}{\Lambda} \\ &\leq \frac{2\sqrt{2} \|\mathbf{P}_{\tilde{\mathbf{L}}} \tilde{\mathbf{L}} - \tilde{\mathbf{L}}\|_F}{\lambda_K} \\ &\leq \frac{8\sqrt{K} \|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}\|}{\lambda_K} \\ &\leq \frac{40\sqrt{3K\delta \log(4N/\epsilon)}}{\lambda_K} \end{aligned}$$

with probability at least  $1 - \epsilon$ .



## A.6 Proof of Theorem 3.6

This proof follows the arguments given in Qin and Rohe (2013). First, note that all population centroids are orthonormal since  $\mathbf{C}_i = \mathbf{Z}_i \boldsymbol{\mu}$ , as shown in the proof of Lemma 3.2. For  $\forall \mathbf{Z}_j \neq \mathbf{Z}_i$ , a sufficient condition for one observed centroid to be closest to the population centroid is

$$\|\mathbf{C}_i \boldsymbol{\mathcal{O}}^T - \mathbf{c}_i\|_2 < \frac{1}{\sqrt{2}} \Rightarrow \|\mathbf{C}_i \boldsymbol{\mathcal{O}}^T - \mathbf{c}_i\|_2 < \|\mathbf{C}_i \boldsymbol{\mathcal{O}}^T - \mathbf{c}_j\|_2.$$

Let  $\mathcal{G} = \{i : \|\mathbf{C}_i \boldsymbol{\mathcal{O}}^T - \mathbf{c}_i\|_2 \geq \frac{1}{\sqrt{2}}\}$ , so  $\mathcal{M} \subset \mathcal{G}$ . Define  $\mathbf{Q} \in \mathbb{R}^{N \times K}$  where the  $i$ th row is  $\mathbf{C}_i$ . By the definition of k-means,  $\|\mathbf{U} - \mathbf{Q}\|_2 \leq \|\mathbf{U} - \mathbf{U}\boldsymbol{\mathcal{O}}\|_2$ . Applying the triangle inequality gives

$$\|\mathbf{Q} - \mathbf{Z}\boldsymbol{\mu}\boldsymbol{\mathcal{O}}\|_2 = \|\mathbf{Q} - \mathbf{U}\boldsymbol{\mathcal{O}}\|_2 \leq \|\mathbf{U} - \mathbf{Q}\|_2 + \|\mathbf{U} - \mathbf{U}\boldsymbol{\mathcal{O}}\|_2 \leq 2\|\mathbf{U} - \mathbf{U}\boldsymbol{\mathcal{O}}\|_2.$$

So,

$$\begin{aligned} \frac{|\mathcal{M}|}{N} &\leq \frac{|\mathcal{G}|}{N} = \frac{1}{N} \sum_{i \in \mathcal{G}} 1 \\ &\leq \frac{2}{N} \sum_{i \in \mathcal{G}} \|\mathbf{C}_i \boldsymbol{\mathcal{O}}^T - \mathbf{c}_i\|_2^2 \\ &= \frac{2}{N} \sum_{i \in \mathcal{G}} \|\mathbf{C}_i - \mathbf{Z}_i \boldsymbol{\mu}\boldsymbol{\mathcal{O}}\|_2^2 \\ &\leq \frac{2}{N} \|\mathbf{Q} - \mathbf{Z}\boldsymbol{\mu}\boldsymbol{\mathcal{O}}\|_F^2 \\ &\leq \frac{8}{N} \|\mathbf{U} - \mathbf{U}\boldsymbol{\mathcal{O}}\|_F^2. \end{aligned}$$

Thus, using the result from Theorem 3.4, with probability at least  $1 - \epsilon$ ,

$$\frac{|\mathcal{M}|}{N} \leq \frac{c_0 K \delta \log(4N/\epsilon)}{N \lambda_k^2},$$

where  $c_0 = 3 \times 5^2 \times 2^9$ .

