Impact of regularization on Spectral Clustering

Antony Joseph*and Bin Yu†

December 5, 2013

Abstract

The performance of spectral clustering is considerably improved via regularization, as demonstrated empirically in Amini et al. [2]. Here, we provide an attempt at quantifying this improvement through theoretical analysis. Under the stochastic block model (SBM), and its extensions, previous results on spectral clustering relied on the minimum degree of the graph being sufficiently large for its good performance. We prove that for an appropriate choice of regularization parameter τ, cluster recovery results can be obtained even in scenarios where the minimum degree is small. More importantly, we show the usefulness of regularization in situations where not all nodes belong to well-defined clusters. Our results rely on the analysis of the spectrum of the Laplacian as a function of τ. As a byproduct of our bounds, we propose a data-driven technique *DK-est* (standing for estimated Davis-Kahn bounds), for choosing the regularization parameter. This technique is shown to work well through simulations and on a real data set.

1 Introduction

The problem of identifying communities, or clusters, in large networks is an important contemporary problem in statistics. Spectral clustering is one of the more popular techniques for such purposes, chiefly due to its computational advantage and generality of application. The algorithm's generality arises from the fact that it is not tied to any modeling assumptions on the data, but is rooted in intuitive measures of community structure such as *sparsest cut* based measures [12], [26], [18], [22]. Other examples of applications of spectral clustering include manifold learning [4], image segmentation [26], and text mining [10].

The canonical nature of spectral clustering also generates interest in variants of the technique. Here, we attempt to better understand the impact of regularized forms of spectral clustering for community detection in networks.

*Department of Genome Dynamics, Lawrence Berkeley National Laboratory, and Department of Statistics, University of California, Berkeley. email: AntonyJoseph@lbl.gov

†Department of Statistics and EECS, University of California, Berkeley. email: binyu@stat.berkeley.edu

In particular, we focus on the Perturbed Spectral Clustering (PSC) procedure proposed in Amini et al. [2]. Their empirical findings demonstrates that the performance of the PSC algorithm, in terms of obtaining the correct clusters, is significantly better for certain values of the regularization parameter. An alternative form of regularization was studied in Dasgupta et al. [9], Chaudhuri et al. [7], and Qin and Rohe [24].

This paper provides an attempt to provide a theoretical understanding for the regularization in the PSC algorithm under the stochastic block model (SBM) framework. We also address the practical issue of the choice of regularization parameter. The following are the main contributions of the paper.

(a) In Section 3 we demonstrate improvements in eigenvector perturbation bounds through regularization. In particular, for a graph with n nodes, previous theoretical analyses for spectral clustering, under the SBM and its extensions, [25],[7], [27], [11] assumed that the minimum degree of the graph scales at least by a polynomial power of $\log n$. Even when this assumption is satisfied, the dependence on the minimum degree is highly restrictive when it comes to making inferences about cluster recovery. Our analysis provides bounds on the perturbation of eigenvectors of the regularized Laplacian. These bounds, when optimized over the regularization parameter, potentially do not depend on the above mentioned constraint on the minimum degree. As an example, for an SBM with two blocks (clusters), our bounds are inversely related to the maximum degree, as opposed to the minimum degree.

(b) In Section 4 we demonstrate that regularization has the potential of addressing a situation, often encountered in practice, where not all nodes belong to well-defined clusters. Without regularization, these nodes would hamper with the clustering of the remaining nodes in the following way: In order for spectral clustering to work, the top eigenvectors - that is, the eigenvectors corresponding to the largest eigenvalues of the Laplacian - need to be able to discriminate between the clusters. Due to the effect of nodes that do not belong to well-defined clusters, these top eigenvectors do not necessarily discriminate between the clusters with ordinary spectral clustering. With a proper choice of regularization parameter, we show that this problem can be rectified. We also demonstrate this on simulated and real datasets.

(c) In Section 6 we provide a data dependent technique for choosing the regularization parameter based on our bounds. We demonstrate that it works well through simulations and on a real data set.

A crucial ingredient in (a) and (b) is the analysis of the spectrum of the Laplacian as a function of the regularization parameter. Assuming that there are K clusters, an adequate gap between the top K eigenvalues and the remaining eigenvalues, ensures that these clusters can be estimated well [28], [22], [18]. Such a gap is commonly referred to as the *eigen gap*. In the situation considered in (b), an adequate eigen gap may not exist for the unregularized Laplacian.

We show that regularization works by creating a gap, allowing us to recover the clusters.

The paper is divided as follows. In the next section we discuss preliminaries. In particular, in Subsection 2.1 we review the PSC algorithm of [2], while in Subsection 2.2 we review the stochastic block model. Our theoretical results, described in (a) and (b) above, are provided in Sections 3 and 4. Section 5 discusses the regularization studied in the papers [9], [7], [24] in relation to our work. Section 6 describes a data dependent method for choosing the regularization parameter, motivated by our bounds. Section 7 provides the high-level idea behind the proofs of results in Sections 3 and 4.

2 Preliminaries

In this section we review the perturbed spectral clustering (PSC) algorithm of Amini et al. [2] and the stochastic block model framework.

We first introduce some basic notation. A graph with n nodes and edge set E is represented by the $n \times n$ symmetric adjacency matrix $A = ((A_{ij}))$, where $A_{ij} = 1$ if there is an edge between i and j, otherwise A_{ij} is 0. In other words,

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

Given such a graph, the typical community detection problem is synonymous with finding a partition of the nodes. A good partitioning would be one in which there are few edges between the various components of the partition, compared to the number of edges within the components. Various measures for goodness of a partition have been proposed, chiefly the Ratio Cut [12] and Normalized Cut [26] . However, minimization of the above measures is an NP-hard problem since it involves searching over all partitions of the nodes. The significance of spectral clustering partly arises from the fact that it provides a continuous approximation to the above discrete optimization problem [12], [26].

2.1 The PSC Algorithm [2]

We now describe the PSC algorithm [2], which is a regularized version of spectral clustering. Denote by $D = diag(\hat{d}_1, \ldots, \hat{d}_n)$ the diagonal matrix of degrees, where $\hat{d}_i = \sum_{j=1}^{n} A_{ij}$. The normalized (unregularized) symmetric graph Laplacian is defined as

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

Regularization is introduced in the following way: Let J be a constant matrix with all entries equal to $1/n$. Then, in perturbed spectral clustering one constructs a new adjacency matrix by adding τJ to the adjacency matrix A and computing the corresponding Laplacian. In particular, let

$$A_\tau = A + \tau J,$$

where $\tau > 0$ is the regularization parameter. The corresponding regularized symmetric Laplacian is defined as

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

Here, $D_\tau = diag(\hat{d}_{1,\tau}, \ldots, \hat{d}_{n,\tau})$ is the diagonal matrix of 'degrees' of the modified adjacency matrix A_τ. In other words, $\hat{d}_{i,\tau} = \hat{d}_i + \tau$.

The PSC algorithm for finding K communities is described in Algorithm 1. The algorithm first computes V_τ, the $n \times K$ eigenvector matrix corresponding to the K largest (in absolute terms) eigenvalues of L_τ. The columns of V_τ are taken to be orthogonal. The rows of V_τ, denoted by $V_{i,\tau}$, for $i = 1, \ldots, n$, corresponds to the nodes in the graph. Clustering the rows of V_τ provides a clustering of the nodes. We remark that with $\tau = 0$, the PSC Algorithm 1 corresponds to the usual spectral clustering algorithm.

We also remark that there is flexibility in the choice of the clustering procedure in Step 2 of the algorithm we describe below. A natural choice would be the k-means algorithm [22], [25], [17]. This was also used for the PSC algorithm [2]. Algorithm 2, proposed in McSherry [20], provides an alternative procedure that has been used in the literature. This algorithm, along with variants [7], [3], uses pairwise distances of the rows of the eigenvector matrix to do clustering.

Algorithm 1 The PSC algorithm [2] with regularization parameter τ

Input : Laplacian matrix L_τ.
Step 1: Compute the $n \times K$ eigenvector matrix V_τ.
Step 2: Use Algorithm 2 to cluster the rows of V_τ into K clusters.

Our main theoretical results concerns the impact of regularization on perturbation of eigenvectors. In order to translate these results into implications for cluster recovery, we use Algorithm 2 in Step 2 of the PSC Algorithm 1 since it is easier to analyze. Our simulation results will use the k-means algorithm in Step 2 instead.

Algorithm 2 Clustering procedure in McSherry [20] with parameter $t > 0$.

Input : Data points $V_{i,\tau}$, for $i = 1, \ldots, n$.
Set $\hat{k}(1) = 1$ and $\mathcal{S} = \{2, \ldots, n\}$.
while $\mathcal{S} \neq \emptyset$ **do**
  Choose a j in $\mathcal{S}$ at random. Set $\mathcal{S} = \mathcal{S} - \{j\}$.
  if For some $i \in \mathcal{S}^c$, $\|V_{j,\tau} - V_{i,\tau}\| < t$ **then** assign $\hat{k}(j) = \hat{k}(i)$.
  else
    If there are unused labels in $\{1, \ldots, K\}$ then assign a new label at random for $\hat{k}(j)$. Otherwise set $\hat{k}(j) = 0$.
  end if
end while
Output: Function $\hat{k}$ that provides the cluster labels.

In Algorithm 2 we denote by $\hat{k} : \{1, \ldots, n\} \to \{0, 1, \ldots, K\}$ as the function that provides cluster labels for the n nodes. The algorithm returns $\hat{k}(i) = 0$ if node i could not be assigned to any of the K clusters, although in our analysis of clustering performance we show that all nodes are clustered accurately. The appropriate choice of the parameter t, used as input to the algorithm, will be specified in Lemma 2.

Our theoretical results assume that the data is randomly generated from a stochastic block model (SBM), which we review in the next subsection. While it is well known that there are real data examples where the SBM fails to provide a good approximation, we believe that the above provides a good playground for understanding the role of regularization in the PSC algorithm. Recent works [2], [11], [25], [6], [16] have used this model, and its variants, to provide a theoretical analyses for various community detection algorithms.

Notation

We use $\|.\|$ to denote the spectral norm of a matrix. Notice that for vectors this corresponds to the usual ℓ_2-norm. We use A' to denote the transpose of a matrix, or vector, A.

For positive a_n, b_n, we use the notation $a_n \asymp b_n$ if there exists universal constants c_1, $c_2 > 0$ so that $c_1 a_n \leq b_n \leq c_2 a_n$. Further, we use $b_n \lesssim a_n$ if $b_n \leq c_2 a_n$, for some c_2 not depending on n. The notation $b_n \gtrsim a_n$ is analogously defined.

2.2 The Stochastic Block Model

Given a set of n nodes, the stochastic block model (SBM), introduced in [14], is one among many random graph models that has communities inherent in its definition. We denote the number of communities in the SBM by K. Throughout this paper we assume that K is known. The communities, which represent a partition of the n nodes, are assumed to be fixed beforehand. Denote these by $C_1, \ldots, C_K$.

Given the communities, the edges between nodes, say i and j, are chosen independently with probability depending the communities i and j belong to. In particular, for a node i belonging to cluster C_{k_1}, and node j belonging to cluster C_{k_2}, the probability of edge between i and j is given by

$$P_{ij} = B_{k_1, k_2}. \tag{1}$$

Here, the *block probability matrix*

$$B = ((B_{k_1, k_2})), \quad \text{where } k_1, k_2 = 1, \ldots, K$$

is a symmetric full rank matrix, with each entry between $[0, 1]$.

The $n \times n$ matrix $P = ((P_{ij}))$, given by (1), represents the population counterpart of the adjacency matrix A. From (1), it is seen that the rank of P is also K. This is most readily seen if the nodes are ordered according

to the clusters they belong to, in which case P has a block structure with K blocks. The population counterpart for the degree matrix D is denoted by $\mathscr{D} = diag(d_1, \ldots, d_n)$, where $\mathscr{D} = diag(P\mathbf{1})$. Here $\mathbf{1}$ denotes the column vector of all ones.

Similarly, the population version of the symmetric Laplacian L_τ is denoted by $\mathscr{L}_\tau$, where

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

Here $\mathscr{D}_\tau = \mathscr{D} + \tau I$ and $P_\tau = P + \tau J$. The $n \times n$ matrices $\mathscr{D}_\tau$ and P_τ represent the population counterparts to D_τ and A_τ respectively. Notice that since P has rank K, the same holds for $\mathscr{L}_\tau$.

2.2.1 The Population Cluster Centers

We now proceed to define population cluster centers $\mathsf{cent}_{k,\tau} \in \mathbb{R}^K$, for $k = 1, \ldots, K$, for the K block SBM. These points are defined so that the rows of the eigenvector matrix $V_{i,\tau}$, for $i \in C_k$, are expected to be scattered around $\mathsf{cent}_{k,\tau}$.

Denote by $\mathscr{V}_\tau$ an $n \times K$ matrix containing the eigenvectors of the K largest eigenvalues (in absolute terms) of $\mathscr{L}_\tau$. As with V_τ, the columns of $\mathscr{V}_\tau$ are also assumed to be orthogonal.

Notice that both $\mathscr{V}_\tau$ and $-\mathscr{V}_\tau$ are eigenvector matrices corresponding to $\mathscr{L}_\tau$. This ambiguity in the definition of $\mathscr{V}_\tau$ is further complicated if an eigenvalue of $\mathscr{L}_\tau$ has multiplicity greater than one. We do away with this ambiguity in the following way: Let $\mathcal{H}$ denote the set of all $n \times K$ eigenvector matrices of $\mathscr{L}_\tau$ corresponding to the top K eigenvalues. We take,

$$\mathscr{V}_\tau = \arg\min_{H \in \mathcal{H}} \|V_\tau - H\|, \tag{2}$$

The matrix $\mathscr{V}_\tau$, as defined above, represents the population counterpart of the matrix V_τ.

Let $\mathscr{V}_{i,\tau}$ denote the i-th row of $\mathscr{V}_\tau$. Notice that since the set $\mathcal{H}$ is closed under the $\|.\|$ norm, one has that $\mathscr{V}_\tau$ is also an eigenvector matrix of $\mathscr{L}_\tau$ corresponding to the top K eigenvalues. Consequently, the rows $\mathscr{V}_{i,\tau}$ are the same across nodes belonging to a particular cluster (See, for example, Rohe et al. [25] for a proof of this fact). In other words, there are K distinct rows of $\mathscr{V}_{i,\tau}$, with each row corresponding to nodes from one of the K clusters. We denote the K distinct rows of $\mathscr{V}_\tau$ as $\mathsf{cent}_{1,\tau}, \ldots, \mathsf{cent}_{K,\tau}$.

Notice that the $\mathsf{cent}_{1,\tau}, \ldots, \mathsf{cent}_{K,\tau}$ depend on the sample eigenvector matrix V_τ through (2), and consequently is a random quantity. However, the following lemma shows that the pairwise distances between the $\mathsf{cent}_{k,\tau}$'s are non-random and, more importantly, independent of τ.

Lemma 1. *Let $1 \le k, k' \le K$. Then,*

$$\|\mathsf{cent}_{k,\tau} - \mathsf{cent}_{k',\tau}\| = \begin{cases} 0, & \text{if } k = k' \\ \sqrt{\frac{1}{|C_k|} + \frac{1}{|C_{k'}|}}, & \text{if } k \neq k' \end{cases}$$

The above lemma, which is proved in Appendix D.4, states that the pairwise distances between the population cluster centroids only depends on the sizes of the various clusters and not on the regularization parameter τ.

For any node i, denote by $k(i)$ the index of the cluster in which node i belongs to. In other words,

$$k(i) = k, \quad \text{if node } i \text{ belongs to cluster } C_k.$$

2.2.2 Relating Perturbation Of Eigenvectors And Cluster Recovery

Recall that spectral clustering works by clustering the rows of the $n \times K$ sample eigenvector matrix, denoted by $V_{i,\tau}$, for $i = 1, \ldots, n$. If the points $V_{i,\tau}$ occupy K well separated regions in $\mathbb{R}^K$, with each region corresponding to one of $C_1, \ldots, C_K$, then the clustering procedure in Step 2 of the PSC Algorithm 1, when applied to the $V_{i,\tau}$'s, should able to identify $C_1, \ldots, C_K$.

Notice that the cluster center corresponding to a node i is given by $\mathsf{cent}_{k(i),\tau}$. In order for spectral clustering to work, the distance of each $V_{i,\tau}$ from its cluster center $\mathsf{cent}_{k(i),\tau}$, given by

$$\hat{\delta}_\tau = \max_{i=1,\ldots,n} \|V_{i,\tau} - \mathsf{cent}_{k(i),\tau}\| \tag{3}$$

should be small relative to the pairwise distance between the centers. The following quantity represents this relative perturbation:

$$\mathsf{pert}_\tau = \frac{\hat{\delta}_\tau}{\min_{k \neq k'} \|\mathsf{cent}_{k,\tau} - \mathsf{cent}_{k',\tau}\|} \tag{4}$$

If pert_τ is small, then the distance of each $V_{i,\tau}$ from its cluster center, which is at most $\hat{\delta}_\tau$, is small compared to the distances between the centers. In particular, if $\mathsf{pert}_\tau < 1/2$ then this implies that among all the cluster centers $\mathsf{cent}_{k,\tau}$, each $V_{i,\tau}$ is closest to its cluster center, given by $\mathsf{cent}_{k(i),\tau}$. Following the pattern of Rohe et al. [25], we say that no nodes are misclustered if $\mathsf{pert}_\tau < 1/2$ holds.

Under a slightly stronger condition on pert_τ, one can show cluster recovery using Algorithm 2. This shown in the lemma below. The proof of the lemma can be inferred from McSherry [20]. For completeness, we provide its proof below.

Lemma 2. *If $\mathsf{pert}_\tau < 1/4$ then Algorithm 2, with*

$$t = \frac{\min_{k \neq k'} \|\mathsf{cent}_{k,\tau} - \mathsf{cent}_{k',\tau}\|}{2}$$

recovers the clusters $C_1, \ldots, C_K$.

Proof. With t as above, we claim that i and j are in the same cluster iff $\|V_{i,\tau} - V_{j,\tau}\| < t$. For the 'only if' part, assume that i and j are in cluster C_k. Then from triangle inequality,

$$\|V_{i,\tau} - V_{j,\tau}\| \leq \|V_{i,\tau} - \mathsf{cent}_{k,\tau}\| + \|V_{j,\tau} - \mathsf{cent}_{k,\tau}\|.$$

The right side is less than t using $\mathsf{pert}_\tau < 1/4$.

