*The Annals of Statistics*
2016, Vol. 44, No. 4, 1765–1791
DOI: 10.1214/16-AOS1447
© Institute of Mathematical Statistics, 2016

# IMPACT OF REGULARIZATION ON SPECTRAL CLUSTERING[1]

BY ANTONY JOSEPH AND BIN YU

*@WalmartLabs and University of California, Berkeley*

The performance of spectral clustering can be considerably improved via regularization, as demonstrated empirically in Amini et al. [*Ann. Statist.* **41** (2013) 2097–2122]. 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. By examining the scenario where the regularization parameter $\tau$ is large, we show that the minimum degree assumption can potentially be removed. As a special case, for an SBM with two blocks, the results require the maximum degree to be large (grow faster than $\log n$) as opposed to the minimum degree. More importantly, we show the usefulness of regularization in situations where not all nodes belong to well-defined clusters. Our results rely on a 'bias-variance'-like trade-off that arises from understanding the concentration of the sample Laplacian and the eigengap as a function of the regularization parameter. As a byproduct of our bounds, we propose a data-driven technique *DKest* (standing for estimated Davis–Kahan 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 a purpose, 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 [11, 16, 21, 25]. Other examples of applications of spectral clustering include manifold learning [4], image segmentation [25] and text mining [9].

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. In particular, we focus on the regularized spectral clustering (RSC) procedure proposed in Amini

Received July 2014; revised January 2016.

[1] Supported in part by NSF Grants DMS-12-28246 and DMS-11-60319 (FRG), ARO Grant W911NF-11-1-0114, NHGRI Grant 1U01HG007031-01 (ENCODE) and the Center of Science of Information (CSoI), a US NSF Science and Technology Center, under Grant agreement CCF-0939370.

*MSC2010 subject classifications.* Primary 62F12; secondary 62H99.

*Key words and phrases.* Spectral clustering, regularization, network analysis, community detection, stochastic block model.

et al. [2]. Their empirical findings demonstrates that the performance of the RSC 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 Chaudhuri et al. [7] and Qin and Rohe [23].

This paper provides an attempt to provide a theoretical understanding for the regularization in the RSC algorithm. We also propose a practical scheme for choosing the regularization parameter based on our theoretical results. Our analysis focuses on the Stochastic Block Model (SBM) and an extension of this model. Below are the three main contributions of the paper.

(a) We attempt to understand regularization for the stochastic block model. In particular, for a graph with $n$ nodes, previous theoretical analyses for spectral clustering, under the SBM and its extensions, [7, 10, 24, 26] 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 cluster recovery results that potentially do not depend on the above mentioned constraint on the minimum degree. As an example, for an SBM with two blocks (clusters), our results require that the maximum degree be large (grow faster than $\log n$) rather than the minimum degree. This is done in Section 3.

(b) We demonstrate that regularization has the potential of addressing a situation where the lower degree nodes do not belong to well-defined clusters. Our results demonstrate that choosing a large regularization parameter has the effect of removing these relatively lower degree nodes. 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. This is done in Section 4.

(c) Although our theoretical results deal with the 'large' $\tau$ case, it is observed empirically that moderate values of $\tau$ may produce better clustering performance. Consequently, in Section 5 we propose *DKest*, a data dependent procedure for choosing the regularization parameter. We demonstrate that this works well through simulations and on a real data set. This is in Section 5.

Our theoretical results involve understanding the trade-offs between the *eigengap* and the concentration of the sample Laplacian when viewed as a function of the regularization parameter. Assuming that there are $K$ clusters, the eigengap refers to the gap between the $K$th smallest eigenvalue and the remaining eigenvalues. An adequate gap ensures that the sample eigenvectors can be estimated well ([16, 21, 27]) which leads to good cluster recovery. The adequacy of an eigengap for cluster recovery is in turn determined by the concentration of the sample Laplacian.

In particular, a consequence of the Davis–Kahan theorem [5] is that if the spectral norm of the difference of the sample and population Laplacians is small compared to the eigengap then the top $K$ eigenvector can be estimated well. Denoting $\tau$ as the regularization parameter, previous theoretical analyses of regularization ([7, 24]) provided high-probability bounds on this spectral norm. These bounds have a $1/\sqrt{\tau}$ dependence on $\tau$, for large $\tau$. In contrast, our high probability bounds behave like $1/\tau$, for large $\tau$. We also demonstrate that the eigengap behaves like $1/\tau$ for large $\tau$. The end result is that we show that one can get a good understanding of the impact of regularization by understanding the situation where $\tau$ goes to infinity. This also explains empirical observations in [2, 23] where it was seen that performance of regularized spectral clustering does not change for $\tau$ beyond a certain value. Our procedure for choosing the regularization parameter works by providing estimates of the Davis–Kahan bounds over a grid of values of $\tau$ and then choosing the $\tau$ that minimizes these estimates.

The paper is divided as follows. In the next subsection, we discuss preliminaries. In particular, in Section 1.1 we review the RSC algorithm of [2], and also discuss the other forms of regularization in literature. In Section 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 describes our *DKest* data dependent method for choosing the regularization parameter.

1.1. *Regularized spectral clustering*. In this section, we review the regularized spectral clustering (RSC) algorithm of Amini et al. [2].

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, for $1 \le i, j \le n$,

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

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 fewer 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 [11] and Normalized Cut [25]. 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 [11, 25].

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

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

**Algorithm 1** The RSC-$\tau$ Algorithm [2]

```
Input: Laplacian matrix $L_\tau$.
Step 1: Compute the $n \times K$ eigenvector matrix $V_\tau$.
Step 2: Use the $K$-means algorithm to cluster the rows of $V_\tau$ into $K$ clusters.
```

Regularization is introduced in the following way: Let $J$ be a constant matrix with all entries equal to $1/n$. Then, in regularized spectral clustering one constructs a new adjacency matrix by adding $\tau J$ 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

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

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

The RSC algorithm for finding $K$ communities is described in Algorithm 1. In order to bring to the forefront the dependence on $\tau$, we also denote the RSC algorithm as RSC-$\tau$. The algorithm first computes $V_\tau$, the $n \times K$ eigenvector matrix corresponding to the $K$ largest eigenvalues of $L_\tau$. The columns of $V_\tau$ are taken to be orthogonal. The rows of $V_\tau$, denoted by $V_{i,\tau}$, for $i = 1, \dots, n$, corresponds to the nodes in the graph. Clustering the rows of $V_\tau$, for example, using the $K$-means algorithm, provides a clustering of the nodes. We remark that the RSC-0 algorithm corresponds to the usual spectral clustering algorithm.

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 RSC algorithm. Recent works [2, 6, 10, 14, 24] have used this model, and its variants, to provide a theoretical analyses for various community detection algorithms.

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

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

Here, the subscript deg stands for 'degree' since the usual Laplacian is modified by adding $\tau$ to the degree matrix $D$. Notice that for the RSC algorithm the matrix $A$ in the above expression was replaced by $A_\tau$.

As mentioned before, we attempt to understand regularization in the framework of the SBM and its extension. We review the SBM in the next section. Using recent results on the concentration of random graph Laplacians [22], we were able to show concentration results in Theorem 4 for the regularized Laplacian in the RSC algorithm. Previous concentration results for the Laplacian (2), as in [7], provide high probability bounds on the spectral norm of the difference of the sample and population regularized Laplacians that depends inversely on $1/\sqrt{\tau}$. However, for the regularization (1) we show that the dependence is inverse in $\tau$, for large $\tau$. We believe that this holds for the regularization (2) as well. We also demonstrate that the eigengap depends inversely on $\tau$, for large $\tau$. The benefit of this, along with our improved concentration bounds, is that one can understand regularization by looking at the case where $\tau$ is large. This results in a very neat criterion for the cluster recovery with the RSC-$\tau$ algorithm.

## 2. The stochastic block model.

Given a set of $n$ nodes, the stochastic block model (SBM), introduced in [12], 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 $C_1, \dots, C_K$. Let $n_k$, for $k = 1, \dots, K$, denote the number of nodes belonging to each of the clusters.

