Randomized Spectral Clustering in Large-Scale Stochastic Block Models

Hai Zhang, Xiao Guo, and Xiangyu Chang

Department of Modern Statistics, School of Mathematics, Northwest University, Xi’an, China; Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi’an Jiaotong University, Xi’an, China

ABSTRACT

Spectral clustering has been one of the widely used methods for community detection in networks. However, large-scale networks bring computational challenges to the eigenvalue decomposition therein. In this paper, we study the spectral clustering using randomized sketching algorithms from a statistical perspective, where we typically assume the network data are generated from a stochastic block model that is not necessarily of full rank. To do this, we first use the recently developed sketching algorithms to obtain two randomized spectral clustering algorithms, namely, the random projection-based and the random sampling-based spectral clustering. Then we study the theoretical bounds of the resulting algorithms in terms of the approximation error for the adjacency matrix, the misclassification error, and the estimation error for the link probability matrix. It turns out that, under mild conditions, the randomized spectral clustering algorithms lead to the same theoretical bounds as those of the original spectral clustering algorithm. We also extend the results to degree-corrected stochastic block models. Numerical experiments support our theoretical findings and show the efficiency of randomized methods. A new R package called Rclus is developed and made available to the public. Supplementary materials for this article are available online.

1. Introduction

Extraordinary amounts of data are being collected in the form of arrays across many scientific domains, including sociology, physics, and biology, among others. In particular, network data and network data analysis have received a lot of attention because of their wide-ranging applications in these areas (Kolaczyk 2009; Goldenberg et al. 2010; Newman 2018). Community detection is one of the fundamental problems in network analysis, where the goal is to find groups of nodes that are, in some sense, more similar to each other than to the other nodes. Past decades have seen various procedures on community detection including modularity maximization, spectral clustering, likelihood methods, semidefinite programming, among others; see Abbe (2018) for a recent survey. However, large networks, say, networks with millions of nodes, bring great challenges to these community detection procedures despite the increasing computational power. Taking the spectral clustering that we will focus on in this paper as an example, the full eigenvalue decomposition therein is time demanding when the dimension becomes large.

Randomization has become one popular method for modern large-scale data analysis; see Mahoney (2011), Drineas and Mahoney (2016), and references therein. The general idea is that depending on the problem of interest, one uses a degree of randomness to construct a small “sketch” of the full dataset, and then uses the resulting sketched data instead to reduce the computational burden. Random projection and random sampling are the two general approaches to obtain such a sketch matrix. Roughly speaking, random projection reduces the computational cost by projecting the data matrix to a smaller dimensional space in order to approximate the data. While random sampling algorithms lighten the computational burden by sampling and rescaling the data in some manner. The randomization techniques have been applied to the least squares regression (Drineas, Mahoney, and Muthukrishnan 2006; Drineas et al. 2011, 2012), and the low-rank matrix approximation (Mahoney and Drineas 2009; Halko, Martinsson, and Tropp 2011; Witten and Candès 2015; Martinsson 2016), among many others. Most works in this area were analyzed from an algorithmic perspective, where the randomized algorithm could lead to approximately as good performance as the full data at hand does for some problems of interest. However, from a statistical perspective, the aim is not only to obtain randomized algorithms which perform well on a particular dataset but also to understand how well they perform under some underlying mechanisms. In the context of regression, there have been a few works that study the randomized algorithms under underlying regression models – for example, the ordinary linear regression (Ma, Mahoney, and Yu 2015; Raskutti and Mahoney 2016; Wang, Yang, and Stufken 2019a), the logistic regression (Wang, Zhu, and Ma 2018; Wang 2019), the ridge regression (Wang, Gittens, and Mahoney 2017), the constrained regressions (Pilanci and Wainwright 2016, 2017), and the spatial autoregressive (SAR) models (Zhou et al. 2017; Li and Kang 2019), among others.
Just like they have studied how well the randomized algorithms can estimate the underlying regression model, it is natural and important to study how well we can use the randomization techniques to detect the communities in a "true" network model. The stochastic block model (SBM) (Holland, Laskey, and Leinhardt 1983) is a simple but expressive network model that captures the community structure of networks observed in the real world. In an SBM, nodes are partitioned into several distinct communities and conditioned on the underlying community assignments, the edges are generated independently according to the community membership of their end nodes. Nodes within the same community are generally more likely to be connected than the other nodes. The SBM is popular among statisticians because it can be rigorously studied coupling with various network community detection procedures; see Abbe (2018) for an excellent review.