Conversely, if $i \in C_k$ and $j \in C_{k'}$, with $k \neq k'$, then

$$\begin{aligned}\|V_{i,\tau} - V_{j,\tau}\| &\geq \|\mathsf{cent}_{k,\tau} - \mathsf{cent}_{k',\tau}\| - \|V_{i,\tau} - \mathsf{cent}_{k,\tau}\| - \|V_{j,\tau} - \mathsf{cent}_{k,\tau}\| \\ &\geq \frac{\min_{k \neq k'} \|\mathsf{cent}_{k,\tau} - \mathsf{cent}_{k',\tau}\|}{2} = t\end{aligned}$$

□

Recall that from Lemma 1 the denominator in pert_τ (4) does not depend on τ. Consequently, the theoretically best choice of τ would be the one that minimizes the numerator in pert_τ, given by $\hat{\delta}_\tau$ (3), when viewed a function of τ. Note, such a τ cannot be computed in practice since the population centers $\mathsf{cent}_{k,\tau}$ are not known in advance.

3 Perturbation Bounds as a Function of τ

Theorem 3, below, describes our bound for the perturbation $\hat{\delta}_\tau$ (3). This in turn will provide implications for cluster recovery using the PSC Algorithm 1. We first describe the assumptions behind the theorem. Let

$$d_{min} = \min_{i=1,\ldots,n} d_i$$

denote the minimum population degree of the graph. The following quantity will appear frequently in our analysis.

$$\tau_{min} = \max\{\tau, d_{min}\} \tag{5}$$

The regularization parameter τ, which is allowed to depend on n, is taken so that the following is satisfied:

Assumption 1 (Minimum τ)**.**

$$\tau_{min} = \kappa_n \log n, \tag{6}$$

where $\kappa_n > 32$. In other words, $\tau_{min} \gtrsim \log n$.

As mentioned earlier, previous analysis of spectral clustering assumed that the minimum degree d_{min} grows at least as fast as $\log n$. By choosing τ appropriately large, Assumption 1 is satisfied even when the minimum degree is, say, of constant order.

Let

$$1 = \mu_{1,\tau} \geq \ldots \geq \mu_{n,\tau}$$

be the eigenvalues of the regularized population Laplacian $\mathscr{L}_\tau$ arranged in decreasing order. The fact that $\mu_{1,\tau}$ is 1 follows from standard results on the spectrum of Laplacian matrices (see, for example, [28]). As mentioned in the introduction, in order to control the perturbation of the first K eigenvectors,

the eigen gap, given by $\mu_{K,\tau} - \mu_{K+1,\tau}$, must be adequately large, as noted in [28], [22], [18]. Since $\mathscr{L}_\tau$ has rank K, one has $\mu_{K+1,\tau} = 0$. Thus, the eigen gap is simply $\mu_{K,\tau}$. We require the following assumption on the size of the eigen gap.

Assumption 2 (Eigen gap)**.**

$$\mu_{K,\tau} > 20\frac{\sqrt{\log n}}{\sqrt{\tau_{min}}}$$

Notice that both Assumptions 1 and 2 depend on τ. As mentioned above, a large τ will ensure that Assumption 1 is satisfied. However, as such, for a given SBM it is not clear what values of τ allow for Assumption 2 to be satisfied. The next subsection demonstrates that for an appropriately chosen τ, improvements in perturbation bounds can be obtained under assumptions weaker than that used in literature. To do this, we require the following theorem which provides bounds on the perturbation of eigenvectors for any τ satisfying the above two assumptions.

Theorem 3. *Let Assumptions 1 and 2 hold. Then, with probability at least* $1 - (2K + 5)/n$*,*

$$\hat{\delta}_\tau = \max_{i=1,\ldots,n} \|V_{i,\tau} - \mathsf{cent}_{k(i),\tau}\| \leq \delta_{\tau,n} \quad \textit{for } i = 1, \ldots, n \tag{7}$$

where $\delta_{\tau,n}$ *is the maximum over* $i = 1, \ldots, n$ *of*

$$\frac{1}{\mu_{K,\tau}^2}\left[293\frac{\sqrt{\log n}}{\tau_{min}} + 31\frac{\sqrt{\log n}}{\sqrt{\tau_{min}|C_{k(i)}|}} + 12\frac{K\sqrt{\log n}}{\tau_{min}^{3/2}}\right]. \tag{8}$$

The above theorem is proved in Appendix B. The results in Theorem 3 are valid even when the number of clusters K is allowed to grow with n. However, for convenience, in this section we restrict our attention to the case where K fixed. We also assume that $|C_{k(i)}| \asymp n$ for each i. Consequently, the first term $\delta_{\tau,n}$, given in Theorem 3, is larger implying that

$$\delta_{\tau,n} \asymp \frac{\sqrt{\log n}}{(\mu_{K,\tau}\sqrt{\tau_{min}})^2}. \tag{9}$$

Bound (7) also strengthens upon the Davis-Kahan (DK) bound for perturbation of eigenvectors (see for example [28], [25]). Direct application of the DK bound would lead to a weaker $\mu_{K,\tau}\sqrt{\tau_{min}}$ in the denominator of (9), instead of $(\mu_{K,\tau}\sqrt{\tau_{min}})^2$ that we get. The proof technique involves results for the concentration of Laplacian of random graphs [23], [19]. The improvement in Davis-Kahan bounds, given in (7), arises from the extension of the techniques in [3] to normalized graph Laplacians.

The bound in Theorem 3, which relies on the Davis-Kahan theorem, also provides an insight into the role of the regularization parameter τ. As a consequence of the Davis-Kahan theorem, the spectral norm of the difference in the

sample and population eigenvector matrices is dictated by

$$\frac{\|L_\tau - \mathscr{L}_\tau\|}{\mu_{K,\tau}}. \tag{10}$$

Increasing τ will ensure that the Laplacian L_τ will be well concentrated around $\mathscr{L}_\tau$. Indeed, it can be shown that

$$\|L_\tau - \mathscr{L}_\tau\| \lesssim \frac{\sqrt{\log n}}{\sqrt{\tau_{min}}} \tag{11}$$

with high probability. This bound does not require any assumption on the minimum degree, provided τ has the form (6). However, increasing τ also has the effect of decreasing the eigen gap, which in this case is $\mu_{K,\tau}$, since the population Laplacian becomes more like a constant matrix upon increasing τ. Thus the optimum τ results from the balancing out of these two competing effects.

This is akin to a 'bias-variance' trade-off, with the 'variance' term represented by $\|L_\tau - \mathscr{L}_\tau\|$, while the 'bias' term is represented by $1/\mu_{K,\tau}$. In particular, our results indicate that the τ that minimizes δ_τ, or equivalently, maximizes

$$\mu_{K,\tau}\sqrt{\tau_{min}},$$

can be chosen as a proxy for the best choice of regularization parameter for PSC.

Independent of our work, a similar argument for the optimum choice of regularization, using the Davis-Kahan theorem, was given in Qin and Rohe [24] for the regulariztion proposed in [9], [7]. However, a quantification of the benefit of regularization, in terms of improvements of the perturbation bounds, as given in Subsections 3.1 and Section 4, was not given in this work. We provide further comparisons in Section 5.

The next subsection quantifies the improvements via regularization. We do this by comparing the perturbation bound $\delta_{\tau,n}$, for a non-zero τ, with $\delta_{0,n}$, the bound for ordinary spectral clustering.

Let C_{max} denote the largest cluster, with $|C_{max}|$ denoting its size. Notice that from Lemma 1 the distance between any two distinct population centroids is atleast $\sqrt{2/|C_{max}|}$. Consequently, if for some choice of regularization parameter $\tau = \tau_n$, one has

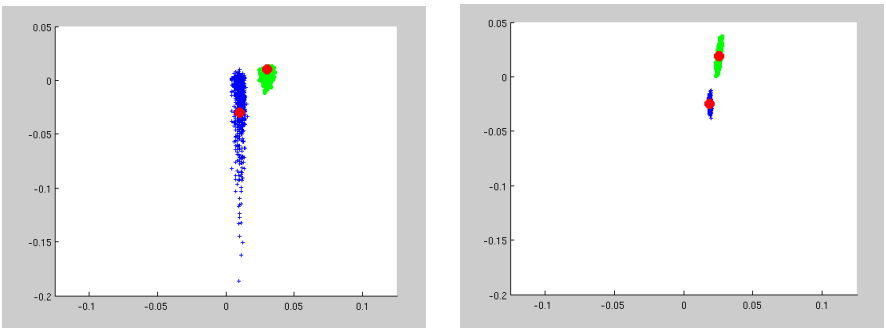
$$\delta_{\tau_n,n} < \frac{\sqrt{2}}{4\sqrt{|C_{max}|}}$$

then from Lemma 2 one gets that the PSC Algorithm 1 recovers the clusters. The following is an immediate consequence of Theorem 3.

Corollary 4. *Let Assumption 1 and 2 be satisfied with regularization parameter $\tau = \tau_n$. If*

$$\frac{(|C_{max}|\log n)^{1/2}}{\mu^2_{K,\tau_n}\tau_{min}} = o(1)$$

then the PSC Algorithm 1 recovers the clusters $C_1, \ldots, C_K$ with probability tending to 1 for large n. Here recall that $\tau_{min} = \min\{\tau_n, d_{min}\}$.



(a) Unregularized (b) Regularized

Figure 1: Scatter plot of first two eigenvectors. Here the block probability matrix B is as in (13). Plot a) corresponds to $\tau = 0$, while b) has $\tau = \gamma_{1,n}$, which in this case is 53. The solid red dot in both plots indicate the population cluster centers. For each τ, the rows of the sample eigen vector matrix, given by $V_{i,\tau}$, are also plotted. The blue '+'s correspond to the $V_{i,\tau}$, with nodes i in the second cluster, while the green circles correspond to the nodes in the first cluster.

3.1 Improvements In Perturbation Bounds

This subsection uses Theorem 3 to demonstrate improvements, as a result regularization, over previous analyses of eigenvector perturbation. In particular, we demonstrate that for a particular choice of regularization parameter, the dependence of the bounds on the minimum degree can be removed. In this subsection we assume that the community sizes are equal. While this assumption is not really needed, it makes the anaylsis considerably less messy.

For the stochastic block model, all nodes in a particular cluster have the same expected degree. In the lemma below we consider a two community stochastic block model. Without loss, assume that each node in cluster 1 has degree $\gamma_{1,n}$ that is larger than the degree of nodes in cluster 2, given by $\gamma_{2,n}$. Standard results on spectral clustering provide bounds on the perturbation of the eigenvectors that are dictated by the minimum degree, which in this case is $\gamma_{2,n}$. However, we show that, through regularization, it is dictated by the maximum degree $\gamma_{1,n}$, without any assumption on the magnitude of the minimum degree.

Lemma 5. *Consider the two community stochastic block model with*

$$B = \begin{pmatrix} p_{1,n} & q_n \\ q_n & p_{2,n} \end{pmatrix}.$$

Assume $q_n \leq p_{2,n} < p_{1,n}$. Then, the second eigenvalue $\mu_{2,\tau}$ is a decreasing function of τ and for $\tau_n = \gamma_{1,n}$,

$$\mu_{2,\tau_n} \geq \frac{\mu_{2,0}}{4}. \tag{12}$$

Further, if $\sqrt{\log n}/(\mu_{2,0}\sqrt{\gamma_{1,n}}) = o(1)$ then

$$\max_i \|V_{i,\tau_n} - \mathsf{cent}_{k(i),\tau_n}\| \lesssim \frac{\sqrt{\log n}}{(\mu_{2,0}\sqrt{\gamma_{1,n}})^2}$$

with probability tending to 1 for large n.

The above lemma states that with the regularization parameter set as $\tau_n = \gamma_{1,n}$, the perturbation of eigenvectors is dictated by the maximum degree $\gamma_{1,n}$, instead of the minimum degree $\gamma_{2,n}$. Direct application of our bounds to ordinary spectral clustering would lead to the larger

$$\frac{\sqrt{\log n}}{(\mu_{2,0}\sqrt{\gamma_{2,n}})^2},$$

along with the stronger requirement that $\sqrt{\log n}/(\mu_{2,0}\sqrt{\gamma_{2,n}}) = o(1)$. The following is an immediate consequence of Corollary 4.

Corollary 6. *For the two block SBM, if the maximum population degree $\gamma_{1,n}$ grows faster than*

$$\frac{\sqrt{n \log n}}{\mu_{2,0}^2}$$

then the PSC Algorithm 1, with regularization parameter $\tau_n = \gamma_{1,n}$, recovers the clusters C_1 and C_2 with probability tending to 1 for large n.

Figure 1 illustrates this with $n = 2000$ and edge probability matrix

$$B = \begin{pmatrix} .003 & .003 \\ .003 & .05 \end{pmatrix}. \tag{13}$$

The figure provides the scatter plot of the first two eigenvectors of the unregularized and regularized sample Laplacians. The figure on the left corresponds to the usual spectral clustering, while the plot on the right corresponds to regularized spectral clustering with $\tau_n = \gamma_{1,n}$, as suggested in Lemma 5. Notice that there is considerably less scattering for points in cluster 2 with regularization. Also note that, as predicted by the theory, the distance between the population cluster centers does not change with regularization.

For the two block model, the eigenvalue $\mu_{2,\tau}$ could be explicitly computed. We extend the above to the K block model, with constant interaction between clusters. In other words, let

$$B = \begin{pmatrix} p_{1,n} & q_n & \cdots & q_n \\ q_n & p_{2,n} & \cdots & q_n \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & q_n & p_{K,n} \end{pmatrix}. \tag{14}$$

The number of communities K is assumed to be fixed. Without loss, assume that $p_{1,n} \geq p_{2,n} \ldots \geq p_{K,n}$ and let $q_n < p_{K,n}$. We demonstrate that the perturbation of the eigenvectors is dictated by the $\gamma_{K-1,n} = d_i$, where i is any node belonging to cluster $K-1$. In other words, it is dictated by the second smallest degree, as opposed to the smallest degree.

Lemma 7. *Let B be as in* (14) *and assume $q_n = o(p_{K-1,n})$. If $\log n/\gamma_{K-1,n} = o(1)$, then, with $\tau_n = \gamma_{K-1,n}$ the eigenvalue μ_{K,τ_n} is bounded away from zero. Further,*

$$\max_{i=1,\ldots,n} \|V_{i,\tau_n} - \mathsf{cent}_{k(i),\tau_n}\| \lesssim \frac{\sqrt{\log n}}{\gamma_{K-1,n}}.$$

with probability tending to one as n tends to infinity.

Notice that since $|C_{max}| \asymp n$, the above result, along with Corollary 4, leads to the following analogue of Corollary 6.

Corollary 8. *Let B as in* (14) *satisfy the assumption of Lemma 7. If $\gamma_{K-1,n}$ grows at a rate faster than*

$$(n \log n)^{1/2},$$

then, for large n, the PSC Algorithm 1 with $\tau_n = \gamma_{K-1,n}$, recovers the clusters with high probability.

Both Lemmas 5 and 7 are proved in Appendix E. It may seem surprising that the performance does not depend on the degree of the lowest block, that is, $\gamma_{K,n}$. One way of explaining this is that if one can do a good job identifying the top $K-1$ highest degree clusters then the cluster with the lowest degree can also be identified simply by eliminating nodes not belonging to this cluster. We remark that the fact that our results do not depend on the minimum degree is not due to our proof technique, but because of the regularization. Indeed, plots such as in Figure 1 demonstrates that the perturbation of eigenvectors depends on the minimum degree with ordinary spectral clustering.

4 Selection Of Strong Clusters

In many practical situations, not all nodes belong to clusters that can be estimated well. As mentioned in the introduction, these nodes interfere with the clustering of the remaining nodes in the sense that none of the top eigenvectors might discriminate between the nodes that do belong to well-defined clusters. As an example of a real life data set, we consider the political blogs data set, which has two clusters, in Subsection 6.2. With ordinary spectral clustering, the top two eigenvectors do not discriminate between the two clusters. Infact, it is only the third eigenvector that discriminates between the two clusters. This results in bad clustering performance when the first two eigenvectors are considered. However, regularization rectifies this problem by 'bringing up' the important eigenvector, thereby allowing for much better performance.

We model the above situation in the following way: Consider a stochastic block model, as in (14), with $K + K_w$ blocks. In particular, let the block probability matrix be given by

$$B = \begin{pmatrix} B_s & B_{sw} \\ B'_{sw} & B_w \end{pmatrix}, \tag{15}$$

where B_s is a $K \times K$ matrix with $(p_{1,n}, \ldots, p_{K,n})$ in the diagonal and $q_{s,n}$ in the off-diagonal. Further, B_{sw}, B_w are $K \times K_w$ and $K_w \times K_w$ dimensional matrices respectively.

In the above $(K+K_w)$-block SBM, the top K blocks corresponds to the well-defined clusters, while the bottom K_w blocks corresponds to less well-defined clusters. We refer to the K well-defined clusters as *strong clusters*. The K_w less well-defined clusters are called *weak clusters*. These are formalized below. The matrix B_s models the distribution of edges between the nodes belonging to the strong clusters, while the matrix B_w has the corresponding role for the weak clusters. The matrix B_{sw} models the interaction between the strong and weak clusters.

We only assume that the rank of B_s is K. Thus, the rank of B is at least K. We remark that if $\text{rank}(B) = K$, then the model (15) encompasses certain degree-corrected stochastic block models (see Karrer and Newman [16] for a description of the model). We provide further remarks on this in Section 8.

As before, we assume that K is known and does not grow with n. The number of weak clusters, K_w, need not be known and is allowed to grow arbitrarily with n. We do not even place any restriction on the sizes of a weak cluster. Indeed, we even entertain the case that each of the K_w clusters has one node. In other words, the nodes in the weak clusters do not even need to form clusters.

We now formalize our notion of strong and weak clusters. As before, let $\gamma_{1,n}, \ldots \gamma_{K,n}$ denote the degrees of the nodes in the K strong clusters. For ease of analysis, we make the following simplifying assumptions. Assume that $p_{k,n} = p_{K,n}$, for each $k = 1, \ldots K$, and that the strong clusters $C_1, \ldots, C_K$ have equal sizes. Notice that in this case $\gamma_{k,n} = \gamma_{K,n}$, for $k = 1, \ldots, K$.