Given the communities, the edges between nodes, say $i$ and $j$, are chosen independently with probability depending on 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

$$P_{ij} = B_{k_1,k_2}.$$

Here, the *block probability matrix*

$$B = \big((B_{k_1,k_2})\big) \qquad \text{where } k_1, k_2 = 1, \dots, K$$

is a symmetric full rank matrix, with each entry between [0, 1]. The $n \times n$ edge probability matrix $P = ((P_{ij}))$, given by (3), represents the population counterpart of the adjacency matrix $A$.

Denote $Z = ((Z_{ik}))$ as the $n \times K$ binary matrix providing the cluster memberships of each node. In other words, each row of $Z$ has exactly one 1, with $Z_{ik} = 1$ if node $i$ belongs to $C_k$. Notice that

$$P = ZBZ'. \tag{3}$$

Here, $Z'$ denotes the transpose of $Z$. Consequently, from (3), it is seen that the rank of $P$ is also $K$.

The population counterpart for the degree matrix $D$ is denoted by $\mathscr{D} = \mathrm{diag}(d_1, \dots, d_n)$, where $\mathscr{D} = \mathrm{diag}(P\mathbf{1})$. Here, $\mathbf{1}$ denotes the column vector of all ones. Similarly, the population version of the symmetric Laplacian $L_\tau$ 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 $P_\tau$ represent the population counterparts to $D_\tau$ and $A_\tau$, respectively. Notice that since $P$ has rank $K$, the same holds for $\mathscr{L}_\tau$.

2.1. *Notation.* We use $\|\cdot\|$ to denote the spectral norm of a matrix. Notice that for vectors this corresponds to the usual $\ell_2$-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 $c_1 a_n \le b_n \le c_2 a_n$. Further, we use $b_n \lesssim a_n$ if $b_n \le c_2 a_n$, for some positive $c_2$ not depending on $n$. The notation $b_n \gtrsim a_n$ is analogously defined.

The quantities

$$d_{\min,n} = \min_{i=1,\dots,n} d_i, \qquad d_{\max,n} = \max_{i=1,\dots,n} d_i$$

denote the minimum and maximum expected degrees of the nodes.

2.2. *The population cluster centers.* We now proceed to define population cluster centers $\mathsf{cent}_{k,\tau} \in \mathbb{R}^K$, for $k = 1, \dots, K$, 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 of the population Laplacian $\mathscr{L}_\tau$. As with $V_\tau$, 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

$$\mathscr{V}_\tau = \arg\min_{H \in \mathcal{H}} \|V_\tau - H\|, \tag{4}$$

where recall that $\|\cdot\|$ denotes the spectral norm. The matrix $\mathscr{V}_\tau$, as defined above, represents the population counterpart of the matrix $V_\tau$.

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 $\|\cdot\|$ 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, e.g., Rohe et al. [24] 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.

Notice that the matrix $\mathscr{V}_{i,\tau}$ depends on the sample eigenvector matrix $V_\tau$ through (4), and consequently is a random quantity. However, the following lemma shows that the pairwise distances between the rows of $\mathscr{V}_{i,\tau}$ are non-random and, more importantly, independent of $\tau$.

LEMMA 1. *Let $i \in C_k$ and $i' \in C_{k'}$. Then*

$$\|\mathscr{V}_{i,\tau} - \mathscr{V}_{i',\tau}\| = \begin{cases} 0 & \text{if } k = k', \\ \sqrt{\dfrac{1}{n_k} + \dfrac{1}{n_{k'}}} & \text{if } k \neq k'. \end{cases}$$

The above lemma is proved in the supplementary material [13]. From the above lemma, there are $K$ distinct rows of $\mathscr{V}_\tau$ corresponding to the $K$ clusters. We denote these as $\mathsf{cent}_{1,\tau}, \ldots, \mathsf{cent}_{K,\tau}$. We also call these the population cluster centers since, intuitively, in an idealized scenario the data points $V_{i,\tau}$, with $i \in C_k$, should be concentrated around $\mathsf{cent}_{k,\tau}$.

2.3. *Cluster recovery using $K$-means algorithm.* Recall that the RSC-$\tau$ Algorithm 1 works by performing $K$-means clustering on the rows of the $n \times K$ sample eigenvector matrix, denoted by $V_{i,\tau}$, for $i = 1, \ldots, n$. In this section, in particular Corollary 3, we relate the fraction of mis-clustered nodes using the $K$-means algorithm to the various parameters in the SBM.

In general, the $K$-means algorithm can be described as follows: Assume one wants to find $K$ clusters, for a given set of data points $x_i \in \mathbb{R}^K$, for $i = 1, \ldots, K$. Then the $K$-clusters resulting from applying the $K$-means algorithm corresponds to a partition $\hat{\mathcal{T}} = \{\hat{T}_1, \ldots, \hat{T}_K\}$ of $\{1, \ldots, n\}$ that aims to minimize the following objective function over all such partitions:

$$\mathsf{Obj}(\mathcal{T}) = \sum_{k=1}^{K} \sum_{i \in T_k} \|x_i - \bar{x}_{T_k}\|^2. \tag{5}$$

Here, $\mathcal{T} = \{T_1, \ldots, T_K\}$ is a partition $\{1, \ldots, n\}$, and $\bar{x}_{T_k}$ corresponds to the vector of component-wise means of the $x_i$, for $i \in T_k$.

In our situation, there is also an underlying true partition of nodes into clusters, given by $\mathcal{C} = \{C_1, \ldots, C_K\}$. Notice that $\mathcal{C} = \hat{\mathcal{T}}$ iff there is a permutation $\pi$ of $\{1, \ldots, K\}$ so that $C_k = \hat{T}_{\pi(k)}$, for $k = 1, \ldots, K$. In general, we use the following measure to quantify the closeness of the outputted partition $\hat{\mathcal{T}}$ and the true partition $\mathcal{C}$: Denote the *clustering error* associated with $\hat{T}_1, \ldots, \hat{T}_K$ as

$$\hat{f} = \min_{\pi} \max_{k} \frac{|C_k \cap \hat{T}^c_{\pi(k)}| + |C^c_k \cap \hat{T}_{\pi(k)}|}{n_k}. \tag{6}$$

The clustering error measures the maximum proportion of nodes in the symmetric difference of $C_k$ and $\hat{T}_{\pi(k)}$.

In many situations, such as ours, there exists population quantities associated with each cluster around which the $x_i$'s are expected to concentrate. Denote these quantities by $m_1, \ldots, m_K$. In our case, $m_k = \mathsf{cent}_{k,\tau}$. If the $x_i$'s, for $i \in C_k$, concentrate well around $m_k$, and the $m_k$'s are sufficiently well separated, then it is expected the $K$-means algorithm recovers the clusters with small error $\hat{f}$.

Denote $X$ as the $n \times K$ matrix with $x_i$'s as rows. In our case, the $x_i = V_{i,\tau}$, and $X = V_\tau$. Further, denote as $M$ the $n \times K$ matrix with the $m_k$'s as rows. In our case, $M = \mathscr{V}_\tau$. Recent results on cluster recovery using the $K$-means algorithm, as given in Kumar and Kannan [15] and Awasthi and Sheffet [3], provide conditions on $X$ and $M$ for the success of $K$-means. The following lemma is implied from Theorem 3.1 in Awasthi and Sheffet [3].

LEMMA 2. *Let $\delta > 0$ be a small quantity. If for each $1 \le k \ne k' \le K$, one has*

$$\|m_k - m_{k'}\| \ge \left(\frac{1}{\delta}\right)\sqrt{K}\|X - M\|\left(\frac{1}{\sqrt{n_k}} + \frac{1}{\sqrt{n_{k'}}}\right) \tag{7}$$

*then the clustering error $\hat{f} = O(\delta^2)$ using the $K$-means algorithm.*

REMARK. In general, minimizing the objective function (5) is not computationally feasible. However, the results in [3, 15] can be extended to partitions $\hat{\mathcal{T}}$ that approximately minimize (5). The condition (7), called the *center separation* condition in [3], provides lower bounds on the pairwise distances between the population cluster centers that depend on the perturbation of data points around the population centers (represented by $\|X - M\|$) and the cluster sizes.