## A.7 Investigation of the Value of $h$ Suggested by Theorem 3.6

We use two approaches to find a value of  $h$  suggested by theory, and show that both yield the same result. The first finds a value of  $h$  that maximizes the population eigengap under the constraint of the sparsity condition (ii) in Theorem 3.6. The second uses some simplifying assumptions to demonstrate that this same  $h$  also gives the smallest bound in Theorem 3.6, under the theorem assumptions.

To show that the eigengap is non-decreasing in  $h$ , note that both  $\mathcal{L}$  and  $\tilde{\mathcal{L}}$  are rank  $K$  and their eigenvectors span the same subspace  $\mathcal{S}$  based on the results of Lemma 3.2. Thus,  $\forall h, \delta h \geq 0$ ,

$$\begin{aligned} \lambda_K(\tilde{\mathcal{L}}(h + \delta h)) &= \min_{u^T u = 1, u \in \mathcal{S}} u^T (\mathcal{L} + (h + \delta h) \mathbf{x} \mathbf{x}^T) u \\ &\geq \min_{u^T u = 1, u \in \mathcal{S}} u^T (\mathcal{L} + h \mathbf{x} \mathbf{x}^T) u + \delta h \min_{u^T u = 1, u \in \mathcal{S}} u^T \mathbf{x} \mathbf{x}^T u \\ &\geq \begin{cases} \lambda_K(\tilde{\mathcal{L}}(h)) + \delta h \lambda_K(\mathbf{x} \mathbf{x}^T), & R \geq K \\ \lambda_K(\tilde{\mathcal{L}}(h)), & \text{otherwise} \end{cases}. \end{aligned}$$

Hence, the eigengap is nondecreasing and  $h$  should have the largest value that satisfies the sparsity condition. The second term in the sparsity condition is the limiting quantity when  $h$  is large. Hence, it is required that

$$\begin{aligned} \frac{\delta}{S^2} &> 3 \log(4N/\epsilon) \\ \frac{1}{(d + \tau)h^2 N^2} + \frac{8}{N^2} \sum_k \left( \sum_i \mathcal{X}_{ik}^{(2)} \sum_l (\mathcal{X}_{lk}^{(2)} - \mathcal{X}_{lk}^2) + \mathcal{X}_{ik}^{(4)} \right) &> 3 \log(4N/\epsilon) \\ \frac{1}{(d + \tau)h^2 N^2} &> 3 \log N + c_1 \\ h &< \frac{1}{N \sqrt{(3 \log N + c_1)(d + \tau)}}. \end{aligned}$$

Given the assumption on node degrees that  $d + \tau > 3 \log(4N/\epsilon)$ , then  $h = O((N \log N)^{-1})$ . Since we want the largest eigengap possible, this suggests that a reasonable value for the tuning parameter is  $h = \Theta((N \log N)^{-1})$ .

In order to investigate the mis-clustering bound and the accompanying conditions, we make some simplifying assumptions. Assume  $B_{i,i} = p, \forall i$  and  $B_{i,j} = q, \forall i \neq j$ ; in addition,  $M_{i,i} = m_1, \forall i$ ;  $M_{i,j} = m_2, \forall i \neq j$ ; and  $R > 1$ . Also, assume that each block has the same number of nodes  $N/K$ . Recall,  $\tilde{\mathcal{L}} = \mathbf{Z}(\mathcal{D}_B^{-1/2} \mathbf{B} \mathcal{D}_B^{-1/2} + h \mathbf{M} \mathbf{M}^T) \mathbf{Z}^T = \mathbf{Z} \tilde{\mathbf{B}} \mathbf{Z}^T$ . Therefore,

$$\begin{aligned} \tilde{\mathbf{B}} &= \frac{1}{N(p + (K-1)q)/K + \tau} ((p-q)\mathbf{I} + q\mathbf{1}_K \mathbf{1}_K^T) + h((m_p - m_q)\mathbf{I} + m_q \mathbf{1}_K \mathbf{1}_K^T) \\ &= \frac{1}{N(p + (K-1)q)/K + \tau} ((p-q) + h(N(p + (K-1)q)/K + \tau)(m_p - m_q))\mathbf{I} \\ &\quad + (q + h(N(p + (K-1)q)/K + \tau)m_q)\mathbf{1}_K \mathbf{1}_K^T, \end{aligned}$$