Let b_{sw} and b_w be defined as the maximum of the elements in B_{sw} and B_w respectively. Denoting by C_{K+1} the set of nodes belong to a weak cluster, we define

$$\gamma_{K+1,n} = (n - |C_{K+1}|)b_{sw} + |C_{K+1}|b_w.$$

The quantity $\gamma_{K+1,n}$ is a bound on the maximum degree of a node in a weak cluster. We make the following assumptions,

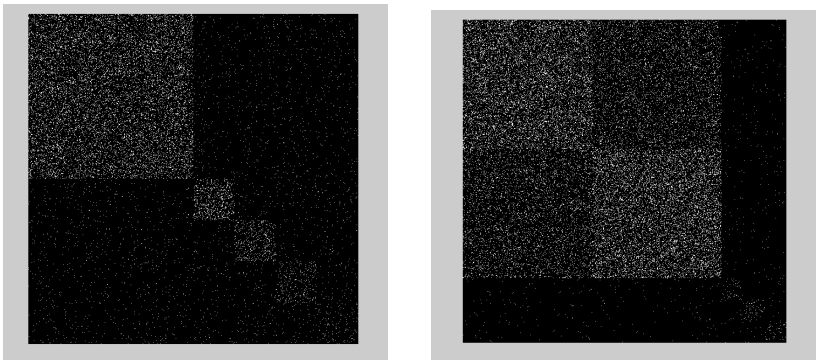
$$\frac{\gamma_{K+1,n}}{\gamma_{K,n}} = o(1). \tag{16}$$

$$b_{sw} \lesssim b_w \tag{17}$$

$$\frac{q_{s,n}}{p_{K,n}} \leq \kappa < 1, \tag{18}$$

where κ is a quantity that does not depend on n. Assumption (16) simply states that the strong clusters have degrees that is of a high order of magnitude

than the weak clusters. Assumption (17) states that the interaction between the strong and weak clusters, denoted by b_{sw}, is not too large. More precisely, it is allowed to be at most the same order of b_w, which is a proxy of how well connected each weak cluster is. Further, Assumption (18) states that, in the absence of the weak clusters, one can distinguish between the strong clusters easily.



(a) $|C_{K+1}| \asymp n$ (b) $|C_{K+1}| = o(n)$

Figure 2: Adjacency matrices for a block model with strong and weak clusters. a) Here $n = 2000$, with one strong cluster ($K = 1$) and four weak clusters ($M = 4$). The first 1000 nodes are taken to belong to the strong cluster 1, while the remaining 1000 nodes were evenly split between the weak clusters 2 to 5. The matrix B has diagonal elements (.025, .012, .009, .006, .004) and off-diagonal element .0025. b) Here $n = 2000$, two strong clusters ($K = 2$), three weak clusters ($M = 3$). The first 1600 nodes are evenly split between the two strong clusters, with the remaining nodes split evenly between the weak clusters. The matrix B_s in (15) has diagonal elements .025 and off-diagonal elements .015. The diagonal elements of B_w are taken as (.007, .0071, .0069). The remaining elements of B are taken to be .001.

For a given assignment of nodes in one of the $K + K_w$ clusters, we denote L_τ, $\mathscr{L}_\tau$ to be the sample, population regularized Laplacians respectively. As before, let $\mu_{k,\tau}$ for $k = 1, \ldots, n$, be the magnitude of the eigenvalues of $\mathscr{L}_\tau$, arranged in decreasing order. Note that $\mu_{k,\tau} = 0$ for $k > K + K_w$ since $\mathscr{L}_\tau$ has rank at most $K + K_w$. We demonstrate the potential of regularization in removing the effect of the weak clusters by consider two scenarios, namely, a) $|C_{K+1}| \asymp n$, and b) $|C_{K+1}| = o(n)$. Example of adjacency matrices drawn from these two cases are shown in Figure 2. The following lemma elucidates why these cases are treated differently.

Lemma 9. *The following holds with $\tau_n = \gamma_{K,n}$:*

Claim 1: If $|C_{K+1}| \asymp n$ then μ_{K+1,τ_n} is bounded away from zero, while μ_{K+2,τ_n} goes to zero for large n.

Claim 2: If $|C_{K+1}| = o(n)$ then μ_{K,τ_n} is bounded away from zero, while μ_{K+1,τ_n} goes to zero for large n.

Lemma 9 states that for regularization parameter $\tau_n = \gamma_{K,n}$, the eigen gap is bounded away from zero. Here the eigen gap is defined as $\mu_{K+1,\tau_n} - \mu_{K+2,\tau_n}$ for the $|C_{K+1}| \asymp n$ case, while it is taken as $\mu_{K,\tau_n} - \mu_{K+1,\tau_n}$ in the $|C_{K+1}| = o(n)$ case. Thus Lemma 9 allows one to control the perturbation of the top $K+1$ sample eigenvectors in the $C_{K+1} \asymp n$ case, and the top K eigenvectors in the $C_{K+1} = o(n)$ case. Note, since such an eigen gap need not exist in the unregularized case, one may not be able to get perturbation results for the top eigenvectors without regularization.

The next essential ingredient is to demonstrate that with $\tau_n = \gamma_{K,n}$, the top population eigenvectors do indeed discriminate between the strong clusters. This is elucidated in Figure 3. The figure deals with the population version of the adjacency matrix in Figure 2(b), where there are 5 ($K = 2$ strong, $M = 3$ weak) clusters. Figures 3(a) and 3(b) show the first 3 eigenvectors of the population Laplacian in the regularized and unregularized cases. We plot the first 3 instead of the first 5 eigenvectors in order to facilitate understanding of the plot. In both cases the first eigenvector is not able to distinguish between the two strong clusters. This makes sense since the first eigenvector of the Laplacian has elements whose magnitude is proportional to square root of the population degrees (see, for example, [28] for a proof of this fact). Consequently, as the population degrees are the same for the two strong clusters, the values for this eigenvector is constant for nodes belonging to the strong clusters.

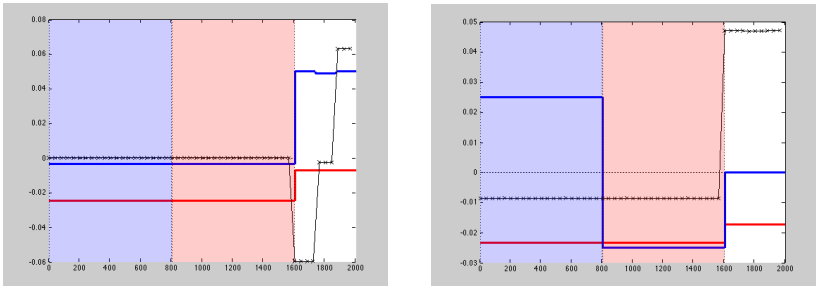
The situation is different for the second population eigenvector. In the regularized case, the second eigenvector is able to distinguish between these two clusters. However, this is not the case for the unregularized case. From Figure 3(a), not even the third unregularized eigenvector is able to distinguish between the strong and weak clusters. Indeed, it is only the fifth eigenvector that distinguishes between the two strong clusters in the unregularized case.

The above provides a different perspective on the role of regularization: Regularization is able to bring out the 'useful' eigenvectors as the 'leading' eigenvectors. From hereon we will assume $\tau_n = \gamma_{K,n}$, unless otherwise specified.

Denote by V_{τ_n} the matrix of top $K+1$ eigenvectors of L_{τ_n}. Similarly, denote by U_{τ_n} the matrix of top K eigenvectors of L_{τ_n}. Lemma 9 allows us to control the perturbation of V_{τ_n} in the $|C_{K+1}| \asymp n$, and that of U_{τ_n} in $|C_{K+1}| = o(n)$ regime. Subsections 4.1 and 4.2 deals with the cases $|C_{K+1}| \asymp n$ and $|C_{K+1}| = o(n)$ respectively.

As before, let $k(i)$ denote the cluster index of node i. In other words, $k(i) = k$, when i is in cluster C_k, and $k = 1, \ldots, K+1$. As before, we use V_{i,τ_n}, U_{i,τ_n} to denote the i-th row of V_{τ_n}, U_{τ_n} respectively.

For the $|C_{K+1}| \asymp n$ scenario, Subsection 4.1 provides results showing that the rows V_{i,τ_n}, for $i = 1, \ldots, n$, are clustered around $K+1$ points $\mathsf{cent}_1, \ldots, \mathsf{cent}_{K+1}$ in $\mathbb{R}^{K+1}$, with the cent_k's being well separated. In particular, we show that the



(a) Unregularized (b) Regularized

Figure 3: First three population eigenvectors corresponding to the adjacency matrix in Figure 2(b). In both plots, the x-axis provides the node indices, while the y-axis gives the eigenvector values. The regularization parameter was taken to be $\gamma_{K,n} = 32.4$. The shaded blue and pink regions corresponds to the nodes belonging to the two strong clusters. The solid red line, solid blue line and $-\times-$ black lines correspond to the first, second and third population eigenvectors respectively.

relative perturbation (see Subsection 2.2.2)

$$\mathsf{pert} = \frac{\max_{i=1,\ldots,n} \|V_{i,\tau_n} - \mathsf{cent}_{k(i)}\|}{\min_{1\le k\neq k'\le K+1} \|\mathsf{cent}_k - \mathsf{cent}_{k'}\|} \tag{19}$$

is small. This allows for recovery of the strong clusters $C_1, \ldots, C_K$, as well as the set C_{K+1} using the PSC Algorithm 1.

The $|C_{K+1}| = o(n)$ scenario is addressed in Subsection 4.2. We show that there are K well separated points $\mathsf{cent}_1^K, \ldots, \mathsf{cent}_K^K$ in $\mathbb{R}^K$ so that the relative perturbation

$$\mathsf{pert}' = \frac{\max_{i\notin C_{K+1}} \|U_{i,\tau_n} - \mathsf{cent}_{k(i)}^K\|}{\min_{1\le k\neq k'\le K} \|\mathsf{cent}_k^K - \mathsf{cent}_{k'}^K\|} \tag{20}$$

is small. In other words, the rows of the matrix U_{τ_n} that are not in C_{K+1} are concentrated around the cent_k^K's. Note, since we say nothing about the nodes in C_{K+1}, this result is in a sense weaker than that in Subsection 4.1. However, since we are dealing with the situation where the size of C_{K+1} is small compared to n, this is not expected to have a significant impact on clustering.

The following quantity will appear as a bound on the perturbation of eigenvectors.

$$\delta_n = \frac{1}{\sqrt{\gamma_{K,n}}} \max\left\{\frac{\sqrt{\log n}}{\sqrt{\gamma_{K,n}}}, \frac{\gamma_{K+1,n}}{\gamma_{K,n}}\right\} \tag{21}$$

Notice that δ_n goes to zero if $\gamma_{K,n}$ grows faster than $\log n$.

4.1 $|C_{K+1}| \asymp n$

Lemma 9 demonstrates that with $\tau_n = \gamma_{K,n}$ there is a gap between the $K+1$-th and $K+2$-th smallest eigenvalues, given by μ_{K+1,τ_n} and μ_{K+2,τ_n} respectively.

This eigen gap allows us to characterize the perturbation of the first $K+1$ eigenvectors of L_{τ_n}. In particular, we demonstrate that with the first $K+1$ eigenvectors of the sample Laplacian matrix L_{τ_n} one can reliably recover the strong clusters. In essence, the eigenvectors treat C_{K+1} as one composite cluster. Theorem 10, below, describes our results.

Theorem 10. *Assume* $\log n/\gamma_{K,n} = o(1)$ *and* $\tau_n = \gamma_{K,n}$. *Then, there exists* $K+1$ *points* $\mathsf{cent}_1, \ldots, \mathsf{cent}_{K+1}$ *in* $\mathbb{R}^{K+1}$ *with*

$$\|\mathsf{cent}_{k'} - \mathsf{cent}_k\| = \sqrt{\frac{1}{|C_k|} + \frac{1}{|C_{k'}|}} \quad \textit{for } k' \neq k \tag{22}$$

so that

$$\max_{i=1,\ldots,n} \|V_{i,\tau_n} - \mathsf{cent}_{k(i)}\| \lesssim \delta_n. \tag{23}$$

with probability tending to one for large n. *Here* δ_n *is as in* (21).

As mentioned earlier, we only wish to identify between the nodes in the strong clusters. The corollary below states the result for recovery of the strong clusters, as well as a C_{K+1}. In Step 2 of Algorithm 1, we set Algorithm 2 to output $K+1$ clusters and take

$$t = \frac{\min_{1\leq k\neq k'\leq K+1} \|\mathsf{cent}_{k'} - \mathsf{cent}_k\|}{2}.$$

We have the following.

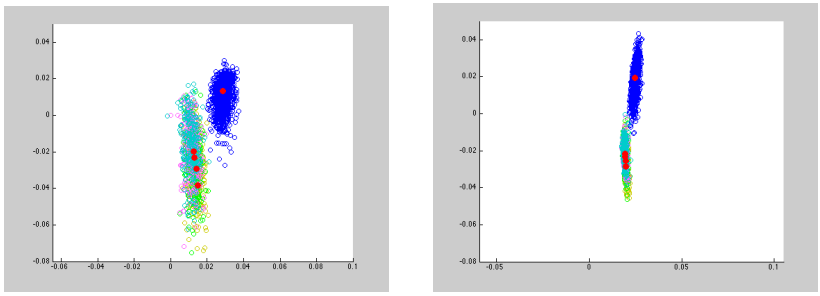
Corollary 11. *If* $\sqrt{n}\delta_n = o(1)$, *then with* $\tau_n = \gamma_{K,n}$, *the PSC Algorithm 1 recovers the strong clusters* $C_1, \ldots, C_K$, *as well as the cluster* C_{K+1}, *with probability tending to one for large* n.

The above follows from (22) and (23), and using the fact that $|C_k| \asymp n$ for each k. Consequently, $\mathsf{pert} \lesssim \sqrt{n}\delta_n$, where pert is given by (19). The result is completed using Lemma 2.

Example sizes of $\gamma_{K,n}$ and $\gamma_{K+1,n}$ for the condition in Corollary 11 to hold are:

1. $\gamma_{K,n}$ grows linearly with n, while $\gamma_{K+1,n}$ is $o(n)$.
2. $\gamma_{K,n}$ grows faster than $\sqrt{n \log n}$, while $\gamma_{K+1,n}$ is $O(n^{1/4}(\log n)^{3/4})$

Figure 4 illustrates the above theorem. Here, $K = 1$ and $M = 4$. Since the four weak clusters are relatively indistinguishable, as can be seen from Figure 2(a), we only wish to separate the strong cluster from the set of weak clusters. Figure 4 shows the scatter plot of the first two eigenvectors of the sample Laplacian matrices. Without regularization, the rows of the eigenvector matrix corresponding to the weak clusters are fairly spread out, as can be seen from Figure 4(a). With regularization, these rows are less spread out, as predicted from the above results. This is shown in Figure 4(b). Indeed, running k-means, with $k = 2$, on the above resulted in a mis-classification of 9.2% of the nodes in the unregularized case, compared with only 1.6% in the regularized case.



(a) Unregularized (b) Regularized

Figure 4: Scatter plot of the first two eigenvectors of the sample Laplacian corresponding to the adjacency matrix in Figure 2(a). The colors denote the different clusters, with blue representing the strong cluster. The solid red points denote the population cluster centers (see Subsection 2.2.1). For (b) we take the regularization parameter $\tau_n = \gamma_{1,n}$, which in this case is 22.5.

4.2 $|C_{K+1}| = o(n)$

For the case $|C_{K+1}| = o(n)$, one gets from Lemma 9 that the difference between μ_{K,τ_n} and μ_{K+1,τ_n} is bounded away from 0. This allows one to bound the perturbation of the first K eigenvectors of L_{τ_n}. Recall that U_{τ_n} is the $n \times K$ matrix of first K eigenvectors of the sample Laplacian L_{τ_n}. The theorem below states that the rows of U_{τ_n}, corresponding to a strong cluster, are close to one of the K points $\mathsf{cent}_1^K, \ldots, \mathsf{cent}_K^K$, depending on the cluster the row belongs to.

Theorem 12. *Assume* $\log n/\gamma_{K,n} = o(1)$ *and* $\tau_n = \gamma_{K,n}$. *Then there exists* K *points* $\mathsf{cent}_1^K, \ldots, \mathsf{cent}_K^K$, *with each* $\mathsf{cent}_k^K \in \mathbb{R}^K$, *so that*

$$\|\mathsf{cent}_{k'}^K - \mathsf{cent}_k^K\| = \sqrt{\frac{1}{|C_k|} + \frac{1}{|C_{k'}|}} \quad \text{for } k' \neq k \tag{24}$$

Further,

$$\max_{i \notin C_{K+1}} \|U_{i,\tau_n} - \mathsf{cent}_{k(i)}^K\| \lesssim \delta_n. \tag{25}$$

with probability tending to one for large n.

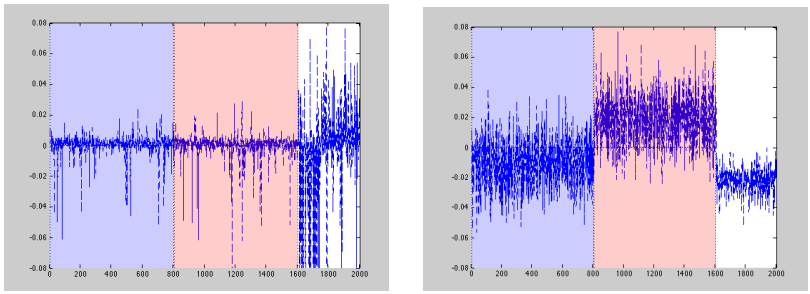
Note that unlike Theorem 10, the above theorem states that the nodes belonging to the strong clusters are concentrated around K points. It says nothing about the nodes belonging to the weak clusters, that is, those in C_{K+1}. In that sense, the results of this subsection are weaker than that in Subsection 4.1. However, since the size C_{K+1} is small compared to n, this should not interfere with the clustering of the rows U_{i,τ_n}.

The weaker nature of Theorem 12 precludes a result on cluster recovery as in Corollary 11. However, notice from Theorem 12 that $\mathsf{pert}' \lesssim \sqrt{n}\delta_n$ with high

probability, where pert' as in (20). Consequently, if we take

$$t = \frac{\min_{1 \le k \ne k' \le K} \|\mathsf{cent}_{k'}^K - \mathsf{cent}_k^K\|}{2}$$

in Algorithm 2 then one has that for nodes i and j in strong clusters, $\|U_{i,\tau_n} - U_{j,\tau_n}\| < t$ iff i and j belong to the same strong cluster. Thus, the nodes in the strong clusters are well separated.



(a) Unregularized (b) Regularized

Figure 5: Second sample eigenvector corresponding to situation in Figure 3. As before, in both plots, the x-axis provides the node indices, while the y-axis gives the eigenvector values. As before, the shaded blue and pink regions corresponds to the nodes belonging to the two strong clusters. For plots (a) & (b) the blue line correspond to the second eigenvector of the respective sample Laplacian matrices.

In Figure 5(a) and 5(b) we show the second sample eigenvector for the two cases in Figure 3(a) and 3(b). Note, we do not show the first sample eigenvector since from Figure 3(a) and 3(b), the corresponding population eigenvectors are not able to distinguish between the two strong clusters. As expected, it is only for the regularized case that one sees that the second eigenvector is able to do a good job in separating the two strong clusters. Running k-means, with $k = 2$, resulted in a mis-classification of 49% of the nodes in the strong clusters in the unregularized case, compared with 16.25% in the regularized case.