Let

$$1 = \mu_{1,\tau} \ge \cdots \ge \mu_{n,\tau}$$

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, e.g., [27]). As mentioned in the Introduction, in order to control the perturbation of the first $K$ eigenvectors the eigengap, given by $\mu_{K,\tau} - \mu_{K+1,\tau}$, must be adequately large, as noted in [16, 21, 27]. Since $\mathscr{L}_\tau$ has rank $K$ one has $\mu_{K+1,\tau} = 0$. Thus the eigengap is simply $\mu_{K,\tau}$. For our $K$-block SBM framework, the following is an immediate consequence of Lemma 2 and the Davis–Kahan theorem for the perturbation of eigenvectors.

COROLLARY 3. *Let $\tau \ge 0$ be fixed. For the RSC-$\tau$ algorithm the clustering error, given by* (6), *is*

$$O\left(\frac{K\|L_\tau - \mathscr{L}_\tau\|^2}{\mu_{K,\tau}^2}\right).$$

PROOF. Use Lemma 2 with $m_k = \mathsf{cent}_{k,\tau}$, $X = V_\tau$, $M = \mathscr{V}_\tau$, and notice that from Lemma 1 that $\|m_k - m_{k'}\|$ is $\sqrt{1/n_k + 1/n_{k'}}$.

Consequently, using $1/\sqrt{n_k} + 1/\sqrt{n_{k'}} \ge \sqrt{1/n_k + 1/n_{k'}}$ one gets from (7) that if

$$\|V_\tau - \mathscr{V}_\tau\| \le \frac{\delta}{\sqrt{K}}, \tag{8}$$

for some $\delta > 0$, then at most $O(\delta^2)$ fraction of nodes are misclassified with the RSC-$\tau$ algorithm.

From the Davis–Kahan theorem [5], one has

$$\|V_\tau - \mathscr{V}_\tau\| \lesssim \frac{\|L_\tau - \mathscr{L}_\tau\|}{\mu_{K,\tau}}. \tag{9}$$

Consequently, if we take $\delta = (\sqrt{K}\|L_\tau - \mathscr{L}_\tau\|)/\mu_{K,\tau}$ then relation (8) is satisfied using (9). This proves the corollary. □

**3. Regularization in the *K* block SBM.** In this section, we will use Corollary 3 to quantify improvements in clustering performance via regularization. If the number of clusters $K$ is fixed (does not grow with $n$), then the quantity

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

in Corollary 3 provides an insight into the role of the regularization parameter $\tau$. Clearly, an ideal choice of $\tau$ would be the one that minimizes (10). Note, however, that this is not practically possible since $\mathscr{L}_\tau$, $\mu_{K,\tau}$ are not known in advance.

Increasing $\tau$ will ensure that the Laplacian $L_\tau$ will be well concentrated around $\mathscr{L}_\tau$. This is demonstrated in Theorem 4 below. However, increasing $\tau$ also has the effect of decreasing the eigengap, which in this case is $\mu_{K,\tau}$, since the population Laplacian becomes more like a constant matrix upon increasing $\tau$. Thus, the optimum $\tau$ results from the balancing out of these two competing effects.

Independent of our work, a similar argument for the optimum choice of regularization, using the Davis–Kahan theorem, was given in Qin and Rohe [23] for the regulariztion proposed in [7]. They do suggest that choosing $\tau$ to be the average degree (times a multiplicative constant) would be good (see Remark 1 on page 6 of their paper). However, a quantification of the benefit of regularization—in terms of a choice of $\tau$, along with a theoretical demonstration of the clustering improvement resulting from this choice, as given in this section and Section 4—was not provided in this work.

Theorem 4 provides high-probability bounds on the quantity $\|L_\tau - \mathscr{L}_\tau\|$ appearing in the numerator of (10). Previous analysis of the regularization (2), in [7, 23], show high-probability bounds on the aforementioned spectral norm that have a $1/\sqrt{d_{\min,n} + \tau}$ dependence on $\tau$. However, for large $\tau$, the theorem below shows that the behavior is $\sqrt{d_{\max,n}}/(d_{\max,n} + \tau)$. We believe this holds for the regularization (2) as well. Thus, our bounds has a $1/\tau$ dependence on $\tau$, for large $\tau$, as opposed to the $1/\sqrt{\tau}$ dependence shown in [7]. This is crucial since the eigengap $\mu_{K,\tau}$ also behaves like $1/\tau$ for large $\tau$ which implies that (10) converges to a quantity as $\tau$ tends to infinity. In Theorem 6, we provide a bound on this quantity. Our claims regarding improvements via regularization will then follow from comparing this bound with the bound on (10) at $\tau = 0$.

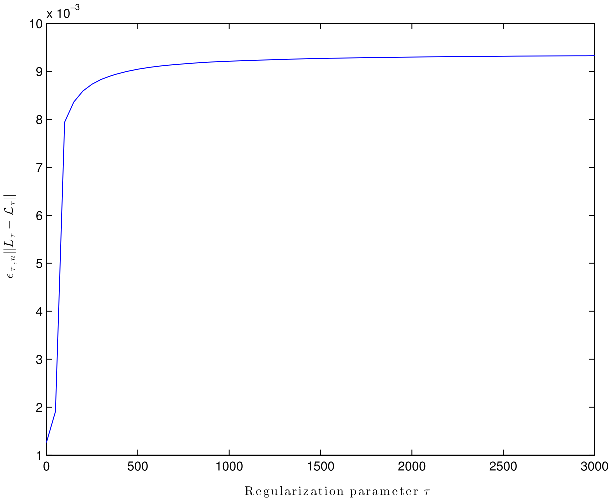


FIG. 1. *Plot of $\varepsilon_{\tau,n}\|L_\tau - \mathscr{L}_\tau\|$ against the regularization parameter $\tau$. Note that the $y$-axis values converges for large $\tau$ thereby demonstrating that $\varepsilon_{\tau,n}$, when viewed as a function of $\tau$, provides the right order of magnitude for large $\tau$.*

THEOREM 4. *With probability at least $1 - 2/n$, for all $\tau$ satisfying*

$$\max\{\tau, d_{\min,n}\} \geq 32\log n, \tag{11}$$

*we have*

$$\|L_\tau - \mathscr{L}_\tau\| \leq \varepsilon_{\tau,n}. \tag{12}$$

*Here*,

$$\varepsilon_{\tau,n} = \begin{cases} \dfrac{10\sqrt{\log n}}{\sqrt{d_{\min,n} + \tau}} & \text{if } \tau \leq 2d_{\max,n}, \\ \dfrac{10\sqrt{\max\{d_{\max,n}, 32\log n\}\log n}}{d_{\min,n} + \tau/2} & \text{if } \tau > 2d_{\max,n}. \end{cases}$$

Notice that in the above result the minimum degree could be of constant order, provided $\tau \gtrsim \log n$. Independent of this work, a similar concentration result in [23] removes the assumption on minimum degree that is traditionally made in spectral clustering results [7, 19, 24] on random graphs. However, as mentioned earlier, their concentration bounds have a $1/\sqrt{\tau}$ dependence on $\tau$, instead of the $1/\tau$ dependence that is shown in the above theorem for $\tau > 2d_{\max,n}$.

Figure 1 demonstrates that $\varepsilon_{\tau,n}$ provides the right order of approximation for large $\tau$ when $n$ is fixed. For the figure, we take $n = 3000$, $K = 2$ and the block probability matrix $B$ given by

$$B = \begin{pmatrix} 0.01 & 0.0025 \\ 0.0025 & 0.003 \end{pmatrix}. \tag{13}$$

We use Theorem 4, along with Corollary 3, to demonstrate improvements from regularization over previous analyses of eigenvector perturbation. Our strategy for this is a follows: Take

$$\delta_{\tau,n} = \frac{\varepsilon_{\tau,n}}{\mu_{K,\tau}}.$$

Notice that from Corollary 3 and Theorem 4, one gets that with probability at least $1 - 2/n$, for all $\tau$ satisfying (11), the clustering error is $O(\delta_{\tau,n}^2)$. Consequently, it is of interest to study the quantity $\delta_{\tau,n}$ as a function of $\tau$. Define

$$\delta_n = \lim_{\tau \to \infty} \delta_{\tau,n}. \tag{14}$$

Although we would have ideally liked to study the quantity,

$$\tilde{\delta}_n = \min_{\max\{\tau, d_{\min,n}\} \gtrsim \log n} \delta_{\tau,n}$$

we study $\delta_n$ since it is easy to characterize as we shall see in Theorem 6 below. Section 5 introduces a data-driven methodology that is based on finding an approximation for $\tilde{\delta}_n$.