where  $m_p = m_1^2 + (R-1)m_2^2$  and  $m_q = m_1 m_2 + \mathbb{1}(R > 1)(m_1 m_2 + (R-2)m_2^2)$ . Note that for matrices of the form  $a\mathbf{I} + b\mathbf{1}_K \mathbf{1}_K^T$ ,  $\lambda_K = a$ . Thus,

$$\lambda_K(\tilde{\mathbf{B}}) = \frac{p-q}{N(p + (K-1)q)/K + \tau} + h(m_p - m_q).$$

Recall that  $\tilde{\mathcal{L}}$  has the same eigenvalues as  $(\mathbf{Z}^T \mathbf{Z})^{1/2} \tilde{\mathbf{B}} (\mathbf{Z}^T \mathbf{Z})^{1/2} = (N/K)^{1/2} \mathbf{I} \tilde{\mathbf{B}} (N/K)^{1/2} \mathbf{I} = (N/K) \tilde{\mathbf{B}}$ . Hence, the population eigengap is given by

$$\lambda_K(\tilde{\mathcal{L}}) = \frac{p-q}{p + (K-1)q + K\tau/N} + \frac{hN(m_p - m_q)}{K}.$$

Using these results, we can find the  $h$  that gives the tightest bound on the mis-clustering rate by minimizing

$$\frac{\delta}{\lambda_K(\tilde{\mathcal{L}})^2} = \frac{h^2 \Theta(N^2) + \Theta(1/(\log N))}{h^2 \Theta(N^2) + h \Theta(N) + \Theta(1)}.$$

The minimum occurs when  $h = \Theta((N \log N)^{-1})$ . Finally, we investigate the eigengap condition.

$$\begin{aligned} \sqrt{3\delta \log(4N/\epsilon)} &\leq \lambda_K/10 \\ \sqrt{c_2 + h^2 O(N^2) \log N} &\leq c_3 + h O(N) \\ h^2 O(N^2 \log N) &\leq c_3^2 + h O(N) + h^2 O(N^2) \\ h &= O\left(\frac{1}{N \log N}\right) \end{aligned}$$

Thus, the mis-clustering bound has a minimum when  $h = \Theta((N \log N)^{-1})$ , and both the eigengap and sparsity conditions require  $h = O((N \log N)^{-1})$ . Hence, this more in depth analysis has the same conclusion as the first argument. Both approaches suggest the tuning parameter should be  $h = \Theta((N \log N)^{-1})$ . This does not agree with the value suggested by the empirical selection procedure in Section 2.4, which is  $h = \Theta(N^{-1})$  based on the population eigenvalues.

The above analysis assumed that  $R$  is constant, but it is also interesting to consider  $R = \Theta(\log N)$ . As above, check what values of  $h$  satisfy the sparsity condition.

$$\begin{aligned}\frac{\delta}{S^2} &> 3 \log(4N/\epsilon) \\ \frac{1}{(d + \tau)h^2 N^2} + \Theta(\log N) &> 3 \log N \\ h &< \frac{1}{\Theta(N \log N)}\end{aligned}$$

Given the assumption on node degrees that  $d + \tau > 3 \log(4N/\epsilon)$ , then  $h = O((N \log N)^{-1})$ . Next, check what values of  $h$  satisfy the eigengap condition.

$$\begin{aligned}\sqrt{3\delta \log(4N/\epsilon)} &\leq \lambda_K/10 \\ \sqrt{c_2 + h^2 O((N \log N)^2)} &\leq c_3 + h O(N \log N) \\ h^2 O((N \log N)^2) &\leq c_3^2 + h O(N \log N) + h^2 O(N^2 \log N) \\ h &= O\left(\frac{1}{N \log N}\right)\end{aligned}$$

Finally, find the  $h$  that minimizes the mis-clustering bound,

$$\frac{\delta}{\lambda_K(\tilde{\mathcal{L}})^2} = \frac{h^2 \Theta(N^2 \log N) + \Theta(1/(\log N))}{h^2 \Theta(N^2 \log N) + h \Theta(N \log N) + \Theta(1)}.$$

The minimum occurs when  $h = \Theta((N \log N)^{-1})$ . Thus, the theory suggests that  $h = \Theta((N \log N)^{-1})$  is a good value when  $R = \Theta(\log N)$ , as well. Unlike for the constant  $R$  case, this result agrees with the value suggested by the empirical procedure in Section 2.4, which yields  $h = \Theta((N \log N)^{-1})$  when  $R = \Theta(\log N)$  based on the population eigenvalues.