5 Comparison With Regularization In [9], [7]

In Dasgupta et al. [9], Chaudhuri et al. [7], the following alternative regularized version of the symmetric Laplacian is proposed:

$$L_{deg,\tau} = D_\tau^{-1/2} A D_\tau^{-1/2}. \tag{26}$$

Here, the subscript *deg* stands for 'degree' since the regular Laplacian is modified by adding τ to the degree matrix D. Notice that for the PSC algorithm, the matrix A in the above expression was replaced by A_τ.

Borrowing from results on the concentration of random graph Laplacians [23], we were able to show concentration results (11) for the regularized Laplacian in the PSC algorithm. This result does not require any assumption on the minimum degree, provided $\tau_{min} \gtrsim \log n$. It was shown in Qin and Rohe [24] that analogous concentration results also hold for the regularized Laplacian $L_{deg,\tau}$. This was shown for the more general degree corrected stochastic block model [16]. However, an analysis of the eigen gap of $L_{deg,\tau}$ (or its population version), as a function of the regularization parameter, was not given in these works. Consequently, it is unclear at this stage whether the benefits of regularization, resulting from the trade-offs between the eigen gap and the concentration bound, as demonstrated in Subsection 3.1 and Section 4 for the PSC algorithm, also hold for the regularization in [7], [24].

Further, it is conjectured in [7], [24] that the regularization parameter taken to be the average degree should be optimal in balancing the bounds. However, for the situation in Lemma 7, the average degree can be too large, especially when there are clusters with very high degree. Indeed, for the K block model considered in Lemma 7, our proof technique also shows that if τ grows faster than $\gamma_{K-1,n}$ then the smallest eigenvalue of $\mathscr{L}_\tau$ goes to 0 for large n. We believe the same to hold true for the regularization (26) as well.

6 Data dependent choice of τ

For the results in Subsection 3.1 and Section 4, the regularization parameter depended on a population quantity which is not known in practice. Here, we propose a data dependent scheme to select the regularization parameter. We also compare it with another scheme that uses the Girvan-Newman modularity [6] . This was suggested in [8]. We use the normalized mutual information criterion (NMI) [2], [29], to quantify the performance of the spectral clustering algorithm in terms of closeness of the estimated clusters to the true clusters. The NMI is a widely used measure of closeness of the estimated clusters to the true clusters.

We now describe our proposed scheme: Our theoretical bounds provide a means to select the regularization parameter τ. One possible route would be to consider the statistic

$$\sqrt{\max\{\tau, \hat{d}_{min}\}}\, \hat{\mu}_{K,\tau}.$$

Here $\hat{d}_{min}$ is the minimum degree of the realized graph and $\hat{\mu}_{K,\tau}$ is K-th smallest eigen value of the sample Laplacian L_τ. From bound (7), it appears that finding the τ that maximizes this criterion should provide a good estimate of the optimum τ. However, the above criterion does not perform well when the average degree of the graph is low, most likely due to the fact the $\hat{\mu}_{K,\tau}$ is a poor substitute for its population counterpart. An alternative criterion, which performs much better, is obtained by directly estimating the quantity in (10). In particular, for each τ in grid, an estimate $\hat{\mathscr{L}}_\tau$ of $\mathscr{L}_\tau$ is obtained using cluster outputted from the PSC algorithm using that τ. Here, the estimate $\hat{\mathscr{L}}_\tau$ is the

population Laplacian corresponding to an edge probability matrix P, as in (1), with an estimated block probability matrix B. In particular, the (k_1, k_2)-th entry of B is taken to be the proportion of edges between the nodes in the estimates of the clusters k_1 and k_2 with the given τ. The following statistic is then considered:

$$DK-est_\tau = \frac{\|L_\tau - \hat{\mathscr{L}}_\tau\|}{\mu_K\left(\hat{\mathscr{L}}_\tau\right)}, \tag{27}$$

where $\mu_K\left(\hat{\mathscr{L}}_\tau\right)$ denotes the the K-th smallest eigenvalue of $\hat{\mathscr{L}}_\tau$. The τ that minimizes the $DK - est_\tau$ criterion is then chosen. Since this criterion provides an estimate of the Davis-Kahan bound, we call it the *DK-est* criterion.

We compare the above to the scheme that uses Girvan-Newman modularity [6], [21], as suggested in [8]. For a particular τ in the grid, the Girvan-Newman modularity is computed for the clusters outputted using the PSC algorithm applied with that τ. The τ that maximizes the modularity value over the grid is then chosen.

Notice that the best possible choice of τ would be the one that simply maximizes the NMI over the selected grid. However, this cannot be computed in practice since calculation of the NMI requires knowledge of the true clusters. Nevertheless, this provides a useful benchmark against which one can compare the other two schemes. We call this the 'oracle' scheme.

6.1 Simulation Results

Figure 6 provides results comparing the two schemes. We perform simulations following the pattern of [2]. If particular, for a graph with n nodes we take the K clusters to be of equal sizes. The $K \times K$ block probability matrix is taken to be of the form

$$B = \mathsf{fac} \begin{pmatrix} \beta w_1 & 1 & ... & 1 \\ 1 & \beta w_2 & ... & 1 \\ ... & ... & ... & ... \\ ... & ... & 1 & \beta w_K \end{pmatrix}.$$

Here, the vector $w = (w_1, \ldots, w_K)$, which are the *inside weights*, denotes the relative degrees of nodes within the communities. Further, the quantity β, which is the *out-in ratio*, represents the ratio of the probability of an edge between nodes from different communities to that of probability of edge between nodes in the same community. The parameter fac is chosen so that the average expected degree of the graph is equal to λ. In the graphs of Figure 6, we denote β and w as OIR and InWei respectively.

Figure 6 compares the two methods of choosing the best τ for various choices of n, K, OIR, InWei and λ. In general, we see that the $DK - est$ selection procedure performs at least as well, and in some cases much better, than the procedure that used the Girvan-Newman modularity. The performance of the two methods is much closer when the average degree is small.

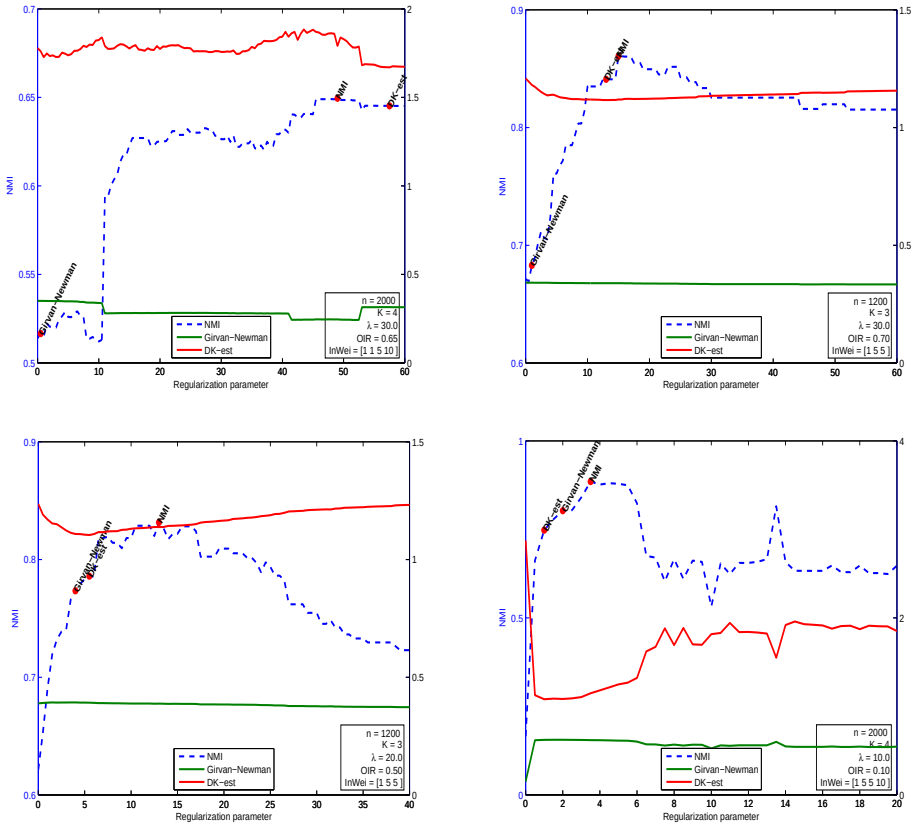


Figure 6: Performance of spectral clustering as a function of τ for stochastic block model for λ values of 30, 20 and 10. The right y-axis provides values for the Girvan-Newman modularities and *DK est* functions, while the left y-axis provides values for the normalized mutual information (NMI). The 3 labeled dots correspond to values of the NMI at τ values which minimizes the *DK-est*, and maximizes the Girvan-Newman modularity and the NMI. Note, the oracle τ, or the τ that maximizes the NMI, cannot be calculated in practice.

6.2 Analysis Of A Real Dataset

We also studied the efficacy of our procedure on the well studied network of political blogs [1]. The data set aims to study the degree of interaction between liberal and conservative blogs over a period prior to the 2004 U.S Presidential Election. The nodes in the networks are select conservative and liberal blog sites. While the original data set had directed edges corresponding to hyperlinks between the blog sites, we converted it to an undirected graph by connecting two nodes with an edge if there is at least one hyperlink from one node to the

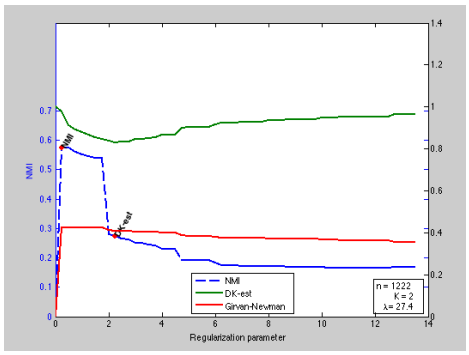
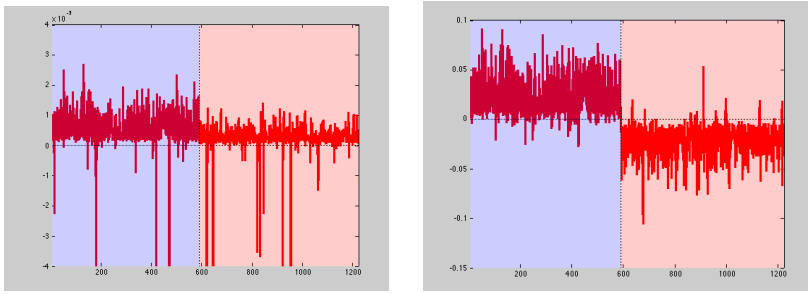


Figure 7: Performance of the three schemes for the political blogs data set [1].

other.



(a) Unregularized (b) Regularized ($\tau = 2.25$)

Figure 8: Second eigenvector of the unregularized and regularized Laplacians for the political blogs data set [1]. The shaded blue and pink regions corresponds to the nodes belonging to the liberal and conservative blogs respectively.

The data set has 1222 nodes with an average degree of 27. Simple spectral clustering, that is with $\tau = 0$, resulted in only 51% of the nodes correctly classified as liberal or conservative. The oracle procedure, with $\tau = 0.5$, resulted in 95% of the nodes correctly classified. The *DK-est* procedure selected $\tau = 2.25$, with an accuracy of 81%. The Girvan-Newman procedure, in this case, outperforms the *DK-est* procedure, providing the same accuracy as the oracle procedure. Figure 7 illustrates these findings.

The results of Section 4 also explain why unregularized spectral clustering performs badly. The first eigenvector in both cases (regularized and unregularized) does not discriminate between the two clusters. In Figure 8, we plot the second eigenvector of the regularized and unregularized Laplacians. The second eigenvector is able to discriminate between the clusters in the regularized case, while it fails to do so in without regularization. Indeed, it is only the third

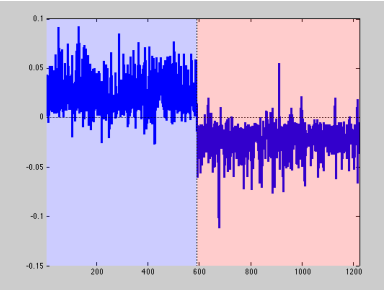


Figure 9: Third eigenvector of the unregularized Laplacian.

eigenvector in the unregularized case that distinguishes between the clusters, as shown in Figure 9.

7 Proof Techniques

Here we discuss the high-level idea behind the proofs of the various theorems. We first discuss the proof of Theorem 3.

7.1 Proof Of Theorem 3

The theorem provides a high-probability bound on

$$\|V_{i,\tau} - \mathscr{V}_{i,\tau}\|, \tag{28}$$

which is the ℓ_2-norm of the difference of the i-th row of the sample and population eigenvectors. From the Davis-Kahan theorem [5], one can get the following bound on the difference of sample and population eigenvector matrices:

$$\|V_\tau - \mathscr{V}_\tau\| \lesssim \frac{\|L_\tau - \mathscr{L}_\tau\|}{\mu_{K,\tau}}$$

Further, from recent results on concentration of Laplacian of random graphs [23], [19], one gets that $\|L_\tau - \mathscr{L}_\tau\| \lesssim \sqrt{\log n}/\sqrt{\tau_{min}}$ with high probability. Consequently,

$$\|V_\tau - \mathscr{V}_\tau\| \lesssim \frac{\sqrt{\log n}}{\mu_{K,\tau}\sqrt{\tau_{min}}} \quad \text{with high probability.} \tag{29}$$

Using the fact that

$$\|V_{i,\tau} - \mathscr{V}_{i,\tau}\| \leq \|V_\tau - \mathscr{V}_\tau\|,$$

one can infer that the right side of (29) provides a bound on (28) as well. However, the above bound it too weak to make inferences about cluster recovery. Consequently, we strengthen these bounds by demonstrating that

$$\|V_{i,\tau} - \mathscr{V}_{i,\tau}\| \lesssim \frac{\sqrt{\log n}}{(\mu_{K,\tau}\sqrt{\tau_{min}})^2} \quad \text{with high probability.} \tag{30}$$

These improvements arise from the extension of the techniques in [3] to normalized Laplacians. We now briefly describe the technique.

For convenience, we remove the subscript τ from the various quantities. Let $M = diag(\lambda_1, \ldots, \lambda_K)$ denote the diagonal matrix of top K eigenvalues of $\mathscr{L}$, where $|\lambda_1| \geq \ldots \geq |\lambda_K| > 0$. Notice that $\mu_{k,\tau} = |\lambda_k|$. Further, let $\hat{M}$ be the corresponding matrix for L. We use the symbol Δ to denote the difference of a sample and population quantity. In particular, let $\Delta L = L - \mathscr{L}$ and $\Delta V = V - \mathscr{V}$. Using $LV = V\hat{M}$, note that

$$\begin{aligned} V - \mathscr{V} &= LV\hat{M}^{-1} - \mathscr{V} \\ &= (\mathscr{L} + \Delta L)(\mathscr{V} + \Delta V)\hat{M}^{-1} - \mathscr{V} \end{aligned}$$

Consequently, using $\mathscr{L}\mathscr{V} = \mathscr{V}M$, one has from the above

$$V - \mathscr{V} = \Delta L \mathscr{V} \hat{M}^{-1} + (\mathscr{L} + \Delta L)\Delta V \hat{M}^{-1} + \mathscr{V}(M - \hat{M})\hat{M}^{-1} \tag{31}$$

In Appendix A, in particular Lemma 14, we provide deterministic bounds on the ℓ_2-norm of the i-th row of each of three terms in the right side of (31). These bounds are applicable to the difference of top-K eigenvector matrices of any two Laplacian matrices. Assuming an SBM framework, Appendix B provides high probability bounds on each of the three terms in (31).

7.2 Proof Of Results In Section 4

We prove in Lemma 19, Appendix C, that with regularization parameter $\tau_n = \gamma_{K,n}$, the Laplacian matrix $\mathscr{L}_{\tau_n}$ is close to a rank $K+1$ Laplacian matrix $\tilde{\mathscr{L}}_{\tau_n}$ in spectral norm. Here $\tilde{\mathscr{L}}_\tau$ is the population regularized Laplacian of a $K+1$-block SBM constructed from clusters $C_1, \ldots, C_K$ and C_{K+1}, and block probability matrix

$$\tilde{B} = \begin{pmatrix} B_s & b_{sw}\mathbf{1} \\ b_{sw}\mathbf{1}'\mathbf{1}' & b_w \end{pmatrix},$$

where the $K \times K$ matrix B_s, and the quantities b_{sw}, b_w are as in (15). Since $\tilde{B}$ has rank $K+1$, the same holds also for $\tilde{\mathscr{L}}_\tau$. We denote by $\tilde{\mu}_{k,\tau}$, for $k = 1, \ldots, n$, to be the magnitude of the eigenvalues of $\tilde{\mathscr{L}}_\tau$ arranged in decreasing order. Notice that $\tilde{\mu}_{k,\tau} = 0$ for $k > K+1$. Explicit expressions for the non-zero eigenvalues of $\tilde{\mathscr{L}}_\tau$ are given in Lemma 21, Appendix C.4.

Consequently, from Lemma 19 one get that the eigenvalues of $\mathscr{L}_{\tau_n}$ are close to that of $\tilde{\mathscr{L}}_{\tau_n}$ via Weyl's inequality [5]. Lemma 9 follows from examining the eigenvalues of $\tilde{\mathscr{L}}_{\tau_n}$, as given in Lemma 21.

In the next subsection we provide the idea behind the proof of Theorem 10. The proof of Theorem 12, although slightly more involved, is similar in spirit. We leave its proof completely to Appendix C.3.

7.2.1 Proof Of Theorem 10

Let $\mathscr{V}_{\tau_n}$ be the $n \times (K+1)$ matrix corresponding to the first $K+1$ eigenvectors of the population Laplacian $\mathscr{L}_{\tau_n}$. Recall that V_{τ_n} is the sample version of $\mathscr{V}_{\tau_n}$.

Further, denote by $\tilde{\mathscr{V}}_{\tau_n}$ as the $n \times (K+1)$ matrix corresponding to the first $K+1$ eigenvectors of $\tilde{\mathscr{L}}_{\tau_n}$.

Since $\tilde{\mathscr{L}}_{\tau_n}$ is the population Laplacian of a $K+1$-block SBM, the matrix $\tilde{\mathscr{V}}_{\tau_n}$ has $K+1$ distinct rows, with the unique rows corresponding to the $K+1$ clusters. Take these distinct rows as $\mathsf{cent}_1, \ldots, \mathsf{cent}_{K+1}$. Then (22) follows from Lemma 1.