Before introducing our main theorem quantifying the performance of RSC-$\tau$ for large $\tau$ we introduce the following definition.

DEFINITION 1. Let $\{\tau_n, n \geq 1\}$ be a sequence of the regularization parameters. For the $K$-block SBM, we say that RSC-$\tau_n$ gives consistent cluster estimates if the error (6) goes 0, with probability tending to 1, as $n$ goes to infinity.

Throughout the remainder of the section, we consider a $K$-block stochastic block model with the following block probability matrix:

$$B = \begin{pmatrix} p_{1,n} & q_n & \cdots & q_n \\ q_n & p_{2,n} & \cdots & q_n \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & q_n & p_{K,n} \end{pmatrix}. \tag{15}$$

Without loss, assume that the *within block probabilities* given by $p_{1,n}, \ldots, p_{K,n}$ satisfy $p_{1,n} \geq p_{2,n} \geq \cdots \geq p_{K,n}$. The *between block probabilities* are taken to be a fixed quantity $q_n$. The number of communities $K$ is assumed to be fixed. Denote $w_k = n_k/n$, for $k = 1, \ldots, K$. The quantity $w_k$ represents the proportion of nodes belonging to the $k$th community. Throughout this section, we assume that $\{\tau_n : n \geq 1\}$ is a sequence of regularization parameters satisfying

$$\frac{(\sum_{k=1}^K 1/w_k) d_{\max,n} \log n}{\tau_n} = o(1). \tag{16}$$

Notice that if the cluster sizes are of the same order, that is $w_k \asymp 1$, then the above condition simply states that $\tau_n$ should grow faster than $d_{\max,n} \log n$.

The following theorem shows that for the stochastic block model regularized spectral clustering would work even when the minimum degree is of constant order. This is an improvement over recent works on unregularized spectral clustering, such as [7, 19, 24], which required the minimum degree to grow at least as fast as $\log n$.

THEOREM 5. *Let the block probability matrix $B$ be as in* (15). *Let $\{\tau_n, n \geq 1\}$ satisfy* (16). *Then RSC-$\tau_n$ gives consistent cluster estimates under the following scenarios*:

(i) *For the $K$-block SBM if $w_k \asymp 1$, for each $k = 1, \dots, K$, and*

$$\frac{(p_{K-1,n} - q_n)^2}{p_{1,n}} \quad \textit{grows faster than} \quad \frac{\log n}{n}. \tag{17}$$

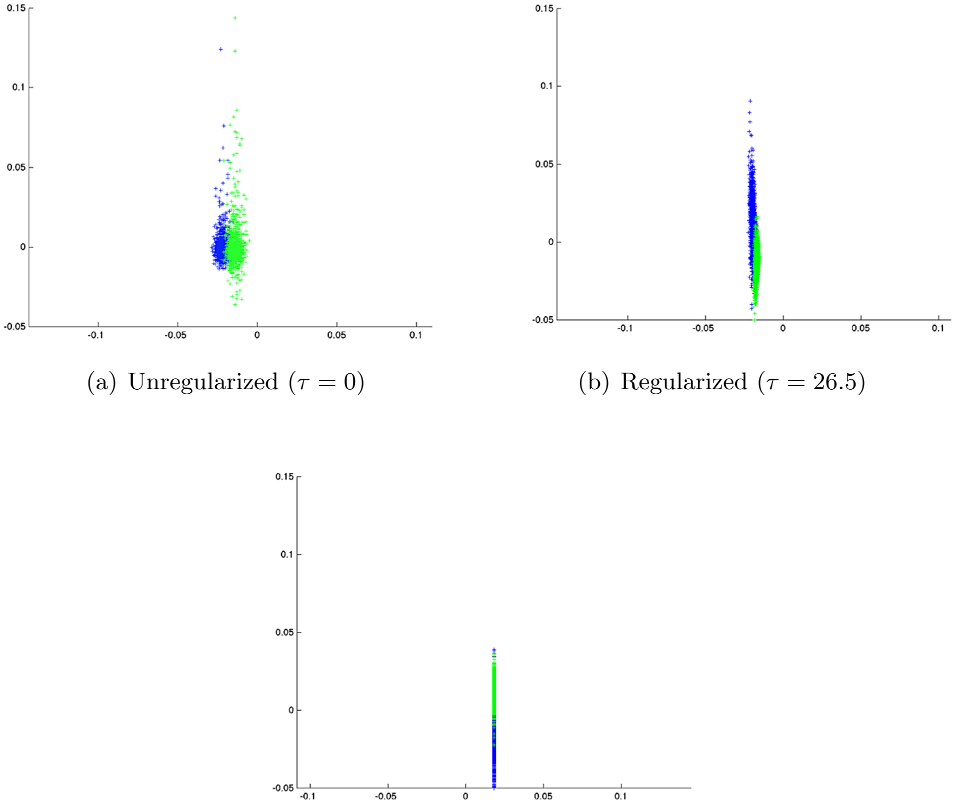
(ii) *For the 2-block SBM if $p_{2,n} = q_n$ and*

$$\frac{(p_{1,n} - q_n)^2}{w_1 p_{1,n} + w_2 q_n} \quad \textit{grows faster than} \quad \frac{\log n}{n(\min\{w_1, w_2\})^2}. \tag{18}$$

REMARK. Regime (i) deals with the situation that the clusters sizes are of the same order of magnitude. Regime (ii), where $p_{2,n} = q_n$ mimics a scenario where there is only one cluster. This is a generalization of the *planted clique* problem where $p_{1,n} = 1$ and $p_{2,n} = q = 1/2$. For the planted clique problem, (18) translates to requiring that $\min\{w_1, w_2\}$ grow faster that $\sqrt{\log n}/\sqrt{n}$ for consistent cluster estimates, which is similar to results in [19].

Notice that in both (17) and (18) the minimum degree could be of constant order. For example, for the two-block SBM if $q_n, p_{2,n} = O(1/n)$ then the minimum degree is of constant order. In this case, ordinary spectral clustering using the normalized Laplacian would perform poorly. RSC performs better since from (17) it only requires that the larger of the two within block probabilities, that is $p_{1,n}$, growing appropriately fast. Figure 2 illustrates this with $n = 3000$ and block probability matrix as in (13). The figure provides the scatter plot of the first two eigenvectors of the unregularized and regularized sample Laplacians. Figure (a) corresponds to the usual spectral clustering, while plots (b) & (c) corresponds to RSC-$\tau$, with $\tau = 26.5$, 3000, respectively. Here, $\tau = 26.5$ was selected using our data-driven methodology for selecting $\tau$ proposed in Section 5. The fraction of mis-classified are 48%, 17.6%, 26.2% for the cases (a), (b), (c), respectively.

From the scatter plots, one sees that there is considerably less scattering for the darker blue points with regularization. This results in improvements in clustering performance. Also, note that the performance in case (c), in which $\tau$ is taken to be very large, is only slightly worse than case (b). For case (c), there is almost no variation in the first eigenvector, plotted along the $x$-axis. This makes sense since



(a) Unregularized ($\tau = 0$)

(b) Regularized ($\tau = 26.5$)

(c) Regularized ($\tau = n$)

FIG. 2. *Scatter plot of first two eigenvectors with $B$ as in* (13). *The $x$, $y$ axes provides values for the first*, *second eigenvectors*, *respectively. The colors corresponds to the cluster memberships of the nodes. Here*, *the block probability matrix $B$ is as in* (13). *Plot* (a) *corresponds to $\tau = 0$.* (b) $\tau = 26.5$, *selected using our data-driven DKest methodology proposed in Section* 5. (c) $\tau = n$.

the first eigenvector is proportional to $(\sqrt{\hat{d}_{1,\tau}}, \dots, \sqrt{\hat{d}_{n,\tau}})$ and for large $\tau$ one has $\sqrt{\hat{d}_{i,\tau}} \approx \sqrt{\tau}$.

It may seem surprising that in Theorem 5, claim (17), the smallest within block probability, that is $p_{K,n}$ does not matter at all. 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.