## A.8 Proof of Theorem 3.7

This proof uses Fano's inequality to derive the lower bound following an approach similar to Chaudhuri et al. (2012). Let  $G_S$  be a partition given by a specific  $S$ , the set of all nodes in the first block, and let  $F$  be the family of all such partitions. Fano's inequality states

$$\sup_{G_S \in F} P_{G_S}(\Psi \neq G_S) \geq 1 - \frac{\beta + \log 2}{\log r},$$

where  $KL(G_S, G_{S'}) \leq \beta$ ,  $r = |F| - 1$ , and  $\Psi$  is the true node partition.

First, by independence the KL-divergence can be written as follows,

$$KL(G_S) = \sum_e KL(\rho, \rho') + \sum_e^N KL(\gamma, \gamma').$$

Let  $\rho$  and  $\rho'$  be the edge distribution and  $\gamma$  and  $\gamma'$  be the node covariate distribution for  $G_S$  and  $G_{S'}$ , respectively. Without loss of generality, assume  $B_{1,1} > B_{2,2} > B_{1,2}$  and let  $b_i \in \{B_{1,1}, B_{2,2}, B_{1,2}\}$ . For a single edge when  $\rho \neq \rho'$ ,

$$\begin{aligned} KL(\rho, \rho') &= \sum_{i>j} b_i \log \frac{b_i}{b_j} + (1 - b_i) \log \frac{1 - b_i}{1 - b_j} + b_j \log \frac{b_j}{b_i} + (1 - b_j) \log \frac{1 - b_j}{1 - b_i} \\ &= \sum_{i>j} (b_i - b_j) \log \left( 1 + \frac{b_i - b_j}{b_j(1 - b_i)} \right) \\ &\leq \sum_{i>j} \frac{(b_i - b_j)^2}{b_j(1 - b_i)} \\ &\leq 3 \frac{(B_{1,1} - B_{1,2})^2}{B_{1,2}(1 - B_{1,1})}. \end{aligned}$$

Now find the KL-divergence of the covariates on a single node. For  $\gamma \neq \gamma'$ ,

$$KL(\gamma, \gamma') = \sum^R KL(\gamma_i, \gamma'_i) \equiv \Gamma.$$

For the case of Bernoulli random variables, this is given by

$$\begin{aligned} KL(\gamma, \gamma') &= \sum^R M_{1,j} \log \frac{M_{1,j}}{M_{2,j}} + (1 - M_{1,j}) \log \frac{1 - M_{1,j}}{1 - M_{2,j}} + M_{2,j} \log \frac{M_{2,j}}{M_{1,j}} + (1 - M_{2,j}) \log \frac{1 - M_{2,j}}{1 - M_{1,j}} \\ &= \sum^R (M_{1,j} - M_{2,j}) \log \frac{M_{1,j}(1 - M_{2,j})}{M_{2,j}(1 - M_{1,j})}. \end{aligned}$$

Therefore, the KL-divergence is bounded by

$$KL(G_S) \leq 3 \binom{N}{2} \frac{(B_{1,1} - B_{1,2})^2}{B_{1,2}(1 - B_{1,1})} + N\Gamma \leq \frac{3N^2}{2} \frac{(B_{1,1} - B_{1,2})^2}{B_{1,2}(1 - B_{1,1})} + N\Gamma.$$