As mentioned in Subsection 2.2, since there are multiple choices of $\mathscr{V}_{\tau_n}$ and $\tilde{\mathscr{V}}_{\tau_n}$, we take $\mathscr{V}_\tau$ to be the eigenvector matrix of $\mathscr{L}_{\tau_n}$ that is closest to V_{τ_n} in spectral norm. With $\mathscr{V}_{\tau_n}$ so defined, we take $\tilde{\mathscr{V}}_{\tau_n}$ to be eigenvector matrix of $\tilde{\mathscr{L}}_{\tau_n}$ that is analogously closest to $\mathscr{V}_{\tau_n}$.

Theorem 10 follows from Theorem 13 below, which is proved in Appendix C.2. It demonstrates that for a particular choice of regularization parameter τ_n, not only is V_{i,τ_n} close to its population counterpart $\mathscr{V}_{i,\tau_n}$, but also $\mathscr{V}_{i,\tau_n}$ is close to $\tilde{\mathscr{V}}_{i,\tau_n}$. As before, the subscript i denotes the i-th row of these matrices.

Theorem 13. *For the regularization parameter $\tau_n = \gamma_{K,n}$, we have*

$$\|\mathscr{V}_{i,\tau_n} - \tilde{\mathscr{V}}_{i,\tau_n}\| \lesssim \frac{1}{\sqrt{\gamma_{K,n}}} \frac{\gamma_{K+1,n}}{\gamma_{K,n}} \tag{32}$$

Further, if $\log n/\gamma_{K,n} = o(1)$ then,

$$\max_i \|V_{i,\tau_n} - \mathscr{V}_{i,\tau_n}\| \lesssim \frac{\sqrt{\log n}}{\gamma_{K,n}} \tag{33}$$

with probability tending to one for large n.

The claim (32) uses the result that $\mathscr{L}_{\tau_n}$ is close to $\tilde{\mathscr{L}}_{\tau_n}$ for $\tau_n = \gamma_{K,n}$. This leads to the fact that the eigenvector matrices $\mathscr{V}_{\tau_n}$ and $\tilde{\mathscr{V}}_{\tau_n}$ are also close. The claim (33) follows from Theorem 17, which is a slightly more general version of Theorem 3. The improvements in the Davis-Kahan bounds are essential ingredients in the proofs of both (32) and (33).

Proof of Theorem 10. Recall that $\mathsf{cent}_1, \ldots, \mathsf{cent}_{K+1}$ are taken to be the $K+1$ distinct rows of $\tilde{\mathscr{V}}_{\tau_n}$, with cent_k corresponding to cluster C_k. The proof of (22) follows from Lemma 1.

Theorem 13 states that if the bounds (32) and (33) are small, then the rows of sample eigenvector matrix V_{τ_n} are concentrated near one of the $K+1$ distinct points representing the clusters $C_1, \ldots, C_{K+1}$. Correspondingly, (23) follows from using

$$\|V_{i,\tau_n} - \tilde{\mathscr{V}}_{i,\tau_n}\| \le \|V_{i,\tau_n} - \mathscr{V}_{i,\tau_n}\| + \|\mathscr{V}_{i,\tau_n} - \tilde{\mathscr{V}}_{i,\tau_n}\|.$$

□

8 Discussion

The paper is an attempt to provide a theoretical quantification for regularization in spectral clustering. Increasing the regularization parameter makes the sample Laplacian better concentrated around its corresponding population version. However, increasing the regularization parameter also changes the eigen gap of the population Laplacian. This was also noticed in [7], [24] for a different form of regularization. The larger this gap, the better is cluster recovery. Intuitively, this gap should be small for large τ as the population Laplacian becomes more like a constant matrix. Consequently, the best choice of regularization parameter is the one that balances these two competing effects. Sections 3 and 4 demonstrate two different ways in which regularization affects this gap. To the best of our knowledge, this is the first paper that incorporates both these effects in the quantification of regularization.

In Subsection 3.1, where the goal was to recover all the clusters, the regularization $\tau_n = \gamma_{K-1,n}$ was chosen since

$$\mu_{K,\tau} \asymp \mu_{K,0} \quad \text{for } \tau \lesssim \gamma_{K-1,n}$$

In other words, the eigen gap at τ_n is the same order of magnitude as that at $\tau = 0$. Consequently, the regularization parameter τ_n increases the performance of clustering since the sample Laplacian L_τ is better concentrated around its population version at $\tau = \tau_n$. Our proof technique also shows that for any alternative larger choice of regularization parameter, say τ_n', with $\tau_n = o(\tau_n')$, the eigen gap $\mu_{K,\tau_n'}$ goes to zero for larger n. For the two block model it can even be shown that such a choice would lead to worse bounds. Consequently, this also hints that for the regularization in [7], [24], the regularizer set as the average degree, as conjecture in [24], is not the best choice, especially when there is large variability in the degrees. However, when the degrees are more or less equal we believe that the average degree should work well since it is close to $\gamma_{K-1,n}$.

More importantly, in Section 4 we show that regularization can help in situations where not all nodes belong to well defined clusters. In such situations, the improvements via regularization are due to two reasons. The first, as mentioned above, is due to better stability of sample Laplacian around its corresponding population counterpart. The second, as demonstrated in Lemma 9, is through the creation of a gap between the top few eigenvalues and the remaining. In this regard, we considered two different regimes depending on the size of the set of nodes belonging to the weak clusters. We also demonstrate in Subsection 4.1 and 4.2, that the top few population eigenvectors are able to distinguish between the nodes of the strong clusters with regularization. This need not be the case without regularization, as illustrated in Figure 3. We also demonstrate this on the political blogs data set.

As remarked in Section 4, if the rank of B, given by (15), is K then the model encompasses specific degree-corrected stochastic block models (D-SBM) [16]. In particular, consider a K-block D-SBM, with degree weight parameter for node i to be θ_i, where $0 < \theta_i \leq 1$. Assume that $\theta_i = 1$ for a large number

of nodes. Take the nodes in the strong clusters to be those with $\theta_i = 1$. The nodes in the strong clusters are associated to one of K clusters depending on the cluster they belong to in the D-SBM. The remaining nodes are taken to be in the weak clusters. Assumptions (16), (17) and (18) puts constraints on the θ_i's which allows one to distinguish between the strong clusters via regularization.

From the results of Section 4, a high value of regularization parameter reduces the influence of the less well defined clusters. We conjecture that these results also extend to situations where there is a hierarchy of clusters. In this case, the less well defined clusters would correspond to those lower down in the hierarchy. An natural way of going about this is to cluster in a hierarchical fashion, using larger values of τ for clusters higher up in the hierarchy. In this regard, there seems to be parallels between this approach and the clustering of points using the level-set approach [13]. We hope to investigate this in a future work.

We provide a data dependent technique for choosing the regularization parameter τ, and compare it the scheme that uses the Girvan-Newman modularity. Since the *DK-est* technique compares the perturbation of the sample Laplacian to the population Laplacian of a stochastic block model, chosen based on the selected clusters, the procedure is similar in spirit to modularity based methods such as Girvan-Newman modularity. From our simulations, our method is seen to perform better than the Girvan-Newman scheme. For the application to the political blogs data set our scheme performs well. However, the scheme that uses the Girvan-Newman modularity outperforms our scheme, most likely due to the large variance in the degrees of the nodes for this dataset. We believe that one can obtain a degree-corrected version of our scheme which performs better in such situations. We leave this for a future work.

Acknowledgments

This work was partly supported by ARO grant W911NF-11-1-0114. We would like to thank Sivaraman Balakrishnan for some very helpful discussions regarding strengthening of the Davis-Kahan bounds. A. Joseph would also like to thank Purnamrita Sarkar for very helpful discussions regarding the results in Section 4, and also Arash A. Amini for sharing the code used in the work [2].

A Bounds On Eigenvectors Differences

In this section we provide deterministic bounds on the difference of eigenvectors from two arbitrary Laplacians matrices. This will be used to provide high probability bounds on the perturbation of eigenvectors. In particular, consider any two Laplacian matrices,

$$L = D^{-1/2} A D^{-1/2}$$

$$\mathscr{L} = \mathscr{D}^{-1/2} P \mathscr{D}^{-1/2},$$

where A, D are the adjacency matrices and the diagonal matrix of degrees corresponding to L. By adjacency matrix we refer to any symmetric matrix with entries in $[0, 1]$. Similarly P, $\mathscr{D}$ are the analogous quantities for $\mathscr{L}$.

Denote,

$$\mathscr{D} = diag(d_1, \ldots, d_n) \quad \text{and} \quad d_{min} = \min_i d_i.$$

Let V, $\mathscr{V}$ be the $n \times K'$ eigenvector matrices corresponding to top (in absolute value) K' eigenvalues of L, $\mathscr{L}$ respectively. As mentioned in Subsection 2.2, since there are multiple choices for V and $\mathscr{V}$, for a given choice of V, we take $\mathscr{V}$ as the eigenvector matrix of $\mathscr{L}$ that is closest to V in spectral norm. We also denote by $\rho_{1,K'}$ and $\rho_{2,K'}$ the K'-th smallest eigenvalue (in magnitude) of L and $\mathscr{L}$ respectively.

Denote the i-th row of the above eigenvector matrices as V_i and $\mathscr{V}_i$. In this section we provide deterministic bounds on $\|\Delta V_i\|$, where

$$\Delta V_i = V_i - \mathscr{V}_i.$$

In general, we use the symbol Δ to denote the difference of two quantities. For example $\Delta A = A - P$ and $\Delta V = V - \mathscr{V}$. We also use the the subscript $i*$ to denote the i-th row of a matrix. The following inequality will be use frequently in the analysis: For matrices H_1, H_2

$$\|H_1 H_2\| \leq \|H_1\| \|H_2\| \tag{34}$$

The lemma below provides deterministic bounds on the ℓ_2 norm of each row of ΔV, given by ΔV_i. Our proof technique involves generalization of perturbation bounds obtained in [3] for eigenvectors of the unnormalized Laplacian, to that of the normalized Laplacians.

Lemma 14. *The following bound holds:*

$$\|\Delta V_i\| \leq \frac{\mathsf{b}_{1,i}}{\rho_{1,K'}} + \frac{\mathsf{b}_{2,i}\|\Delta V\|}{\rho_{1,K'}} + \frac{\mathsf{b}_{3,i}}{\rho_{1,K'}}, \tag{35}$$

where,

$$\begin{aligned}
\mathsf{b}_{1,i} &= R_i \, \|\mathscr{L}_{i*}\| \, \|\Delta R \mathscr{V}\| + |\Delta R_i| \, \|\mathscr{V}_i\| \\
&\quad + \left(R_i \, \|\Delta A_{i*}\| \, \|\Delta R \mathscr{V}\| + R_i \left\| \Delta A_{i*} \mathscr{D}^{-1/2} \mathscr{V} \right\| \right) / \sqrt{d_{min} d_i} \qquad (36) \\
\mathsf{b}_{2,i} &= \|\mathscr{L}_{i*}\| \, (1 + R_i \, \|\Delta R\| + |\Delta R_i|) + \frac{\|R\| \, \|\Delta A_{i*}\|}{\sqrt{d_{min} d_i}} \qquad (37) \\
\mathsf{b}_{3,i} &= \|\mathscr{V}_i\| \|\Delta L\|. \qquad (38)
\end{aligned}$$

Here $R = diag(R_1, \ldots, R_n)$ is $(\mathscr{D}/D)^{1/2}$. Further, $\Delta R = diag(\Delta R_1, \ldots, \Delta R_n)$, where $\Delta R = R - I$.

Proof. Let $M_1 = diag(\lambda_1, \ldots, \lambda_{K'})$ denote the diagonal matrix of eigenvectors of L, where $|\lambda_1| \geq \ldots \geq |\lambda_{K'}|$. Notice that $\rho_{1,K'} = \lambda_{K'}$. Let M_2 be the analogous diagonal matrix of eigenvalues of $\mathscr{L}$. Now,

$$\begin{aligned} \Delta V &= LVM_1^{-1} - \mathscr{V} \\ &= L(\mathscr{V} + \Delta V)M_1^{-1} - \mathscr{V} \end{aligned} \tag{39}$$

The first equality follows from noticing that $LV = VM_1$, since the columns of V are eigen vectors

Correspondingly, from (39) one gets that ΔV is the sum of three terms given by,

$$\Delta V = [\mathscr{L} + \Delta L]\mathscr{V}M_1^{-1} + [\mathscr{L} + \Delta L]\Delta V M_1^{-1} - \mathscr{V} \tag{40}$$

Notice that $\mathscr{L}\mathscr{V} = \mathscr{V}M_2$. Consequently, from (40) one gets the ΔV_i is the sum of three $1 \times K'$ row vectors J_1, J_2, J_3, where

$$\Delta V_i = \underbrace{\Delta L_{i*}\mathscr{V}M_1^{-1}}_{J_1} + \underbrace{[\mathscr{L}_{i*} + \Delta L_{i*}]\Delta V M_1^{-1}}_{J_2} + \underbrace{\mathscr{V}_i(M_2 - M_1)M_1^{-1}}_{J_3} \tag{41}$$

Below, we provide bounds on the ℓ_2 norm each of the above three separately. These would correspond to the three terms appearing in the right side of (35). Before this, we describe how we handle the ΔL_{i*} term appearing in (41).

Notice that,

$$\begin{aligned} \Delta L &= R\mathscr{D}^{-1/2}A\mathscr{D}^{-1/2}R - \mathscr{L} \\ &= \underbrace{R\mathscr{L}R - \mathscr{L}}_{\Delta L^1} + \underbrace{R\mathscr{D}^{-1/2}\Delta A\,\mathscr{D}^{-1/2}R}_{\Delta L^2} \end{aligned} \tag{42}$$

By subtracting and adding $R\mathscr{L}$, write $\Delta L^1 = \Delta L^{11} + \Delta L^{12}$, where

$$\Delta L^{11} = R\mathscr{L}\Delta R \tag{43}$$

$$\Delta L^{12} = \Delta R\mathscr{L} \tag{44}$$

Similarly $\Delta L^2 = \Delta L^{21} + \Delta L^{22}$, where

$$\begin{aligned} \Delta L^{21} &= R\mathscr{D}^{-1/2}\Delta A\,\mathscr{D}^{-1/2}\Delta R \\ \Delta L^{22} &= R\mathscr{D}^{-1/2}\Delta A\,\mathscr{D}^{-1/2} \end{aligned}$$

We now bound the ℓ_2-norm of the three terms in (41). We first bound the ℓ_2-norm of J_1. Notice that,

$$\|\Delta L_{i*}^{11}\mathscr{V}M_1^{-1}\| \leq R_i\|\mathscr{L}_{i*}\|\|\Delta R\mathscr{V}\|\|M_1^{-1}\|$$

and

$$\|\Delta L_{i*}^{12}\mathscr{V}M_1^{-1}\| \leq |\Delta R_i|\|\mathscr{V}_i\|\|M_2M_1^{-1}\|,$$

where for the above we use that $\mathscr{L}\mathscr{V} = \mathscr{V}M_2$. Similarly,

$$\|\Delta L_{i*}^{21}\mathscr{V}M_1^{-1}\| \leq R_i d_i^{-1/2}d_{min}^{-1/2}\|\Delta A_{i*}\|\|\Delta R\mathscr{V}\|\|M_1^{-1}\|$$

and

$$\|\Delta L_{i*}^{22}\mathscr{V}M_1^{-1}\| \leq R_i d_i^{-1/2} d_{min}^{-1/2} \|\Delta A_{i*}\mathscr{D}^{-1/2}\mathscr{V}\| \|M_1^{-1}\|$$

The expression for $\mathsf{b}_{1,i}$ results from using $\|M_1^{-1}\| = 1/\rho_{1,K'}$ and $\|M_2 M_1^{-1}\| \leq 1/\rho_{1,K'}$.

Similarly for ℓ_2 norm of J_2, notice that it is bounded by

$$\left(\|\mathscr{L}_{i*}\| + \|\Delta L_{i*}\|\right) \|\Delta V\| \|M_1^{-1}\|$$

We need to bound $\|\Delta L_{i*}\|$. From the above one sees that,

$$\begin{aligned}
\|\Delta L_{i*}^{11}\| &\leq R_i \|\mathscr{L}_{i*}\| \, \|\Delta R\| \\
\|\Delta L_{i*}^{12}\| &\leq |\Delta R_i| \, \|\mathscr{L}_{i*}\| \\
\|\Delta L_{i*}^{2}\| &\leq \|R^2\| \, \|\Delta A_{i*}\| / \sqrt{d_i d_{min}}
\end{aligned}$$

The claim for the ℓ_2 norm of J_3 follows from using the fact that

$$\|M_2 - M_1\| \leq \|\Delta L\|,$$

which follows from Weyl's inequality [5]. □

In equation (35), one needs to provide an upper bound on $\|\Delta V\|$ in (35). This is achieved using the Davis-Kahan theorem (see, for example, [5]), a version of which we state below. Recall that $\rho_{2,K'}$ denoted the K'-th smallest eigenvalue of $\mathscr{L}$. Similarly, denote as $\rho_{2,K'+1}$ the $K'+1$ smallest eigenvalue (in magnitude) of $\mathscr{L}$. We remark that if $\mathscr{L}$ corresponds to the population Laplacian of K' block SBM, then $\rho_{2,K'+1} = 0$. However, here we made no such assumption on $\mathscr{L}$.

Theorem 15 (Davis-Kahan with spectral norm)**.** *Let* $\|\Delta L\| < (\rho_{2,K'} - \rho_{2,K'+1})/2$. *Then*

$$\|\Delta V\| \leq 2 \frac{\|\Delta L\|}{\rho_{2,K'} - \rho_{2,K'+1}}. \tag{45}$$

The following is an immediate corollary of the above and Lemma 14.

Corollary 16. *If* $\|\Delta L\| \leq (\rho_{2,K'} - \rho_{2,K'+1})/2$ *then*

$$\|\Delta V_i\| \leq \frac{2\mathsf{b}_{1,i}}{\rho_{2,K'}} + \frac{4\mathsf{b}_{2,i}\|\Delta L\|}{\rho_{2,K'}(\rho_{2,K'} - \rho_{2,K'+1})} + \frac{2\mathsf{b}_{3,i}}{\rho_{2,K'}}. \tag{46}$$

Here $\mathsf{b}_{1,i}$, $\mathsf{b}_{2,i}$ *and* $\mathsf{b}_{3,i}$ *are as in Lemma 14.*

Proof. Notice that $|\rho_{2,K'} - \rho_{1,K'}| \leq \|\Delta L\|$ using Weyl's inequality (see for example [5]). Consequently, $\rho_{1,K'} \geq \rho_{2,K'}/2$ since $\|\Delta L\| \leq \rho_{2,K'}/2$. The proof is completed using Theorem 15. □

B Proof of Theorem 3

We apply the results of Appendix A, in particular Corollary 16, to prove a more general version of Theorem 3, which we describe below. This will be required in the proof of Theorem 13.

Consider a K block SBM as in Section 3. In Theorem 17 we use Corollary 16 to bound the perturbation of the first K' eigenvectors of the sample Laplacian matrix. Taking $K' = K$ leads to Theorem 3.