Theorem 5 follows from the following more general theorem for cluster recovery in the SBM. Theorem 5 and its more general version, Theorem 6 below, will be proved in the supplementary material [13].

Denote $\gamma_{k,n} = n_k(p_{k,n} - q_n)$. The following is a more general result regarding the impact of regularization in the $K$-block SBM.

THEOREM 6. *For the $K$ block SBM, with block probability matrix* (15),

$$\delta_n \asymp \frac{(\tilde{m}_{1,n} m_{1,n} - m_{2,n})}{m_{1,n}} \sqrt{d_{\max,n} \log n}. \tag{19}$$

*Here, $\delta_n$ is given by* (14) *and*

$$m_{1,n} = \sum_{k=1}^{K} \frac{w_k}{\gamma_{k,n}}, \tag{20}$$

$$\tilde{m}_{1,n} = \sum_{k=1}^{K} \frac{1}{\gamma_{k,n}}, \tag{21}$$

$$m_{2,n} = \sum_{k=1}^{K} \frac{w_k}{\gamma_{k,n}^2}. \tag{22}$$

*Further, let $\{\tau_n, n \geq 1\}$ satisfy* (16). *If $\delta_n$ goes to* 0, *as $n$ tends to infinity, then RSC-$\tau_n$ gives consistent cluster estimates.*

Theorem 5 and Theorem 6 will be proved in the supplementary material [13].

## 4. Regularization in SBM with strong and weak 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 Section 5.2. With ordinary spectral clustering, the top two eigenvectors do not discriminate between the two clusters (see Figure 3 for explanation). In fact, 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—where there are main clusters as well as outlier nodes—in the following way: Consider a stochastic block model, as in (15), with $K + K_w$ blocks. In particular, let the block probability matrix be given by

$$B = \begin{pmatrix} B_s & B_{sw} \\ B'_{sw} & B_w \end{pmatrix}, \tag{23}$$

where $B_s$ is a $K \times K$ matrix with $(p_{1,n}, \ldots, p_{K,n})$ in the diagonal and $q_n$ 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 or *strong* clusters, while the bottom $K_w$ blocks corresponds to less well-defined or *weak* clusters.

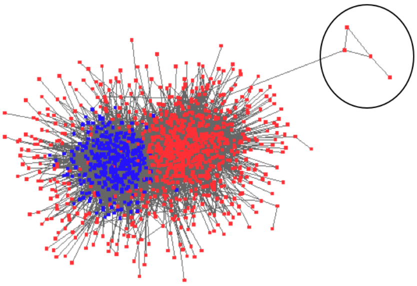


FIG. 3. *Depiction of the political blog network* [1]. *Instead of discriminating between the red and blue nodes, the second eigenvector discriminates the small cluster of* 4 *nodes* (*circled*) *from the remaining. This results in bad clustering performance.*

We now formalize our notion of strong and weak clusters. 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. For ease of analysis, we make the following simplifying assumptions: Assume that $p_{k,n} = p_n^s$, for $k = 1, \dots K$, and that the strong clusters $C_1, \dots, C_K$ have equal sizes, that is, assume $n_k = n^s$ for $k = 1, \dots, K$.

Let $n^w$ be the number of nodes belonging to a weak cluster. In other words, $Kn^s + n^w = n$. We make the following three assumptions:

$$\frac{(p_n^s - q_n)^2}{p_n^s} \quad \text{grows faster than} \quad \frac{\log n}{n}, \tag{24}$$

$$n^w = O\Bigl(\sqrt{p_n^s n \log n}\Bigr). \tag{25}$$

Assumption (24) ensures recovery of the strong clusters if there were no nodes belonging to weak clusters (See Theorem 5 or McSherry [19], Corollary 1). Assumption (25) simply states that the total number of nodes belonging to the weak clusters does not grow faster than $\sqrt{p_n^s n \log n}$.

We only assume that the rank of $B_s$ is $K$. Thus, the rank of $B$ is at least $K$. 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 could be as high as $n^w$. We do not even place any restriction on the sizes of each weak cluster. Indeed, we even entertain the case that each of the $K_w$ clusters has one node. Consequently, we are only interested in recovering the strong clusters.

Theorem 7 presents our theorem for the recovery of the $K$ strong clusters using the RSC-$\tau_n$ Algorithm, with $\{\tau_n, n \geq 1\}$, satisfying

$$\frac{np_n^s \log n}{\tau_n} = o(1). \tag{26}$$

In other words, the regularization parameter is taken to grow faster than $np_n^s \log n$, where notice that $np_n^s$ is of the same order of the expected maximum degree of the graph. Let $\hat{T}_1, \dots, \hat{T}_K$ be the clusters outputted from the RSC-$\tau_n$ algorithm. Let

$$\hat{f} = \min_{\pi} \max_{k} \frac{|C_k \cap \hat{T}^c_{\pi(k)}| + |C^c_k \cap \hat{T}_{\pi(k)}|}{n_k},$$

be as in (6). Notice that the clusters $C_1, \dots, C_K$ do not form a partition of $\{1, \dots, n\}$, while the estimates $\hat{T}_1, \dots, \hat{T}_K$ do. However, since the total size of the weak clusters is $o(n)$ from assumption (25), the quantity $\hat{f}$ still represents a good measure of accuracy of cluster recovery.

THEOREM 7. *Let assumptions* (24) *and* (25) *be satisfied. If* $\{\tau_n, n \geq 1\}$ *satisfies* (26), *then the clustering error* $\hat{f}$ *for RSC-*$\tau_n$ *goes to zero with probability tending to one.*

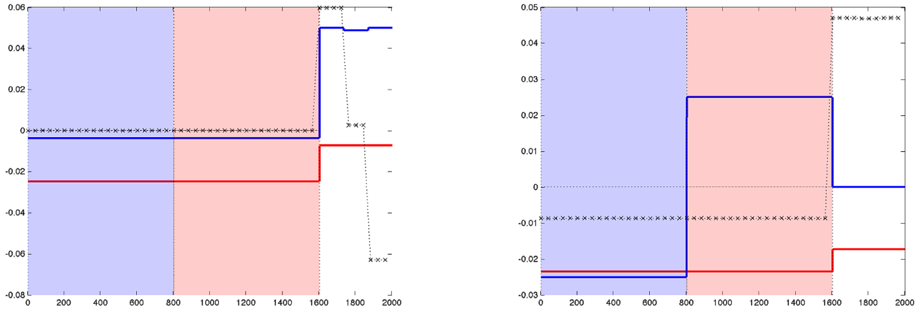
For convenience, we relegate the proof of the theorem to the supplementary material [13]. The theorem states that under assumptions (24) and (25) one can get the same results with regularization that one would get if the nodes belonging to the weak clusters were not present.

Spectral clustering (with $\tau = 0$) may fail under the above assumptions. This is elucidated in Figure 4. Here, $n = 2000$ and there are two strong clusters ($K = 2$) and 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 $B_s$ and $B_w$ are as in (27) and $B_{sw}$ is a matrix with all entries 0.015.

$$B_s = \begin{pmatrix} 0.025 & 0.015 \\ 0.015 & 0.025 \end{pmatrix}, \qquad B_w = \begin{pmatrix} 0.007 & 0.015 & 0.015 \\ 0.015 & 0.0071 & 0.015 \\ 0.015 & 0.015 & 0.0069 \end{pmatrix}. \tag{27}$$

The nodes in the weak clusters have relatively lower degrees, and consequently, cannot be recovered. Figures 4(a) and 4(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, e.g., [27] 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 4(a), not even

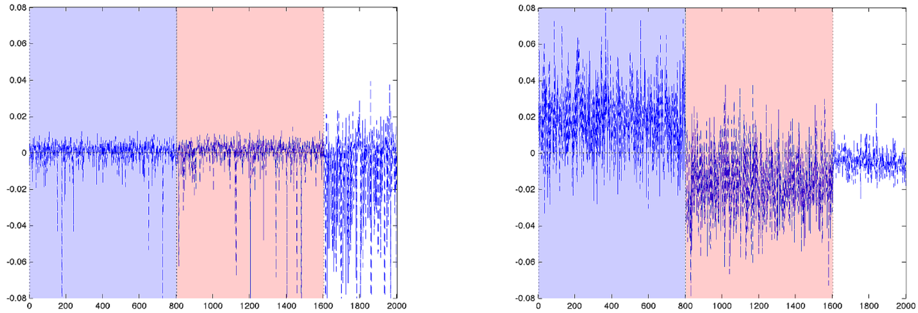


(a) Unregularized (b) Regularized

FIG. 4. *First three population eigenvectors corresponding to $B_s$ and $B_w$ in* (27). *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 $n$. 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.*

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.