The number of partitions can be bounded as follows,

$$\begin{aligned} |F| &= \frac{1}{2} \binom{N}{N/2} = \frac{N!}{2((N/2)!)^2} \\ &\geq \frac{\sqrt{2\pi N}(N/e)^N}{(e\sqrt{N/2}(N/(2e))^{N/2})^2} \\ &\geq \frac{2^{N-2.1}}{\sqrt{N/2}}, \end{aligned}$$

where the first inequality uses  $\sqrt{2\pi N}(N/e)^N \leq N! \leq e\sqrt{N}(N/e)^N$ . Now the log term is bounded by

$$\begin{aligned} \log(|F| - 1) &\geq \log\left(\frac{2^{N-2.1}}{\sqrt{N/2}} - 1\right) \\ &\geq (N - 3) \log 2 - \frac{1}{2} \log(N/2) \\ &\geq \frac{\log 2}{2} N \text{ for } N \geq 8. \end{aligned}$$

Thus, by Fano's inequality, in order to correctly determine the block assignments with probability at least  $1 - \epsilon$  requires

$$\begin{aligned} \epsilon &\geq 1 - \frac{3N^2(B_{1,1} - B_{1,2})^2/(2B_{1,2}^2(1 - B_{1,1})^2) + N\Gamma + \log 2}{(N \log 2)/2} \\ B_{1,1} - B_{1,2} &\geq B_{1,2}(1 - B_{1,1}) \sqrt{\frac{2}{3N} \left( \frac{\log 2}{2} (1 - \epsilon) - \Gamma - \frac{\log 2}{N} \right)}. \end{aligned}$$

Fix  $B_{1,1}$  and let  $\Delta = B_{1,1} - B_{1,2}$ , then rewrite this bound as

$$\Delta \geq \frac{B_{1,1}(1 - B_{1,1})}{\left( \frac{2}{3N} \left( \frac{\log 2}{2} (1 - \epsilon) - \Gamma - \frac{\log 2}{N} \right) \right)^{-1/2} + (1 - B_{1,1})}.$$

## A.9 Comparison of the General Lower Bound to Theorem 3.6

First, simplify the general lower bound given in Theorem 3.7 to make the comparison with Theorem 3.6 easier.

$$\begin{aligned}
\Delta &\geq \frac{B_{1,1}(1 - B_{1,1})}{\left(\frac{2}{3N} \left(\frac{\log 2}{2}(1 - \epsilon) - \mathcal{K} - \frac{\log 2}{N}\right)\right)^{-1/2} + (1 - B_{1,1})} \\
&\geq \frac{B_{1,1}(1 - B_{1,1})}{3/2 \left(\frac{2}{3N} \left(\frac{\log 2}{2}(1 - \epsilon) - \mathcal{K} - \frac{\log 2}{N}\right)\right)^{-1/2}} \\
&\geq B_{1,1}(1 - B_{1,1}) \left(\frac{2}{3}\right) \left(\frac{2}{3N} \left(\frac{\log 2}{2}(1 - \epsilon) - \mathcal{K} - \frac{\log 2}{8}\right)\right)^{1/2} \\
&\geq \frac{c_4}{\sqrt{N}}
\end{aligned}$$

According to Theorem 3.6 to achieve perfect clustering with probability  $1 - \epsilon$ , requires  $\sqrt{c_0 K \delta \log(4N/\epsilon)} < \lambda_K$ . Assuming  $K = 2$  as in Theorem 3.7 and  $B_{1,1} = B_{2,2}$ ,

$$\begin{aligned}
\sqrt{2c_0 \left(\frac{1}{N\Delta/2 + NB_{1,2} + \tau} + h^2\Theta(N^2)\right) \log(4N/\epsilon)} &< \frac{\Delta}{\Delta + 2B_{1,2} + 2\tau/N} + \frac{hN(m_p - m_q)}{2} \\
\frac{\Theta(\log N)}{N\Delta/2 + NB_{1,2} + \tau} + h^2\Theta(N^2 \log N) &< \frac{\Delta^2}{(\Delta + 2B_{1,2})^2} + \frac{\Delta h\Theta(N)}{\Delta + 2B_{1,2}} + h^2\Theta(N^2) \\
\Theta\left(\frac{\log N}{N}\right) + h^2\Theta(N^2 \log N) &< \frac{\Delta^2}{(\Delta + 2B_{1,2})^2} + \frac{\Delta h\Theta(N)}{\Delta + 2B_{1,2}} + h^2\Theta(N^2).
\end{aligned}$$