In the theorem below we use the notation Appendix A with $L = L_\tau$ and $\mathscr{L} = \mathscr{L}_\tau$. We take V, $\mathscr{V}$ in Appendix A to be the eigenvector matrices corresponding to the first K' eigenvectors, where $K' \leq K$, of L, $\mathscr{L}$ respectively. In this case $\rho_{2,K'} = \mu_{K',\tau}$ and $\rho_{2,K'+1} = \mu_{K'+1,\tau}$. We also require the following more general version of Assumption 2, which given the size of the eigen gap as a function of τ.

Assumption 3 (Eigen gap)**.**

$$\rho_{2,K'} - \rho_{2,K'+1} > 20\frac{\sqrt{\log n}}{\sqrt{\tau_{min}}}$$

Notice that with $K' = K$, Assumption 3 is the same as Assumption 2 since $\rho_{2,K'+1} = 0$ for the K block SBM. Then we have the following:

Theorem 17. *Let Assumptions 1 and 3 hold. Then, with probability at least* $1 - (2K' + 5)/n$*,*

$$\max_i \|V_i - \mathscr{V}_i\| \leq \frac{\tilde{\delta}_{\tau,n}}{\rho_{2,K'}(\rho_{2,K'} - \rho_{2,K'+1})} \tag{47}$$

where

$$\tilde{\delta}_{\tau,n} = 293\frac{\sqrt{\log n}}{\tau_{min}} + 31\frac{\sqrt{\log n}}{\sqrt{\tau_{min}}}\|\mathscr{V}_i\| + 12\frac{K'\sqrt{\log n}}{\tau_{min}^{3/2}}. \tag{48}$$

Proof of Theorem 3. Take $K' = K$ in Theorem 17. As mentioned before, in this case Assumption 3 is the same as Assumption 2 since $\rho_{2,K'+1} = \mu_{K+1,\tau} = 0$. Further, $\|\mathscr{V}_i\| = 1/\sqrt{|C_{k(i)}|}$ from Lemma 22. □

We now prove Theorem 17. Notice that $D = diag(\hat{d}_{1,\tau}, \ldots, \hat{d}_{n,\tau})$ and $\mathscr{D} = diag(d_{1,\tau}, \ldots, d_{n,\tau})$. Let $\tau_{min} = \max(\tau, d_{min})$ satisfy the condition in Assumption 1. In other words, recall that

$$\tau_{min} = \kappa_n \log n,$$

where $\kappa_n > c = 32$. Further, let

$$\tau_i = \max\{\tau, d_i\}. \tag{49}$$

We prove Theorem 17 by appealing to Corollary 16. To do this, the following deterministic as well as high probability bounds that are derived in the Subsections D.1 and D.2 in Appendix D would prove to be useful.

Deterministic bounds: With $\mathscr{L} = \mathscr{L}_\tau$, one has

$$\|P_{i*}\| \le \sqrt{d_{i,\tau}} \quad \text{and} \quad \|\mathscr{L}_{i*}\| \le 1/\sqrt{\tau_{min}}.$$

High Probability bounds: We assume that c_2 is a positive number satisfying

$$c_2/\sqrt{c} < 1.$$

Let

$$c_1 = .5c_2^2/(1 + c_2/\sqrt{c}). \quad \text{and} \quad c_3 = c_2/\sqrt{1 - c_2/\sqrt{c}}$$

From Subsection D.2, the following holds with probability at least $1 - (2K' + 3)/n^{c_1 - 1}$: For each $i = 1, \ldots, n$,

1.
$$|\Delta R_i| \le c_3 \frac{\sqrt{\log n}}{\sqrt{\tau_i}} \tag{50}$$

2.
$$\|\Delta A_{i*} \mathscr{D}^{-1/2} \mathscr{V}\| \le c_2 K' \sqrt{\frac{\log n}{\tau_{min}}} \tag{51}$$

3.
$$\|\Delta A_{i*}\|^2 \le d_{i,\tau} + c_2 \sqrt{\tau_i \log n} \tag{52}$$

Notice that (50) implies that $R_i \le 1 + c_3/\sqrt{c}$, using $c \log n \le \tau_i$. Further, (52) implies that

$$\|\Delta A_{i*}\|^2 \le (1 + c_2/\sqrt{c}) d_{i,\tau},$$

using τ_i, as well as $c \log n$, are at most $d_{i,\tau}$.

For convenience we take $c = 32$, $c_2 = 2\sqrt{2}$. Then one has $c_1 > 2$, $c_3 = 4\sqrt{2}$. The following lemma shows that the condition of Corollary 16 is satisfied with high probability.

Lemma 18. *Let Assumption 1 be satisfied and let c, c_2 be chosen as above. Then with probability at least $1 - 2/n^{c_1 - 1}$, the following hold*

$$\|\Delta L\| \le \kappa \sqrt{\frac{\log n}{\tau_{min}}}$$

Here $\kappa = \sqrt{2c_1} + c_2(2 + c_2/\sqrt{c}) < 10$.

Consequently, if Assumption 3 also holds then the condition in Corollary 16 is satisfied with probability at least $1 - 2/n^{c_1 - 1}$.

The above is proved in Appendix D.3. Using the above deterministic and high probability bounds, one gets that with probability at least $1 - (2K + 3)/n^{c_1 - 1}$,

$$\mathsf{b}_{1,i} \le \omega_{11} \frac{\sqrt{\log n}}{\tau_{min}} + \omega_{12} \frac{\sqrt{\log n}}{\sqrt{\tau_{min}}} \|\mathscr{V}_i\| + \omega_{13} \frac{K' \sqrt{\log n}}{\tau_{min}^{3/2}}$$

$$\mathsf{b}_{2,i} \leq \frac{\omega_2}{\sqrt{\tau_{min}}}$$

$$\mathsf{b}_{3,i} \leq \kappa \|\mathscr{V}_i\| \frac{\sqrt{\log n}}{\sqrt{\tau_{min}}},$$

where

$$\omega_{11} = c_3\left(1 + c_3/\sqrt{c}\right)\left(1 + \sqrt{1 + c_2/\sqrt{c}}\right), \quad \omega_{12} = c_3, \quad \omega_{13} = \left(1 + c_3/\sqrt{c}\right) c_2$$

$$\omega_2 = 1 + \frac{c_3}{\sqrt{c}}\left(2 + c_3/\sqrt{c}\right) + \left(1 + c_3/\sqrt{c}\right)\sqrt{1 + c_2/\sqrt{c}}.$$

Using the values of c, c_2, c_3 given before, one gets $\omega_{11} < 26$, $\omega_{12} < 6$, $\omega_{13} < 6$, $\omega_2 < 7$ and $\kappa < 1/10$. Substituting these in bound (46), and using $\rho_{2,K'} - \rho_{2,K'+1} \leq \rho_{2,K'}$, gives the desired expression.

C Proof of Results in Section 4

The following lemma demonstrates that the population Laplacian matrices $\mathscr{L}_{\tau_n}$ and $\tilde{\mathscr{L}}_{\tau_n}$ are close with $\tau_n = \gamma_{K,n}$. Here $\tilde{\mathscr{L}}_{\tau_n}$ is as in Subsection 7.2. In the lemma below, we take $L = \mathscr{L}_{\tau_n}$ and $\mathscr{L} = \tilde{\mathscr{L}}_{\tau_n}$, where $L = D^{-1/2}AD^{-1/2}$ and $\mathscr{L} = \mathscr{D}^{-1/2}P\mathscr{D}^{-1/2}$. The quantities ΔA and ΔR are as in Appendix A.

Lemma 19. *Let $L = \mathscr{L}_{\tau_n}$ and $\mathscr{L} = \tilde{\mathscr{L}}_{\tau_n}$ so that $\Delta L = \mathscr{L}_{\tau_n} - \tilde{\mathscr{L}}_{\tau_n}$. Then, with $\tau_n = \gamma_{K,n}$, the following bounds hold*

1. $$|\Delta R_i| \lesssim \frac{\gamma_{K+1,n}}{\gamma_{K,n}}$$

2. $$\|\Delta A_{i*}\| \lesssim \sqrt{\gamma_{K+1,n}} \quad \textit{and} \quad \|\mathscr{L}_{i*}\| \lesssim 1/\sqrt{\gamma_{K,n}}$$

3. $$\|\Delta L\| \lesssim \frac{\gamma_{K+1,n}}{\gamma_{K,n}}$$

Proof. Recall $\tilde{\mu}_{k,\tau_n}$ for $k = 1, \ldots, n$, are the magnitude of the eigenvalues of $\mathscr{L} = \tilde{\mathscr{L}}_{\tau_n}$ arranged in decreasing order. We write the $D, \mathscr{D}$ in Appendix A as $diag(d_{1,\tau_n}, \ldots, d_{n,\tau_n})$ and $diag(\tilde{d}_{1,\tau_n}, \ldots, \tilde{d}_{n,\tau_n})$ respectively. Correspondingly,

$$\begin{aligned} |\Delta R_i| &\leq \frac{|\tilde{d}_{i,\tau_n} - d_{i,\tau_n}|}{d_{i,\tau_n}^{1/2}(\tilde{d}_{i,\tau_n}^{1/2} + d_{i,\tau_n}^{1/2})} \\ &\lesssim \frac{\gamma_{K+1,n}}{\gamma_{K,n}} \end{aligned}$$

The last relation follows since $|\tilde{d}_{i,\tau_n} - d_{i,\tau_n}| \lesssim \gamma_{K+1,n}$ using (17) and $\tilde{d}_{i,\tau_n}$, $d_{i,\tau_n} \geq \gamma_{K,n}$ as $\tau_n = \gamma_{K,n}$. This proves 1.

Further,

$$\|\Delta A_{i*}\| = O\left(\sqrt{\gamma_{K+1,n}}\right)$$
$$\|\mathscr{L}_{i*}\| \leq 1/\sqrt{\tilde{d}_{i,\tau_n}} = O\left(1/\sqrt{\gamma_{K,n}}\right)$$

The first relation follows from using the $\|\Delta A_{i*}\|^2 \leq (n-|C_{K+1}|)b_{sw}^2+|C_{K+1}|b_w^2$, the right side of which is at most $\gamma_{K+1,n}$. The second relation in the above follows from using the same argument as in Lemma 23. This proves 2.

We need to bound $\|\Delta L\|$. One sees that

$$\|\Delta L\| \leq \|\Delta L^{11}\| + \|\Delta L^{12}\| + \|\Delta L^2\|,$$

where the matrices on the right are defined in (42) - (44). Consequently, using (34), one gets that

$$\|\Delta L\| \lesssim \frac{\gamma_{K+1,n}}{\gamma_{K,n}}.$$

In bounding $\|\Delta L^2\|$, where ΔL^2 as in (42), we use that $\|\Delta A\| \leq \gamma_{K+1,n}$, since $-\Delta A$ corresponds to an adjacency matrix with degree at most $\gamma_{K+1,n}$. □

C.1 Proof of Lemma 9

We first prove Claim 1. We first show that the non-zero eigenvalues of the $\tilde{\mathscr{L}}_{\tau_n}$ are bounded away from zero. Since $\mathscr{L}_{\tau_n}$ is close to $\tilde{\mathscr{L}}_{\tau_n}$ from Lemma 19, this will lead to the claim regarding the eigenvalues of $\mathscr{L}_{\tau_n}$.

From Appendix C.4, the non-zero and non-unitary eigenvalues of $\tilde{\mathscr{L}}_{\tau_n}$ are given by

$$\lambda_1 = \frac{|C_K|}{\gamma_{K,n}+\tau_n}(p_{K,n}-q_{s,n})$$
$$\lambda_2 = \frac{|C_{K+1}|(b_w+\tau_n/n)}{\gamma_{K+1}+\tau_n} - \frac{|C_{K+1}|(b_{sw}+\tau_n/n)}{\gamma_{K,n}+\tau_n}$$

The eigenvalue λ_1 has multiplicity $K-1$. Notice that the numerator of λ_1 above is $(1-\kappa)|C_K|p_{K,n}$, where κ is as in (18). Further, $\gamma_{K,n} \asymp |C_K|p_{K,n}$, using $|C_{K+1}| \asymp n$, $b_{s,w} \leq b_w$, and $\gamma_{K+1,n} = o(\gamma_{K,n})$. Thus the numerator of λ_1 is $\asymp \gamma_{K,n}$. Consequently, with $\tau_n = \gamma_{K,n}$, the eigenvalue λ_1 is bounded away from zero.

Next, we show that λ_2 is bounded away from zero. With $b_{sw} \leq b_w$ one has,

$$\lambda_2 \geq |C_{K+1}|(b_w+\tau_n/n)\left(\frac{1}{\gamma_{K+1,n}+\tau_n} - \frac{1}{\gamma_{K,n}+\tau_n}\right)$$
$$\asymp |C_{K+1}|(b_w+\tau_n/n)/\gamma_{K,n}.$$

The last inequality follows from using $\tau_n = \gamma_{K,n}$ and $\gamma_{K+1,n} = o(\gamma_{K,n})$. Further, using $|C_{K+1}|\tau_n/n \asymp \gamma_{K,n}$, one gets that λ_2 is also bounded away from zero.

We need to show that μ_{K+1,τ_n} is bounded away from zero, while μ_{K+2,τ_n} goes to zero. To see this, notice that

$$\begin{aligned}\mu_{K+1,\tau_n} &\geq \tilde{\mu}_{K+1,\tau_n} - |\mu_{K+1,\tau_n} - \tilde{\mu}_{K+1,\tau_n}| \\ &\geq \tilde{\mu}_{K+1,\tau_n} - \|\Delta L\|,\end{aligned}$$

The second inequality follows from Weyl's inequality [5]. Thus, as $\|\Delta L\| = o(1)$ from Lemma 19, one gets that μ_{K+1,τ_n} is bounded away from zero. Similarly,

$$\mu_{K+2,\tau_n} \leq \tilde{\mu}_{K+2,\tau_n} + \|\Delta L\|.$$

The right side of the above is at most $\|\Delta L\|$ since $\tilde{\mu}_{K+2,\tau_n} = 0$, as $\tilde{\mathscr{L}}_{\tau_n}$ has rank $K+1$. Thus μ_{K+2,τ_n} goes to zero as $\|\Delta L\| = o(1)$. This proves Claim 1.

Next, we prove Claim 2. Notice that λ_1 is bounded away from zero even if $|C_{K+1}| = o(n)$. We show that λ_2 goes to zero if $C_{K+1} = o(n)$. To see this notice that

$$\begin{aligned}\lambda_2 &\leq \frac{|C_{K+1}|(b_w + \tau_n/n)}{\gamma_K + \tau_n} \\ &\leq \frac{(\gamma_{K+1,n} + |C_{K+1}|\gamma_{K,n}/n)}{\gamma_{K,n}}\end{aligned}$$

The right side goes to zero since $\gamma_{K+1,n}/\gamma_{1,n} = o(1)$ and $|C_{K+1}|/n = o(1)$. Thus, as above, using $\|\Delta L\| = o(1)$ and Weyl's inequality, one proves Claim 2.

C.2 Proof of Theorem 13

We first prove (32). We use Corollary 16 with $L = \mathscr{L}_{\tau_n}$ and $\mathscr{L} = \tilde{\mathscr{L}}_{\tau_n}$. Further, take $K' = K+1$ and $\Delta V = V - \mathscr{V}$, where $V = \mathscr{V}_{\tau_n}$ and $\mathscr{V} = \tilde{\mathscr{V}}_{\tau_n}$. Notice that with the above choice of $\mathscr{V}$, one has

$$\|\mathscr{V}_i\| \lesssim 1/\sqrt{n},$$

using Lemma 22 and since the clusters have sizes $\asymp n$. Using the bounds in Lemma 19, along with inequality (34), one gets

$$\begin{aligned}\mathsf{b}_{1,i} &\lesssim \frac{1}{\sqrt{\gamma_{K,n}}}\frac{\gamma_{K+1,n}}{\gamma_{K,n}} \\ \mathsf{b}_{2,i} &= \frac{1}{\sqrt{\gamma_{K,n}}}.\end{aligned}$$

To bound $\mathsf{b}_{3,i}$ we use the bound of $\|\Delta L\|$ given above.

Further, from the proof of Lemma 9, one gets that $\rho_{2,K'} = \tilde{\mu}_{K+1,\tau_n}$ is bounded away from zero with the choice of τ_n. This, combined with the fact that $\rho_{2,K'+1} = 0$ and $\|\Delta L\| = o(1)$, implies that the condition in Corollary 16 is satisfied for large n. This proves (32).

We now prove (33). To do this we apply Theorem 17 with $L = L_{\tau_n}$ and $\mathscr{L} = \mathscr{L}_{\tau_n}$. Further, take $K' = K+1$.We first need to show that Assumption

3 is satisfied for large n. Notice that $\rho_{2,K'} = \mu_{K+1,\tau_n}$ and $\rho_{2,K'+1} = \mu_{K+1,\tau_n}$. Consequently, as $\log(n)/\gamma_{K+1,n} = o(1)$, and $\rho_{2,K'+1} - \rho_{2,K'}$ bounded away from zero from Lemma 9, one gets that Assumption 3 is satisfied for large n. Consequently, Theorem 17 can be applied.

Further, note that from (32)

$$\begin{aligned}\|\mathscr{V}_{i,\tau_n}\| &\leq \|\tilde{\mathscr{V}}_{i,\tau_n}\| + +O\left(\frac{\gamma_{K+1,n}}{\gamma_{K,n}^{3/2}}\right) \\ &= \frac{1}{\sqrt{|C_{k(i)}|}} + O\left(\frac{\gamma_{K+1,n}}{\gamma_{K,n}^{3/2}}\right) \\ &\lesssim \max\left\{\frac{1}{\sqrt{n}}, \frac{\gamma_{K+1,n}}{\gamma_{K,n}^{3/2}}\right\}\end{aligned}$$

Here the second statement follows from Lemma 22, while the third statement follows since $|C_{k(i)}| \asymp n$. The proof is completed by noting that $\tau_{min} \geq \tau_n = \gamma_{K,n}$.

C.3 Proof of Theorem 12

The proof of Theorem 12 is quite similar to that of Theorem 10. We provide here the key steps in the proof. Analogous to Subsection 7.2.1, let U_{τ_n}, $\mathscr{U}_{\tau_n}$ be the $n \times K$ matrices corresponding to the first K eigenvectors of L_{τ_n}, $\mathscr{L}_{\tau_n}$ respectively. Similarly, let $\tilde{\mathscr{U}}_{\tau_n}$ be top K eigenvectors of $\tilde{\mathscr{L}}_{\tau_n}$.