In Figure 5(a) and 5(b), we show the second sample eigenvector for the two cases in Figure 4(a) and 4(b). Note, we do not show the first sample eigenvector since from Figure 4(a) and 4(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) Unregularized (b) Regularized

FIG. 5. *Second sample eigenvector corresponding to situation in Figure* 4. *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) *and* (b), *the blue line correspond to the second eigenvector of the respective sample Laplacian matrices.*

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. ***DKest*: Data dependent choice of $\tau$.**

The results Sections 3 and 4 theoretically examined the gains from regularization for large values of regularization parameter $\tau$. Those results do not rule out the possibility that intermediate values of $\tau$ may lead to better clustering performance. In this section, we propose a data dependent scheme to select the regularization parameter. We compare it with the scheme in [8] that uses the Girvan–Newman modularity [6]. We use the measure $\hat{f}$, given by (6), to quantify the closeness of the estimated clusters to the true clusters.

Our scheme works by directly estimating the quantity in (10) by providing, for each $\tau$ in grid, an estimate $\hat{\mathscr{L}}_\tau$ of $\mathscr{L}_\tau$. The intuition behind this estimate is as follows: Given knowledge of the true clusters, one can get estimates of the within and between block probabilities, assuming that the graph is drawn from an SBM. However, since the clusters are not known, we use the clusters provided by RSC-$\tau$ algorithm and then use these to estimate the within and between block probabilities. The Laplacian $\hat{\mathscr{L}}_\tau$ is simply the Laplacian corresponding to these estimated probabilities.

In particular, let $\hat{C}_{1,\tau}, \ldots, \hat{C}_{K,\tau}$ be the estimates of the clusters $C_1, \ldots, C_K$ produced from running RSC-$\tau$. The estimate $\hat{\mathscr{L}}_\tau$ is taken as the population regularized Laplacian corresponding to an estimated block probability matrix $\hat{B}$ and clusters $\hat{C}_{1,\tau}, \ldots, \hat{C}_{K,\tau}$. More specifically, the $(k_1, k_2)$th entry of $\hat{B}$ is taken as

$$\hat{B}_{k_1,k_2} = \frac{\sum_{i \in \hat{C}_{k_1,\tau}, j \in \hat{C}_{k_2,\tau}} A_{ij}}{|\hat{C}_{k_1,\tau}||\hat{C}_{k_2,\tau}|}. \tag{28}$$

The above is simply the proportion of edges between the nodes in the cluster estimates $\hat{C}_{k_1,\tau}$ and $\hat{C}_{k_2,\tau}$. The following statistic is then considered:

$$DKest_\tau = \frac{\|L_\tau - \hat{\mathscr{L}}_\tau\|}{\mu_K(\hat{\mathscr{L}}_\tau)}, \tag{29}$$

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

We compare the above to the scheme that uses Girvan–Newman modularity [6, 20], as suggested in [8]. For a particular $\tau$ in the grid, the Girvan–Newman modularity is computed for the clusters outputted using the RSC-$\tau$ algorithm. The $\tau$ that maximizes the modularity value over the grid is then chosen.

Notice that the best possible choice of $\tau$ would be the one that simply minimizes the clustering error $\hat{f}$ over the selected grid. However, this cannot be computed in practice since calculation of $\hat{f}$ 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.

TABLE 1
*Performance of spectral clustering as a function of $\tau$ for SBM for $\lambda$ values of* 30, 20 *and* 10

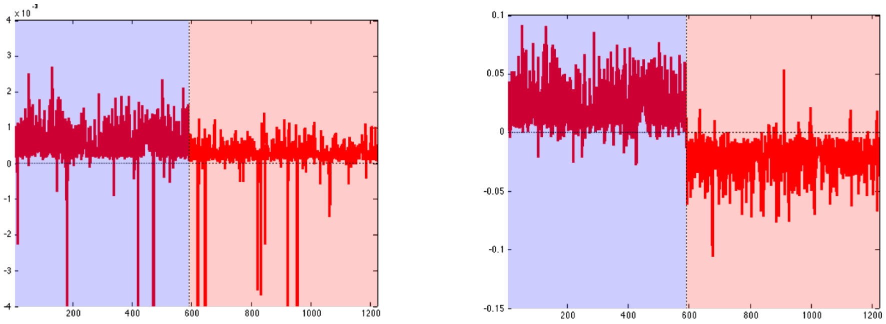
| | | | | | Mean $\hat{f}$ (std. $\hat{f}$) | | |
|---|---|---|---|---|---|---|---|
| **$n$** | **$\lambda$** | **$K$** | **$w$** | **$\beta$** | **Oracle** | ***DKest*** | **Girvan–Newman** |
| 1200 | 20 | 3 | [1 5 5] | 0.70 | 0.044<br>(0.015) | 0.078<br>(0.028) | 0.108<br>(0.032) |
| 1200 | 20 | 3 | [1 5 10] | 0.30 | 0.000<br>(0.001) | 0.003<br>(0.003) | 0.010<br>(0.006) |
| 1200 | 30 | 3 | [1 5 5] | 0.70 | 0.014<br>(0.003) | 0.019<br>(0.004) | 0.026<br>(0.007) |
| 1200 | 30 | 3 | [1 5 10] | 0.70 | 0.006<br>(0.003) | 0.016<br>(0.010) | 0.023<br>(0.011) |
| 2000 | 10 | 3 | [1 5 10] | 0.50 | 0.007<br>(0.003) | 0.077<br>(0.066) | 0.097<br>(0.065) |
| 2000 | 10 | 3 | [1 5 10] | 0.70 | 0.008<br>(0.005) | 0.277<br>(0.129) | 0.119<br>(0.119) |
| 2000 | 20 | 3 | [1 5 10] | 0.70 | 0.012<br>(0.013) | 0.068<br>(0.030) | 0.045<br>(0.020) |
| 2000 | 20 | 4 | [1 5 5 10] | 0.60 | 0.001<br>(0.001) | 0.018<br>(0.011) | 0.021<br>(0.017) |
| 2000 | 30 | 4 | [1 5 5 10] | 0.60 | 0.001<br>(0.001) | 0.008<br>(0.008) | 0.013<br>(0.008) |

5.1. *Simulation results*. Table 1 provides results comparing the three schemes, namely, *DKest*, Girvan–Newman and ‘oracle’ schemes. We perform simulations following the pattern of [2]. In 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

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

Here, the vector $w = (w_1, \dots, w_K)$, which are the *inside weights*, denotes the relative degrees of nodes within the communities. Further, the quantity $\beta$, 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 scalar parameter $\mathsf{fac}$ is chosen so that the average expected degree of the graph is equal to $\lambda$.

Table 1 compares the two methods of choosing the best $\tau$ for various choices of $n, K, \beta, w$ and $\lambda$. For each of these choices of parameters, a random graph is generated according to an SBM with block probability matrix given by (28). Regularized spectral clustering is conducted for $\tau$ ranging from 0 to $\lambda$, in steps of



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

FIG. 6. *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.*

0.5. Next, the regularization parameter $\tau$ is selected using the *DKest* and Girvan–Newman methodologies, and the corresponding clustering error $\hat{f}$ is computed. For comparison purposes, we also compute $\hat{f}$ using the optimal, though impractical, oracle scheme. The above process is repeated 10 times, and the mean and standard deviations of the clustering error $\hat{f}$ using the oracle, *DKest* and Girvan–Newman procedures are reported in Table 1.

In general, we see that the *DKest* selection procedure performs comparably, 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.

5.2. *Analysis of the political blogs dataset.* Here, we investigate the performance of *DKest* 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 other.