Note that, due to the sparsity condition, we require  $h = O((N \log N)^{-1})$ . Accordingly, lower order terms can be dropped. First, consider  $h = \Theta((N \log N)^{-1})$ .

$$\begin{aligned}
\Theta\left(\frac{\log N}{N}\right) + \Theta\left(\frac{1}{\log N}\right) &< \frac{\Delta^2}{(\Delta + 2B_{1,2})^2} + \frac{\Delta}{\Delta + 2B_{1,2}} \Theta\left(\frac{1}{\log N}\right) + \Theta\left(\frac{1}{(\log N)^2}\right) \\
\Theta\left(\frac{1}{\log N}\right) &< \Delta^2 \\
\Delta &> \Theta\left(\frac{1}{\sqrt{\log N}}\right)
\end{aligned}$$

If we choose  $h = \Theta(N^{-3/2})$ ,

$$\begin{aligned}\Theta\left(\frac{\log N}{N}\right) &< \frac{\Delta^2}{(\Delta + 2B_{1,2})^2} + \frac{\Delta}{\Delta + 2B_{1,2}}\Theta\left(\frac{1}{\sqrt{N}}\right) + \Theta\left(\frac{1}{N}\right) \\ \Theta\left(\frac{\log N}{N}\right) &< \Delta^2 \\ \Delta &> \Theta\left(\sqrt{\frac{\log N}{N}}\right).\end{aligned}$$

These two results correspond to two different levels of sparsity. If  $d + \tau > 3 \log(4N/\epsilon)$ , then the theoretically suggested value of  $h$  is  $h = \Theta((N \log N)^{-1})$  resulting in  $\Delta > \Theta((\log N)^{-1/2})$ . If  $d + \tau > \Theta(N/(\log N))$ , then the suggested value is  $h = \Theta(N^{-3/2})$  resulting in  $\Delta > \Theta(\sqrt{(\log N)/N})$ . Hence, in the sparse case there is a factor of  $\sqrt{N/(\log N)}$  difference between our bound and the general lower bound, and in the dense case there is a factor of  $\sqrt{\log N}$  difference. Note that  $h = \Theta(N^{-3/2})$  still satisfies all assumptions when  $d + \tau > 3 \log(4N/\epsilon)$ , but does not give the lowest bound on the mis-clustering rate.

The above analysis considered  $R$  to be constant. If  $R = \Theta(\log N)$ , then the results change substantially. Observe that the general lower bound from Theorem 3.7 now becomes,

$$\Delta \geq B_{1,1}(1 - B_{1,1}) \left(\frac{2}{3}\right) \left(\frac{2}{3N} \left(\frac{\log 2}{2}(1 - \epsilon) - \Theta(\log N) - \frac{\log 2}{8}\right)\right)^{1/2}.$$

Hence, for a sufficiently large  $N$ ,  $\Delta$  can be zero. The results from Theorem 3.6 now become,

$$\Theta\left(\frac{\log N}{N}\right) + h^2 \Theta((N \log N)^2) < \frac{\Delta^2}{(\Delta + 2B_{1,2})^2} + \frac{\Delta h \Theta(N \log N)}{\Delta + 2B_{1,2}} + h^2 \Theta((N \log N)^2).$$

For  $h = \Theta((N \log N)^{-1})$ ,

$$\begin{aligned}\Theta\left(\frac{(\log N)^2}{N}\right) + \Theta(1) &< \frac{\Delta^2}{(\Delta + 2B_{1,2})^2} + \frac{\Delta}{\Delta + 2B_{1,2}}\Theta(1) + \Theta(1) \\ \Delta &> \sqrt{\Theta\left(\frac{(\log N)^2}{N}\right) - \Theta(1)}.\end{aligned}$$

Thus, both the lower and upper bound can reach zero for a sufficiently large  $N$ , but with a different dependence on the model parameters.