Since there may be ambiguity in the choice of U_{τ}, $\mathscr{U}_{\tau}$ and $\tilde{\mathscr{U}}_{\tau}$, this situation is dealt with as in Subsection 7.2.1 for V_τ, $\mathscr{V}_\tau$ and $\tilde{\mathscr{V}}_\tau$. Theorem 20 below, gives the analogue of Theorem 13.

Theorem 20. *For the regularization parameter $\tau_n = \gamma_{K,n}$, we have*

$$\max_{i \notin C_{K+1}} \|\mathscr{U}_{i,\tau_n} - \tilde{\mathscr{U}}_{i,\tau_n}\| \lesssim \frac{1}{\sqrt{\gamma_{K,n}}} \frac{\gamma_{K+1,n}}{\gamma_{K,n}} \tag{53}$$

Further, if $\log n/\gamma_{K,n} = o(1)$ then,

$$\max_{i \notin C_{K+1}} \|U_{i,\tau_n} - \mathscr{U}_{i,\tau_n}\| \lesssim \frac{\sqrt{\log n}}{\gamma_{K,n}} \tag{54}$$

with probability tending to one for large n.

Proof of Theorem 12. As $\tilde{\mathscr{U}}_{\tau_n}$ are the top K eigenvectors of a $K+1$ SBM, there are K distinct values of $\tilde{\mathscr{U}}_{i,\tau_n}$ for $i \notin C_{K+1}$. Denote these by $\mathsf{cent}_1^K, \ldots, \mathsf{cent}_K^K$. These correspond to each of the K strong clusters. Further, from Lemma 21 one has for $i, i' \notin C_{K+1}$

$$\|\tilde{\mathscr{U}}_{i,\tau} - \tilde{\mathscr{U}}_{i',\tau}\| = \|\tilde{\mathscr{V}}_{i,\tau} - \tilde{\mathscr{V}}_{i',\tau}\|.$$

This follows from the claim that the eigenvector corresponding to λ_2 is constant for $i \notin C_{K+1}$. Thus (24) follows from applying Lemma 1 for a $K+1$ block SBM. Further, (25) is proved by noting that

$$\|U_{i,\tau_n} - \tilde{\mathscr{U}}_{i,\tau_n}\| \leq \|U_{i,\tau_n} - \mathscr{U}_{i,\tau_n}\| + \|\mathscr{U}_{i,\tau_n} - \tilde{\mathscr{U}}_{i,\tau_n}\|.$$

□

Proof of Theorem 20. This proof is similar to the of Theorem 13. Note, as with the proof Theorem 13, we use Corollary 16 for proving (53). In particular, we take $K' = K$, $L = \mathscr{L}_{\tau_n}$ and $\mathscr{L} = \tilde{\mathscr{L}}_{\tau_n}$. As before, from the proof of Lemma 9 the K-th eigenvalue of $\tilde{\mathscr{L}}_{\tau_n}$ is bounded away from zero, while its $K+1$-th smallest eigenvalue goes to zero. Consequently, using $\|\Delta L\| = o(1)$, the condition in Corollary 16 is satisfied for large n.

The bounds for $\mathsf{b}_{1,i}$, $\mathsf{b}_{2,i}$ and $\mathsf{b}_{3,i}$ are as in the proof of Theorem 13. Bound (53) follows from noting that for $i \notin C_{K+1}$, one has

$$\|\tilde{\mathscr{U}}_{i,\tau_n}\| = O(1/\sqrt{n})$$

as $\|\tilde{\mathscr{U}}_{i,\tau_n}\| \leq \|\tilde{\mathscr{V}}_{i,\tau_n}\| = 1/\sqrt{|C_{k(i)}|}$.

The claim (54), as with (33), follows from an application of Theorem 17, with $K' = K$. □

C.4 Eigen Analysis Of A $K+1$ block SBM

We investigate the eigenvalues of the $K+1$ community stochastic block model with block probability matrix

$$\tilde{B} = \begin{pmatrix} B_s & b_{sw}\mathbf{1} \\ b_{sw}\mathbf{1}'\mathbf{1}' & b_w \end{pmatrix}$$

As in the paper, the community assignment is given by the set $C_1, \ldots, C_{K+1}$. Denote the corresponding population Laplacian by $\tilde{\mathscr{L}}$. Then we have the following,

Lemma 21. *If $|C_1| = |C_1| = \ldots = |C_K|$, then the non-zero eigenvalues of $\tilde{\mathscr{L}}$ are 1, λ_1 and λ_2, where*

$$\lambda_1 = \frac{|C_K|}{\gamma_K}(p_{K,n} - q_{s,n}) \tag{55}$$

$$\lambda_2 = \frac{|C_{K+1}|b_w}{\gamma_{K+1}} - \frac{|C_{K+1}|b_{sw}}{\gamma_K}, \tag{56}$$

where λ_1 has multiplicity $K-1$. Further, the eigenvector corresponding to λ_2 is constant across nodes not in C_{K+1}.

Proof. Recall that from Subsection D.4 the non-zero eigenvalues of $\mathscr{L}$ are the same as that of

$$\tilde{B}_{eig} = (Z'RZ)^{1/2}B(Z'RZ)^{1/2}$$

Now,

$$Z'RZ = diag\left(\frac{|C_K|}{\gamma_K}, \ldots, \frac{|C_K|}{\gamma_K}, \frac{|C_{K+1}|}{\gamma_{K+1}}\right)$$

Consequently,

$$\tilde{B}_{eig} = \begin{pmatrix} \frac{|C_K|}{\gamma_K} B_s & \left(\frac{|C_K||C_{K+1}|}{\gamma_K \gamma_{K+1}}\right)^{1/2} b_{sw}\mathbf{1} \\ \left(\frac{|C_K||C_{K+1}|}{\gamma_K \gamma_{K+1}}\right)^{1/2} b_{sw}\mathbf{1}' & \frac{|C_{K+1}|}{\gamma_{K+1}} b_w \end{pmatrix},$$

One sees that

$$v_1 = (\sqrt{|C_K|\gamma_K}, \ldots, \sqrt{|C_K|\gamma_K}, \sqrt{|C_{K+1}|\gamma_{K+1}})'$$

is an eigenvector of $\tilde{B}_{eig}$ with eigenvalue 1. Next, consider a vector $v_2 = (v_{21}', 0)'$. Here v_{21} is a $K \times 1$ dimensional vector that is orthogonal to the constant vector. We claim that v_2 so defined is also an eigenvector of $\tilde{B}_{eig}$. To see this notice that

$$\tilde{B}_{eig}\, v_2 = \frac{|C_K|}{\gamma_K} \begin{pmatrix} B_s v_{21} \\ 0 \end{pmatrix},$$

Here we use the fact that $\mathbf{1}'v_{21} = 0$ as v_{21} is orthogonal to $\mathbf{1}$. Next, notice that

$$B_s = ((p_{K,n} - q_{s,n})I + q_{s,n}\mathbf{1}\mathbf{1}')$$

Consequently,

$$B_s v_{21} = (p_{s,n} - q_{s,n})v_{21}$$

The above implies that v_2 is an eigenvector of $\tilde{B}_{eig}$ with eigenvalue λ_1, given by 55.

Notice that from the above construction can construct $K-1$ orthogonal eigenvectors v_k, for $k = 2, \ldots, K$, such that v_k's are also orthgonal to v_1. Essentially, for $k \geq 2$, each $v_k = (v_{k1}', 0)'$, where $v_{k1}'\mathbf{1} = 0$. There are $K-1$ orthogonal choices of the v_{k1}'s.

Given that 1 and λ_1 are eigenvalues of $\tilde{B}_{eig}$, with the latter having multiplicity $K-1$, the remaining eigenvalue is given by,

$$\begin{aligned} \lambda_2 &= trace(\tilde{B}_{eig}) - 1 - (K-1)\lambda_1 \\ &= \frac{|C_K|p_{K,n}}{\gamma_K} + (K-1)\frac{|C_K|}{\gamma_K} q_{s,n} + \frac{|C_{K+1}|b_w}{\gamma_{K+1}} - 1 \\ &= \frac{|C_{K+1}|b_w}{\gamma_{K+1}} - \frac{|C_{K+1}|b_{sw}}{\gamma_K} \end{aligned}$$

The claim regarding the eigenvector corresponding to λ_2 follows from seeing that this should be the case since it is orthogonal to eigenvectors $v_1, \ldots, v_K$ defined above. □

D Analysis of SBM with K blocks

Throughout this section we assume that we have samples from a K block SBM. Denote the sample and population regularized Laplacian as L_τ, $\mathscr{L}_\tau$ respectively. For ease of notation, we remove the subscript τ from the various matrices such as L_τ, $\mathscr{L}_\tau$, A_τ, D_τ, $\mathscr{D}_\tau$. We also remove the subscript τ in the $\hat{d}_{i,\tau}, d_{i,\tau}$'s and denote these as $\hat{d}_i$, d_i respectively. However, in some situations we may need to refer to these quantities at $\tau = 0$. In such cases, we make this clear by writing them as $\hat{d}_{i,0}$, for $i = 1, \ldots, n$ and $d_{i,0}$ for $i = 1, \ldots, n$.

We divide the bounds into two types, viz. deterministics bounds and high probability bounds.

D.1 Deterministic bounds

The following are bounds on the deterministic quantities in the above expression.

Lemma 22.

$$\|\mathscr{V}_i\| = \frac{1}{\sqrt{|C_{k(i)}|}}, \tag{57}$$

here $C_{k(i)}$ is the cluster containing node i, and $|.|$ denotes its size.

Proof. This follows from facts about the eigenvector of the symmetric Laplacian for the stochastic block model. See Appendix D.4 for the proof. □

Lemma 23. *The following bounds hold.*

$$\|P_{i*}\| \leq \sqrt{d_i} \tag{58}$$

$$\|\mathscr{L}_{i*}\| \leq 1/\sqrt{\tau_{min}} \tag{59}$$

Proof. To see (58), notice that

$$\begin{aligned} \|P_{i*}\|^2 &= \sum_{j=1}^{n} p_{ij}^2 \\ &\leq \sum_{j=1}^{n} p_{ij} = d_i \end{aligned}$$

Further, (59) follows from noting that $\mathscr{L}_{i*} = d_i^{-1/2} P_{i*} \mathscr{D}^{-1/2}$. Then one has

$$\begin{aligned} \|\mathscr{L}_{i*}\| &\leq d_i^{-1/2} \|P_{i*}\| \|\mathscr{D}^{-1/2}\| \\ &\leq d_i^{-1/2} d_i^{1/2} \tau_{min}^{-1/2} = \tau_{min}^{-1/2} \end{aligned}$$

□

D.2 High probability bounds

We need probabilistic bounds on the weigthed sum of Bernoulli random variables. The following lemma is proved in [15].

Lemma 24. *Let W_j, $1 \leq j \leq N$ be N independent Bernoulli(r_j) random variables. Furthermore, let α_j, $1 \leq j \leq N$ be non-negative weights that sum to 1 and let $N_\alpha = 1/\max_j \alpha_j$. Then the weighted sum $\hat{r} = \sum_j \alpha_j W_j$, which has mean given by $r^* = \sum_j \alpha_j r_j$, satisfies the following large deviation inequalities. For any r with $0 < r < r^*$,*

$$P(\hat{r} < r) \leq \exp\{-N_\alpha D(r\|r^*)\} \tag{60}$$

and for any $\tilde{r}$ with $r^* < \tilde{r} < 1$,

$$P(\hat{r} > \tilde{r}) \leq \exp\{-N_\alpha D(\tilde{r}\|r^*)\} \tag{61}$$

where $D(r\|r^*)$ denotes the relative entropy between Bernoulli random variables of success parameters r and r^*.

The following is an immediate corollary of the above.

Corollary 25. *Let W_j be as in Lemma 24. Let β_j, for $j = 1, \ldots, N$ be non-negative weights, and let*

$$W = \sum_{j=1}^{N} \beta_j W_j.$$

Then,

$$P\left(W - E(W) > \delta\right) \leq \exp\left\{-\frac{1}{2\max_j \beta_j} \frac{\delta^2}{(E(W) + \delta)}\right\} \tag{62}$$

and

$$P\left(W - E(W) < -\delta\right) \leq \exp\left\{-\frac{1}{2\max_j \beta_j} \frac{\delta^2}{E(W)}\right\} \tag{63}$$

Proof. Here we use the fact that

$$D(r||r^*) \geq (r - r^*)^2/(2r), \tag{64}$$

for any $0 < r, r^* < 1$. We prove (62). The proof of (63) is similar. The event under consideration may be written as

$$\{\hat{r} - r^* > \tilde{\delta}\},$$

where $\hat{r} = W/\sum_j \beta_j$, $r^* = E(W)/\sum_j \beta_j$ and $\tilde{\delta} = \delta/\sum_j \beta_j$. Correspondingly, using Lemma 24 and (64), one gets that

$$P\left(W - E(W) > \delta\right) \leq \exp\left\{-\frac{\sum_j \beta_j}{\max_j \beta_j} \frac{\tilde{\delta}^2}{2(r^* + \tilde{\delta})}\right\}.$$

Substituting the values of $\tilde{\delta}$ and r^* results in bound (62). □

The following lemma provides high probability bounds on the degree.

Lemma 26. *On a set E_1 of probability at most $1 - 2/n^{c_1-1}$, one has*

$$|\hat{d}_{i,\tau} - d_{i,\tau}| \leq c_2\sqrt{\tau_i \log n} \quad \textit{for each } i = 1, \ldots, n.,$$

where $c_1 = .5c_2^2/(1 + c_2/\sqrt{c})$.

Proof. Use the fact that $\hat{d}_{i,\tau} - d_{i,\tau} = \hat{d}_{i,0} - d_{i,0}$, and

$$P(|\hat{d}_{i,0} - d_{i,0}| \leq c_2\sqrt{\tau_i \log n} \quad \forall i) \leq \sum_{i=1}^{n} P(|\hat{d}_{i,0} - d_{i,0}| \leq c_2\sqrt{\tau_i \log n})$$

Notice that $\hat{d}_{i,0} = \sum_{j=1}^{n} A_{ij}$. Apply Corollary 25 with $\beta_j = 1$ and $W_j = A_{ij}$, and $\delta = c_2\sqrt{\tau_{min} \log n}$ to bound each term in the sum of the right side of the above equation.

The error exponent can be bounded by,

$$2n \exp\left\{-\frac{1}{2}\frac{\delta^2}{(E(W) + \delta)}\right\}. \tag{65}$$

We claim that,

$$E(W) + \delta \leq (1 + c_2/\sqrt{c})\tau_i. \tag{66}$$

Substituting the above bound in the error exponent (65) will complete the proof.

To see the claim, notice that $E(W) = d_{i,0}$. Now, consider the case $d_{i,0} \geq \tau$. In this case, $\tau_i = d_{i,0}$ and $\log n < d_{i,0}/c$. Correspondingly, $E(W) + \delta$ is at most $d_{i,0}(1 + c_2/\sqrt{c})$.

Next, consider the case $d_{0,i} < \tau$. In this case $\tau_i = \tau_{min}$, which is at least $c \log n$ from Assumption 1. Consequently,

$$E(W) + \delta \leq c \log n + c_2\sqrt{c} \log n.$$

The right side of the above can be bounded by $(1 + c_2/\sqrt{c})\tau$. This proves the claim. □

The following is an immediate consequence of the above lemma.

Corollary 27. *On the set E_1 of Lemma 26, one has*

$$|\Delta R_i| \leq c_3 \frac{\sqrt{\log n}}{\sqrt{\tau_i}} \quad \textit{for each } i = 1, \ldots, n,$$

where recall that

$$c_3 = \left(\frac{c_2}{\sqrt{1 - c_2/\sqrt{c}}}\right).$$

Proof. From Lemma 26 one gets that with probability at least $1 - 2/n^{c_1 - 1}$, the following two events hold for each $i = 1, \ldots, n$.

$$|\hat{d}_i - d_i| \leq c_2 \sqrt{\tau_i \log n} \tag{67}$$

and

$$\hat{d}_i > \max \left\{ \tau, d_i \left(1 - c_2 \frac{\sqrt{\log n}}{\sqrt{\tau_i}} \right) \right\} \tag{68}$$

Now use the fact that τ_i is at least $c \log n$ to get that $\hat{d}_i$ is at least $(1 - c_2/\sqrt{c})\tau_i$. Now, notice that

$$\begin{aligned} \Delta R_i &= \frac{d_i^{1/2} - \hat{d}_i^{1/2}}{\hat{d}_i^{1/2}} \\ &= \frac{d_i - \hat{d}_i}{\hat{d}_i^{1/2}(\hat{d}_i^{1/2} + d_i^{1/2})} \end{aligned}$$

Correspondingly, with high probability,

$$|\Delta R_i| \leq \frac{c_2 \sqrt{\tau_i \log n}}{\sqrt{1 - c_2/\sqrt{c}} \tau_i}.$$

This leads to completion of the proof. □

In the lemma below we bound the quantity $\|\Delta A_{i*}\|$.

Lemma 28. *On a set E_2 of probability at least $1 - 1/n^{c_1 - 1}$*

$$\|\Delta A_{i*}\|^2 \leq d_i + c_2 \sqrt{\tau_i \log n} \quad \textit{for } i = 1, \ldots, n.$$

Here $c_1 = .5c_2^2/(1 + c_2/\sqrt{c})$.

Proof. Recall that

$$\|\Delta A_{i*}\|^2 = \sum_{j=1}^{n} (A_{ij} - p_{ij})^2.$$

For the time being, assume that $p_{ij} \leq 1/2$. The random variable $(A_{ij} - p_{ij})^2$ may be expressed as,

$$(A_{ij} - p_{ij})^2 = W_{ij}(1 - p_{ij})^2 + (1 - W_{ij})p_{ij}^2,$$

where W_{ij} is a Bernoulli random variable with success probability p_{ij}. Simplifying, one sees that

$$(A_{ij} - p_{ij})^2 = (1 - 2p_{ij})W_{ij} + p_{ij}^2.$$

If $p_{ij} > 1/2$, take W_{ij} to be Bernoulli$(1 - p_{ij})$, to get that

$$(A_{ij} - p_{ij})^2 = (2p_{ij} - 1)W_{ij} + (1 - p_{ij})^2.$$

Correspondingly,

$$(A_{ij} - p_{ij})^2 = \beta_{ij} W_{ij} + \min(p_{ij}, 1 - p_{ij})^2,$$

where $\beta_{ij} = \max(1 - 2p_{ij}, 2p_{ij} - 1)$, and W_{ij} is Bernoulli($\min(p_{ij}, 1 - p_{ij})$). Notice that $0 \leq \beta_{ij} \leq 1$. Consequently,

$$\|\Delta A_j\|^2 = \sum_{j=1}^{n} \beta_{ij} W_{ij} + \sum_{i=1} \min(p_{ij}, 1 - p_{ij})^2 \tag{69}$$

Now use the fact that $\max_i \beta_{ij} \leq 1$ and $\sum_{i=1}^{n} \beta_{ij} E(W_{ij}) \leq d_{0,i}$ in Corollary 25, along with arguments similar to Lemma 26, to get that with probability $1 - 1/n^{c_1 - 1}$, for each $i = 1, \ldots, n$, one has

$$\sum_{j=1}^{n} \beta_{ij} W_{ij} \leq c_2 \sqrt{\tau_i \log n}.$$

Now, use

$$\begin{aligned} \sum_{j=1}^{n} \min(p_{ij}, 1 - p_{ij})^2 &\leq \sum_{i=1}^{n} p_{ij}^2 \\ &\leq d_i \end{aligned}$$

in (69) to prove the lemma. □

The next lemma will require Hoeffding's inequality, which we state below

Lemma 29 (Hoeffding's inequality)**.** *If W_j be Bernoulli(p_j), and let β_j be real numbers. Then,*

$$P\left(|\sum_{j=1}^{N} \beta_j (W_j - E(W_j))| > t \right) \leq 2 \exp\left\{ - \frac{-t^2}{\sum_{i=1}^{n} \beta_j^2} \right\}.$$

We use the above to bound $\|A_{i*} \mathscr{D}^{-1/2} \mathscr{V}\|$.