The data set has 1222 nodes with an average degree of 27. Spectral clustering ($\tau = 0$) 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 *DKest* procedure selected $\tau = 2.25$, with an accuracy of 81%. The Girvan–Newman (GN) procedure, in this case, outperforms the *DKest* procedure providing the same accuracy as the oracle procedure. As predicted by our theory, the performance becomes insensitive for large $\tau$. In this case, 70% of the nodes are correctly clustered for large $\tau$.

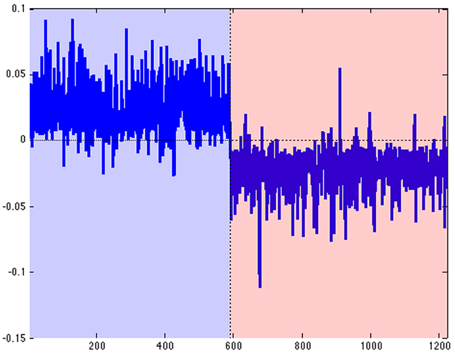


FIG. 7. *Third eigenvector of the unregularized Laplacian.*

We remark that the *DKest* procedure does not perform as well as the GN procedure most likely because our estimate $\hat{\mathscr{L}}_\tau$ in (29) assumes that the data is generated from an SBM, which is a poor model for the data due to the large heterogeneity in the node degrees. A better model for the data would be the degree corrected stochastic block model (D-SBM) proposed by Karrer and Newman [14]. If we use D-SBM based estimates in *DKest*, then the selection of $\tau$ matches that of the GN Newman and the oracle procedure. See Section 6 for a discussion on this.

The results of Section 4 also explain why unregularized spectral clustering performs badly (see Figure 3). The first eigenvector in both cases (regularized and unregularized) does not discriminate between the two clusters. In Figure 6, 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 eigenvector in the unregularized case that distinguishes between the clusters, as shown in Figure 7.

**6. Discussion.** The paper provides a theoretical justification for regularization. In particular, we show why choosing a large regularization parameter can lead to good results. The paper also partly explains empirical findings in Amini et al. [2] showing that the performance of regularized spectral clustering becomes insensitive for larger values of regularization parameters. It is unclear at this stage whether the benefits of regularization, resulting from the trade-offs between the eigengap and the concentration bound, hold for the regularization in [7, 23] as they hold for the regularization in Amini et al. [2] (as demonstrated in Sections 3 and 4).

Recent results in [17] demonstrates that it is possible to remove the $\sqrt{\log n}$ appearing in the numerator of the concentration bound (12). Although the concentration bound of [17] has a $1/\sqrt{\tau}$ dependence on $\tau$, it is very likely that one could combine the techniques in [17], along with our results, to improve the concentration bound in (12).

Even though our theoretical results focus on larger values of the regularization parameter it is very likely that intermediate values of $\tau$ produce better clustering performance. Consequently, we propose a data-driven methodology for choosing the regularization parameter. We hope to quantify theoretically the gains from using intermediate values of the regularization parameter in a future work.

We remark that community detection in not all datasets can be improved via regularization. Note, our theoretical results do not demonstrate improvement using regularization for block models where the *within block probabilities* are all equal, and the *between block probabilities* are equal as well. This is confirmed by simulation results we conducted. This suggests that regularization only has an effect in block models with heterogeneous degrees. It would be interesting to characterize the scenarios where regularization has a dramatic effect.

For the extension of the SBM proposed in Section 4, if the rank of $B$, given by (23), is $K$ then the model encompasses specific degree-corrected stochastic block models (D-SBM) [14] where the edge probability matrix takes the form

$$P = \Theta Z B Z' \Theta.$$

Here, $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ models the heterogeneity in the degrees. In particular, consider a $K$-block D-SBM with $0 < \theta_i \leq 1$, for each $i$. Assume that $\theta_i = 1$ for the most of the 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. Assumption (25) states that the number of nodes $i$ with $\theta_i < 1$ is $o(n)$. It would be interesting to investigate the effect of regularization in more general versions of the D-SBM, especially where there are high as well as low degree nodes.

The *DKest* methodology for choosing the regularization parameter works by providing estimates of the population Laplacian assuming that the data is drawn from an SBM. From our simulations, it is seen that the performance of *DKest* does not change much if we take the matrix norm in the numerator of (29) to be the Frobenius norm, which is much faster to compute.

It is seen that the performance of *DKest* improves for the political blogs data set by taking $\hat{\mathscr{L}}_\tau$ to be the estimate assuming that the data is drawn from the more flexible D-SBM. Indeed, if we take $\hat{\mathscr{L}}_\tau$ to be such an estimate then the performance of *DKest* is seen to be as good as the oracle scheme (and the GN scheme) for this data set. We describe how we construct this estimate in the supplementary material [13].

## APPENDIX: 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 $L_\tau, \mathscr{L}_\tau$, respectively. For ease of notation, we remove the subscript $\tau$ from the various matrices such as

$L_\tau, \mathscr{L}_\tau, A_\tau, D_\tau, \mathscr{D}_\tau$. We also remove the subscript $\tau$ 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 $i = 1, \dots, n$ and $d_{i,0}$ for $i = 1, \dots, n$.

The following lemma provides high probability bounds on the degree. Let $\tau$ be of the form $\max\{d_{\min,n}, c\log n\}$ and $\delta_{i,c} = \max\{d_{i,0}, c\log n\}$.

LEMMA 8. *On a set $E_1$ of probability at most $1 - 2/n^{c_1 - 1}$, one has*

$$|\hat{d}_{i,\tau} - d_{i,\tau}| \le c_2\sqrt{\delta_{i,c}\log n} \qquad \text{for each } i = 1, \dots, n,$$

*where $c_1 = 0.5c_2^2/(1 + c_2/\sqrt{c})$.*

The above is proved in the supplementary material [13].

**A.1. Concentration of Laplacian.** Below we provide the proof of Theorem 4. Throughout this section we assume that the quantities $c, c_2$ appearing in Lemma 8 are given by $c = 32$ and $c_2 = 2\sqrt{2}$. Notice that this makes $c_1 > 2$, where $c_1$ as in Lemma 8.

From Lemma 8, with probability at least $1 - n^{-1}$,

$$\max_i |\hat{d}_i - d_i|/d_i \le \max_i c_2\sqrt{\delta_{i,c}\log n}/d_i.$$

We claim that the right-hand side of the above is at most $1/2$. To see this, notice that

$$\begin{aligned}\sqrt{\delta_{i,c}\log n}/d_i &\le \sqrt{\delta_{i,c}\log n}/\delta_{i,c}\\ &= \sqrt{\log n}/\sqrt{\delta_{i,c}}\\ &\le 1/\sqrt{c}.\end{aligned}$$

Here, the first inequality follows from noting that $d_i = d_{i,0} + \tau$, which is at most $\max\{d_{i,0}, c\log n\}$, using $\tau \ge c\log n$. The third inequality follows from using $\delta_{i,c} \ge c\log n$. Consequently, $\max_i |\hat{d}_i - d_i|/d_i \le 1/2$ using $c_2 = 2\sqrt{2}$ and $c = 32$.

PROOF OF THEOREM 4. Our proof has parallels with the proof in [22]. Write $\tilde{L} = \mathscr{D}^{-1/2}A\mathscr{D}^{-1/2}$. Then

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

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

$$\begin{aligned}\|L - \tilde{L}\| &\le \|L - FL\| + \|FL - \tilde{L}\|\\ &\le \|I - F\|\|L\| + \|F\|\|L\|\|I - F\| \qquad (30)\\ &\le \|I - F\|\big(2 + \|I - F\|\big).\end{aligned}$$

Notice that

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

Further, using $\max_i |\hat{d}_i - d_i|/d_i \le 1/2$, and the fact that $\sqrt{1+x} - 1 \le x$ for $x \in [-3/4, 3/4]$, as in [22], one gets that

$$\|F - I\| \le c_2 \frac{\max_i \sqrt{\delta_{i,c} \log n}}{d_i}$$

with high probability. Consequently, using (30), one gets that

$$\|L - \tilde{L}\| \le c_2 \max_i \frac{\sqrt{\delta_{i,c} \log n}}{d_i} \left(2 + c_2 \max_i \frac{\sqrt{\delta_{i,c} \log n}}{d_i}\right) \tag{31}$$

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

$$\max_i \frac{\sqrt{\delta_{i,c}}}{d_i} \le \tilde{\varepsilon}_{\tau,n} = \begin{cases} \dfrac{1}{\sqrt{d_{\min,n} + \tau}} & \text{if } \tau \le 2d_{\max,n}, \\ \dfrac{\sqrt{\max\{d_{\max,n}, c \log n\}}}{d_{\min,n} + \tau/2} & \text{if } \tau > 2d_{\max,n}. \end{cases}$$

To see this notice, that $\delta_{i,c} \le d_{i,0} + \tau = d_i$, using $\max\{\tau, d_{i,0}\} \ge c \log n$. Consequently, $\sqrt{\delta_{i,c}}/d_i \le 1/\sqrt{d_{i,0} + \tau}$, which is at most $1/\sqrt{d_{\min,n} + \tau}$.

Further,

$$\max_i \frac{\sqrt{\delta_{i,c}}}{d_i} = \max\left\{\max_i \left\{\frac{\sqrt{d_{i,0}}}{d_{i,0} + \tau}, \frac{\sqrt{c \log n}}{d_{i,0} + \tau}\right\}\right\}.$$

Now,

$$\max_i \left\{\frac{\sqrt{d_{i,0}}}{d_{i,0} + \tau}\right\} \le \frac{\sqrt{d_{\max,n}}}{d_{\max,n} + \tau} \qquad \text{for } \tau > d_{\max,n}$$

and $d_{i,0} + \tau d_{\min,n} + \tau$. Consequently,

$$\max_i \frac{\sqrt{\delta_{i,c}}}{d_i} \le \frac{\sqrt{\max\{d_{\max,n}, c \log n\}}}{d_{\min,n} + \tau}$$

for $\tau > d_{\max,n}$. Consequently, from (31), one gets that

$$\|L - \tilde{L}\| \le c_2 \tilde{\varepsilon}_{\tau,n} \sqrt{\log n}(2 + c_2/\sqrt{c}) \tag{32}$$

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

Next, we bound $\|\tilde{L} - \mathscr{L}\|$. We get high probability bounds on this quantity using results in [18, 22]. In particular, as in [22],

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

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

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

Further, $\|Y_{ij}\| \leq 1/(d_{\min,n} + \tau)$. Let $\sigma^2 = \|\sum_{i \leq j} E(Y_{ij}^2)\|$. We claim that $\sigma^2 \leq \tilde{\varepsilon}_{\tau,n}^2$. As in [22], page 15, notice that

$$\sum_{i \leq j} E(Y_{ij}^2) = \sum_{i=1}^{n} \frac{1}{d_{i,0} + \tau} \left( \sum_{j=1}^{n} \frac{P_{ij}(1 - P_{ij})}{d_{j,0} + \tau} \right) e_i e_i^T. \tag{33}$$

Clearly,

$$\left( \sum_{j=1}^{n} \frac{P_{ij}(1 - P_{ij})}{d_{j,0} + \tau} \right) \leq \frac{d_{i,0}}{d_{\min,n} + \tau}.$$

Consequently, for each $i$ the right-hand side of (33) is at most $1/(d_{\min,n} + \tau)$ leading to the fact that $\sigma^2 \leq 1/(d_{\min,n} + \tau)$.

For $\tau > 2d_{\max,n}$, we can get improvements in the bound for $\sigma^2$. By using the fact that $d_{j,0} + \tau > d_{\max,n} + \tau/2$ for $\tau > 2d_{\max,n}$, one gets that

$$\left( \sum_{j=1}^{n} \frac{P_{ij}(1 - P_{ij})}{d_{j,0} + \tau} \right) \leq \frac{d_{i,0}}{d_{\max,n} + \tau/2}$$

for $\tau > 2d_{\max,n}$. Consequently, using $d_{i,0}/(d_{i,0} + \tau) \leq d_{\max,n}/(d_{\max,n} + \tau)$, one gets that $\sigma^2 \leq d_{\max,n}/(d_{\max,n} + \tau/2)^2$ for $\tau > 2d_{\max,n}$. Thus, $\sigma \leq \tilde{\varepsilon}_{\tau,n}$.

Applying Corollary 4.2 in [18], one gets

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

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

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

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

$$\|\tilde{L} - \mathscr{L}\| \leq \sqrt{2c_1 \log n}\, \tilde{\varepsilon}_{\tau,n}. \tag{34}$$

As a result, combining (32) and (34), one gets that with probability at least $1 - 2/n^{c_1 - 1}$, one has

$$\|L_\tau - \mathscr{L}_\tau\| \leq \sqrt{\log n}\, \tilde{\varepsilon}_{\tau,n} \left[ \sqrt{2c_1} + c_2 \left( 2 + (c_2/\sqrt{c}) \right) \right].$$

Substituting the values of $c_2$, $c$, and noting that $c_1 > 2$ one gets the expression in the theorem. □

**Acknowledgments.** A. Joseph would like to thank Sivaraman Balakrishnan and Puramrita Sarkar for some very helpful discussions, Arash A. Amini for sharing the code used in the work [2] and Liza Levina for pointing out a mistake in an earlier version of this draft.

## SUPPLEMENTARY MATERIAL

**Supplementary Material: Supplement to "Impact of regularization on spectral clustering"** (DOI: 10.1214/16-AOS1447SUPP; .pdf). The supplementary file contains the proof of the claims in the paper that were not included in the main body.

## REFERENCES

[1] ADAMIC, L. A. and GLANCE, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the* 3*rd International Workshop on Link Discovery* 36–43. ACM, New York.

[2] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859

[3] AWASTHI, P. and SHEFFET, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Lecture Notes in Computer Science* **7408** 37–49. Springer, Heidelberg. MR3003539

[4] BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

[5] BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[6] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[7] CHAUDHURI, K., CHUNG, F. and TSIATAS, A. Spectral clustering of graphs with general degrees in the extended planted partition model. *J. Mach. Learn. Res.* **2012** 1–23.

[8] CHEN, A., AMINI, A., BICKEL, P. and LEVINA, L. (2012). Fitting community models to large sparse networks. In *Joint Statistical Meetings, San Diego.*

[9] DHILLON, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. Seventh ACM SIGKDD Inter. Conf. on Know. Disc. and Data Mining* 269–274. ACM, New York.

[10] FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990

[11] HAGEN, L. and KAHNG, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **11** 1074–1085.

[12] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

[13] JOSEPH, A. and YU, B. (2016). Supplement to "Impact of regularization on spectral clustering." DOI:10.1214/16-AOS1447SUPP.

[14] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107. MR2788206

[15] KUMAR, A. and KANNAN, R. (2010). Clustering with spectral norm and the $k$-means algorithm. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science FOCS* 2010 299–308. IEEE Computer Soc., Los Alamitos, CA. MR3025203

[16] KWOK, T. C., LAU, L. C., LEE, Y. T., OVEIS GHARAN, S. and TREVISAN, L. (2013). Improved Cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. In *STOC'*13—*Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 11–20. ACM, New York. MR3210762

[17] LE, C. M. and VERSHYNIN, R. (2015). Concentration and regularization of random graphs. Available at arXiv:1506.00669.

[18] MACKEY, L., JORDAN, M. I., CHEN, R. Y., FARRELL, B. and TROPP, J. A. (2012). Matrix concentration inequalities via the method of exchangeable pairs. Available at arXiv:1201.6002.

[19] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742

[20] NEWMAN, M. E. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys*. *Rev*. *E* **69** 026113.

[21] NG, A. Y., JORDAN, M. I., WEISS, Y. et al. (2002). On spectral clustering: Analysis and an algorithm. *Adv*. *Neural Inf*. *Process*. *Syst*. **2** 849–856.

[22] OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at arXiv:0911.0600.

[23] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. Available at arXiv:1309.4111.

[24] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann*. *Statist*. **39** 1878–1915. MR2893856

[25] SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **22** 888–905.

[26] SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J*. *Amer*. *Statist*. *Assoc*. **107** 1119–1128. MR3010899

[27] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat*. *Comput*. **17** 395–416. MR2409803

@WALMARTLABS
SAN BRUNO, CALIFORNIA 94066
USA
E-MAIL: AJoseph0@walmartlabs.com

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: binyu@stat.berkeley.edu