Lemma 30. *On a set E_3 of probability at least $1 - 2K/n^{c_1 - 1}$*

$$\|\Delta A_{i*} \mathscr{D}^{-1/2} \mathscr{V}\| \leq c_2 K \sqrt{\frac{\log n}{\tau_{min}}} \quad \textit{for } i = 1, \ldots, n.$$

Here $c_1 = .5 c_2^2 / (1 + \sqrt{c_2}/c)$.

Proof. We prove that for each k,

$$P\left(|\Delta A_{i*} \mathscr{D}^{-1/2} \mathscr{V}_{*k}| > c_2 \sqrt{\frac{K \log n}{\tau_{min}}} \quad \text{for } i = 1, \ldots, n \right) \leq 2/n^{c_1 - 1}. \tag{70}$$

Applying the union bound over k will complete the proof. To see the above, use Lemma 29 with $\beta_j = d_j^{-1/2} \mathscr{V}_{jk}$ and $W_j = A_{ij}$. Notice that,

$$\begin{aligned}\sum_{j=1}^{n} \beta_j^2 &= \sum_{j=1}^{n} d_j^{-1} \mathscr{V}_{jk}^2 \\ &\leq \sum_{j=1}^{n} d_j^{-1} \|\mathscr{V}_{j*}\|^2 \\ &= trace(\mathscr{V}^T \mathscr{D}^{-1} \mathscr{V}) \\ &\leq K/\tau_{min}\end{aligned}$$

The inequality uses that the columns of $\mathscr{V}$ are orthogonal and $d_j \geq \tau_{min}$ for each j. Substituting the above in the Hoeffding's bound of Lemma 29 and using $c_2^2/2 > c_1$, one gets (70). Consequently, taking a union bound over k completes the proof of the Lemma. □

D.3 Concentration of Laplacian

Below we provide the proof of Lemma 18.

Proof of Lemma 18. For completeness, we give the outline of the proof in [23], adapted to our case. Write $\tilde{L} = \mathscr{D}^{-1/2} A \mathscr{D}^{-1/2}$. Then,

$$\|L - \mathscr{L}\| \leq \|L - \tilde{L}\| + \|\tilde{L} - \mathscr{L}\|.$$

We first bound $\|L - \tilde{L}\|$. Let $F = D^{1/2} \mathscr{D}^{-1/2}$. Then $\tilde{L} = FLF$. Correspondingly,

$$\begin{aligned}\|L - \tilde{L}\| &\leq \|L - FL\| + \|FL - \tilde{L}\| \\ &\leq \|I - F\| \|L\| + \|F\| \|L\| \|I - F\|\end{aligned}$$

As in [23], we use the fact that $\sqrt{1+x} - 1 \leq x$ for $x \in [-3/4, 3/4]$. Notice that

$$F - I = (I + (\Delta D)\mathscr{D}^{-1})^{1/2} - I.$$

From Lemma 26, notice that

$$|\Delta \hat{d}_i|/d_i \leq c_2 \sqrt{\tau_i \log n}/d_i$$

with probability at least $1 - n^{1-c_1}$. The right side of the above is at most $1/2$ (which is less than $3/4$) by choosing c and c_2 as in the theorem. Correspondingly,

$$\|F - I\| \leq c_2 \max_i \sqrt{\tau_i \log n}/d_i$$

with high probability. Use $\tau_i \leq d_i$ and $d_i \geq \tau_{min}$ to get that

$$\|F - I\| \leq c_2 \sqrt{\log n}/\sqrt{\tau_{min}}.$$

Use the fact that, $\|L\| \leq 1$ to get that,

$$\|L - \tilde{L}\| \leq c_2\sqrt{\frac{\log n}{\tau_{min}}}\left(2 + c_2\sqrt{\frac{\log n}{\tau_{min}}}\right)$$

with probability at least $1 - 1/n^{c_1-1}$.

Next, we bound $\|\tilde{L} - \mathscr{L}\|$. Notice that,

$$\tilde{L} - \mathscr{L} = \sum_{i \leq j} Y_{ij},$$

where $Y_{ij} = \mathscr{D}^{-1/2} X_{ij} \mathscr{D}^{-1/2}$, with

$$X_{ij} = \begin{cases} (A_{ij} - P_{ij})\left(e_i e_j^T + e_j e_i^T\right), & \text{if } i \neq j \\ (A_{ij} - P_{ij}) e_i e_i^T & \text{if } i = j \end{cases}$$

as in [23]. Here e_i is the i-th column of the $n \times n$ identity matrix. In our case one has,

$$\|Y_{ij}\| \leq 1/\sqrt{d_i d_j} \leq 1/\tau_{min}.$$

Further, let $\sigma^2 = \|\sum_{i \leq j} E(Y_{ij}^2)\|$, which at most $1/\tau_{min}$ from the above. Then, applying Corollary 4.2 in [19], one gets

$$P\left(\|\tilde{L} - \mathscr{L}\| \geq t\right) \leq n e^{-t^2/2\sigma^2}.$$

Consequently, with probability at least $1 - 1/n^{c_1-1}$ one has,

$$\|\tilde{L} - \mathscr{L}\| \leq \sqrt{\frac{2c_1 \log n}{\tau_{min}}}.$$

Thus with probability at least $1 - 2/n^{c_1-1}$, one has that $\|L - \mathscr{L}\|$ is bounded by

$$\sqrt{\frac{\log n}{\tau_{min}}}\left(\sqrt{2c_1} + c_2(2 + c_2/\sqrt{c})\right),$$

where we use the fact that $\log n/\tau_{min} \leq 1/c$, since $\tau_{min} > c \log n$. □

D.4 Proof of Lemmas 1 and 22

Notice that the population regularized Laplacian $\mathscr{L}_\tau$ corresponds to the population Laplacian of an ordinary stochastic block model with block probability matrix

$$B_\tau = B + vv',$$

where $v = (\sqrt{\tau/n})\mathbf{1}$. Correspondingly, we can use the following facts of the population eigenvectors and eigenvalues given for a SBM.

Let Z be the community membership matrix, that is, the $n \times K$ matrix with entry (i, k) being 1 if node i belongs to cluster C_k. Then, the following is proved in [25].

1. Let $R = \mathscr{D}_\tau^{-1}$. Then, the non-zero eigen values of $\mathscr{L}_\tau$ are the same as that of
$$B_{eig} = B_\tau(Z'RZ), \tag{71}$$
or equivalently, $\tilde{B}_{eig} = (Z'RZ)^{1/2} B_\tau (Z'RZ)^{1/2}$.

2. Define $\mu = R^{1/2} Z (Z'RZ)^{-1/2}$. Let,
$$\tilde{B}_{eig} = H \Lambda H^T,$$
where the right side of the above gives the singular value decomposition of the matrix on the right. Then, the eigenvectors of $\mathscr{L}_\tau$ are given by μH.

Further, since in the stochastic block model the expected node degrees are the same for all nodes in a particular cluster, one can write $R^{1/2} Z = ZQ$, where Q^{-2} is the $K \times K$ diagonal matrix of population degrees of nodes in a particular community. Consequently, one sees that
$$\mu H = Z(Z^T Z)^{-1/2} H.$$
Lemmas 1 and 22 follows from noting that
$$\mu H (\mu H)^T = Z(Z^T Z)^{-1} Z^T$$
and the fact that $(Z^T Z)^{-1} = diag(1/|C_1|, \ldots, 1/|C_K|)$.

E Proof of results in Subsection 3.1

Here we prove results of Subsection 3.1. The proof of both lemmas use (9). With τ_n as in the lemmas, one sees that $\tau_{min} = \tau_n$. Thus, Lemma 5 follows from (12).

E.1 Proof of Lemma 5

We first prove (12). For convenience we remove the subscript n from the various quantities. Further, at the risk of ambiguity of notation, we introduce the subscript τ whenever the quantities depend upon τ. Recall, p_1, p_2 are the within community probability and q the between community probability.
$$B = \begin{pmatrix} p_1 & q \\ q & p_2 \end{pmatrix}$$
Let $p_{1,\tau}$, $p_{2,\tau}$ and q_τ be equal to τ/n added to p_1, p_2 and q respectively. Then the elements of the 2×2 matrix B_{eig}, given in (71), are given by,
$$\tilde{B}_{11} = \frac{p_{1,\tau}|C_1|}{\gamma_{1,\tau}} \qquad \tilde{B}_{12} = q_\tau \frac{|C_1|}{\gamma_{1,\tau}}$$
$$\tilde{B}_{21} = q_\tau \frac{|C_2|}{\gamma_{2,\tau}} \qquad \tilde{B}_{22} = p_{2,\tau} \frac{|C_2|}{\gamma_{2,\tau}}$$

Here $\gamma_{1,\tau}$, $\gamma_{2,\tau}$ are the degrees of the two communities after regularization, that is, $\gamma_{1,\tau} = \gamma_1 + \tau$ and $\gamma_{2,\tau} = \gamma_2 + \tau$, where recall that,

$$\gamma_1 = p_1|C_1| + q|C_2|$$
$$\gamma_2 = q|C_1| + p_2|C_2|$$

Using the fact that 1 is an eigenvalue of B_{eig}, the smallest eigenvalue can be computed easily. One sees that,

$$\mu_{2,\tau} = .5\frac{a_1}{\gamma_1 + \tau} + .5\frac{a_2}{\gamma_2 + \tau},$$

where $a_1 = p_{1,\tau}|C_1| - q_\tau|C_2|$ and $a_2 = p_{2,\tau}|C_2| - q_\tau|C_1|$. In the case of equal community size, that is $|C_1| = |C_2| = C$, one has that $a_1 = (p_1 - q)C$ and $a_2 = (p_2 - q)C$. In other words, a_1 and a_2 do not depend upon τ. It is then seen that $\mu_{2,\tau}$ is a decreasing function of τ and (12) holds. Substituting $\tau = \gamma_1$, it is seen that $\mu_{2,\tau}$ is atleast $\mu_{2,0}/4$.

Further, from the assumption of Lemma 5, one gets Assumption 1 and 2 holds for large n. The proof of the lemma is completed by using (9)

E.2 Proof of Lemma 7

We first prove that μ_{K,τ_n} is bounded away from zero for $\tau_n = \gamma_{K-1,n}$. Consider the stochastic block model with K communities, where K is fixed. The matrix B is given by (14). We first consider the case that $q_n = 0$, that is, there is no interaction between the clusters. Once again, we remove the subscript n from the various quantities, and introduce τ in the subscript whenever the quantities depend on τ.

Recall that from (71), $B_{eig} = (B + vv')F$, where $v = \sqrt{\tau/n}\mathbf{1}$ and $F = Z'RZ$. We need to find the K-th smallest eigenvalue of B_{eig}. This is also the inverse of the largest eigenvalue of B_{eig}^{-1}.

Consequently, we now show that the maximum eigenvalue of B_{eig}^{-1} is bounded from above. Use the fact that since K is fixed,

$$\lambda_{max}(B_{eig}^{-1}) \asymp trace(B_{eig}^{-1}).$$

Correspondingly, we now proceed to calculate the trace of B_{eig}^{-1} and show that it is bounded from above. Notice,

$$B_{eig}^{-1} = F^{-1}(B + vv')^{-1},$$

where

$$F^{-1} = diag\left(\frac{\gamma_1 + \tau}{C}, \ldots, \frac{\gamma_K + \tau}{C}\right),$$

where C denotes the size of the communities, which are assumed to be equal. Here, for convenience, we remove the subscript n and denote $\gamma_{i,n}$ by simply γ_i. Using Sherman-Morrison formula

$$(B + vv')^{-1} = B^{-1} - \frac{(B^{-1}v)(B^{-1}v)'}{1 + v'B^{-1}v}$$

One sees that,

$$B^{-1}v = \sqrt{\tau/n}(1/p_1, \ldots, 1/p_K)'$$

Correspondingly,

$$v'B^{-1}v = \frac{\tau}{n}\sum_i 1/p_1 = \tau m_1,$$

where $m_1 = (1/K)\sum_k 1/\gamma_k$, using $\gamma_k = p_k C$ and $n = KC$. Further, the diagonal entries of the matrix $(B^{-1}v)(B^{-1}v)'$ can be written as

$$\frac{\tau}{n} diag(1/p_1^2, \ldots, 1/p_K^2).$$

We need the trace of B_{eig}^{-1}. Using the above, one sees that this is the same as

$$\sum_k \frac{\gamma_k + \tau}{\gamma_k} - \frac{(\tau/n)\sum_i(\gamma_k + \tau)/(Cp_k^2)}{1 + \tau m_1},$$

using $\gamma_k = p_k C$. Consequently,

$$trace(B_{eig}^{-1}) = \sum_k \frac{\gamma_k + \tau}{\gamma_k} - \frac{\tau m_1 + \tau^2 m_2}{1 + \tau m_1},$$

where $m_2 = (1/K)\sum_k /\gamma_k^2$. Thus we have,

$$trace(B_{eig}^{-1}) = K\tau m_1 - \frac{\tau m_1 + \tau^2 m_2}{1 + \tau m_1}$$

The above is leads to,

$$\begin{aligned} trace(B_{eig}^{-1}) &= \frac{K\tau m_1 + K\tau^2 m_1^2 - \tau m_1 - \tau^2 m_2}{1 + \tau m_1} \\ &= \frac{(K-1)\tau m_1}{1 + \tau m_1} + \frac{(K-1)\tau^2 m_1^2 - \tau^2 v}{1 + \tau m_1} \\ &= \frac{(K-1)\tau m_1}{1 + \tau m_1} + \tau^2 m_1^2 \frac{\left[(K-1) - c_v^2\right]}{1 + \tau m_1} \end{aligned} \tag{72}$$

Here $v = 1/K\sum_i(1/\gamma_i - m_1)^2$ and $c_v = \sqrt{v}/m_1$ is the associated coefficient of variation. We claim that $(K-1) - c_v \asymp \gamma_K/\gamma_{K-1}$. To see this, notice that,

$$v/m_1^2 = K\frac{\sum_{k=1}^K 1/\gamma_k^2}{\left(\sum_{k=1}^K 1/\gamma_k\right)^2} - 1.$$

One sees that,

$$1 - \frac{\sum_{k=1}^K 1/\gamma_k^2}{\left(\sum_{k=1}^K 1/\gamma_k\right)^2} \asymp \gamma_K/\gamma_{K-1}.$$

Further, since the first term in (72) is bounded, one gets that,

$$trace(B_{eig}^{-1}) \lesssim \frac{\tau^2 m_1^2 \gamma_K/\gamma_{K-1}}{1+\tau m_1}. \\ \lesssim \frac{(\tau/\gamma_{K-1})(\tau/\gamma_K)}{1+\tau/\gamma_K},$$

where for the last relation we use the fact that $m_1 \asymp 1/\gamma_K$. Consequently, one gets that if $\tau = \gamma_{K-1}$ then $trace(B_{eig}^{-1}) \asymp 1$. This implies that $\mu_{K,\tau}$ is bounded below from zero.

As a consequence, Assumption 2 is satisfied for large n if $\log n/\gamma_{K-1} = o(1)$. We remark that the above results also holds if $\tau_n \lesssim \gamma_{K-1,n}$ as well. This completes the proof of the lemma for $q_n = 0$.

Now consider the K block model with off-diagonal elements of B equal to q. Notice that

$$B_\tau = B_0 + \tilde{v}(\tilde{v})^T,$$

where $B_0 = diag(p_1 - q, \ldots, p_K - q)$ and $\tilde{v} = \sqrt{\tilde{\tau}/n}\mathbf{1}$, where $\tilde{\tau} = \tau + nq$. Thus applying the above result for the diagonal block model one gets that if $\tilde{\tau} = C(p_{K-1} - q)$, the quantity $\mu_{K,\tau}$ is bounded away from 0. With $\tilde{\tau}$ as above, one has $\tau \asymp \gamma_{K-1}$, as $nq = o(\gamma_{K-1})$ since $q = o(p_{K-1})$.

References

[1] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[2] A.A. Amini, A. Chen, P.J. Bickel, and E. Levina. Fitting community models to large sparse networks. *Ann. Statist*, 41(4):2097–2122, 2013.

[3] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. *Advances in Neural Information Processing Systems*, 25(3), 2011.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[5] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer, 1997.

[6] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[7] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012:1–23.

[8] A. Chen, A. Amini, P. Bickel, and L. Levina. Fitting community models to large sparse networks. In *Joint Statistical Meetings, San Diego*, 2012.

[9] Anirban Dasgupta, John E Hopcroft, and Frank McSherry. Spectral analysis of random graphs with skewed degree distributions. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 602–610. IEEE, 2004.

[10] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. seventh ACM SIGKDD inter. conf. on Know. disc. and data mining*, pages 269–274. ACM, 2001.

[11] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

[12] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11(9): 1074–1085, 1992.

[13] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[14] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[15] A. Joseph and A.R. Barron. Fast sparse superposition codes have near exponential error probability for $R < C$. *IEEE. Trans. Inform. Theory, to appear*, 2013.

[16] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[17] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 299–308. IEEE, 2010.

[18] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. Improved cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. *arXiv preprint arXiv:1301.5584*, 2013.

[19] L. Mackey, M.I. Jordan, R.Y. Chen, B. Farrell, and J.A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *arXiv preprint arXiv:1201.6002*, 2012.

[20] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[21] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[22] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[23] R.I. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[24] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *arXiv preprint arXiv:1309.4111*, 2013.

[25] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[26] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pat. Analysis and Mach. Intel.*, 22(8):888–905, 2000.

[27] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[28] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[29] YY Yao. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Springer, 2003.