In this work, we focus on studying how randomization can be used to reduce the computational cost of spectral clustering, and understanding how well the resulting randomized spectral clustering algorithms perform under the SBMs. Spectral clustering is a popular and simple algorithm for clustering which consists of the following two steps. One first conducts the eigenvalue decomposition of the adjacency matrix or the Laplacian matrix and then runs the $k$-means on several leading eigenvectors to obtain the nodes clusters or communities (Von Luxburg 2007). It is well known that the full eigenvalue decomposition in the first step generally requires $O(n^3)$ time where $n$ denotes the number of nodes, which is time demanding when $n$ becomes huge. Regardless of the computational issues, it has been shown to enjoy good theoretical properties within the SBM framework; see, Rohe, Chatterjee, and Yu (2011), Choi, Wolfe, and Airoldi (2012), Qin and Rohe (2013), Lei and Rinaldo (2015), Sarkar and Bickel (2015), Joseph and Yu (2016), Su, Wang, and Zhang (2019), Yang et al. (2020), Tang, Cape, and Priebe (2021), Deng, Ling, and Strohmer (2021) and Levin et al. (2021), among many others. Facing large networks, it is thus, desirable to study whether these properties would retain under certain randomization of the algorithms. In this article, we use the idea of randomization to obtain two kinds of randomized spectral clustering algorithms; namely, the random projection-based and the random sampling-based spectral clustering, and in particular, we study their theoretical properties under the SBMs.

We focus on the adjacency matrix $A$ of the network. The random projection-based method is motivated as follows. Note that the adjacency matrix inherits a low-rank structure approximately since it is assumed to be sampled from a SBM (Rohe, Chatterjee, and Yu 2011; Lei and Rinaldo 2015). Therefore, if one can make use of such low-rank structure to derive a matrix with a lower dimension which captures the essential information of $A$, then the eigenvalue decomposition of this matrix can help to derive that of $A$, which in turn reduces the computational cost. Indeed, the recently developed randomized low-rank matrix approximation algorithms provide a powerful tool for performing such low-rank matrix approximation (Halko, Martinsson, and Tropp 2011; Witten and Candès 2015; Martinsson 2016). Specifically, these techniques utilize some amount of randomness to compress the columns and rows of $A$ to $l$ ($l \ll n$) linear combinations of the columns and rows of $A$. The eigenvalue decomposition on the resulting $l$-dimensional matrix can be largely reduced since $l$ is far smaller than $n$. The random projection-based spectral clustering refers to the original spectral clustering with its first step replaced by the randomized eigenvalue decomposition. On the other hand, the computational cost of the original spectral clustering can be reduced via the random sampling. Note that we only need to find a few leading eigenvectors of $A$, which can be obtained using many fast iterative methods, such as the orthogonal iteration and Lanczos iteration; see Baglama and Reichel (2005); Calvetti, Reichel, and Sorensen (1994), among others. And it is well known that the time complexity of iterative algorithms is in direct proportion to the number of nonzero elements of $A$ multiplied by the number of iterations. Therefore, if we sample the elements of $A$ in some way to obtain a sparser matrix, then the time for computing its leading eigenvectors will be largely reduced. There have been a few works on the randomized matrix sparsification; see Gittens and Tropp (2009), Achlioptas and McSherry (2007), Arora, Hazan, and Kale (2006) and Li, Levina, and Zhu (2020b), among others. In particular, Li, Levina, and Zhu (2020b) apply the sampling technique to study the network cross-validation problem. In this work, we use a simple sampling strategy to obtain a sparsified matrix; that is, sample pair $(i,j)$’s of nodes with probability $p_{ij}$’s, then use the iteration method of Calvetti, Reichel, and Sorensen (1994) to find its leading vectors, and after that perform the $k$-means algorithm on these eigenvectors, which we refer to the random sampling-based spectral clustering.

We theoretically justify the randomized spectral clustering algorithms in terms of the approximation error that measures the deviation of the randomized matrix $\tilde{A}$ of the adjacency matrix $A$ from the population matrix $P$ and the misclassification error. In addition, although the spectral clustering is nonparametric in nature, we develop a simple method to estimate the link probability matrix $B$ based on the output clusters where $B_{kl}$ is the edge probability between any node pairs in communities $k$ and $l$ and provide its theoretical bound. It is worth noting that our analysis does not rely on the common assumption in most SBM literatures that $B$ is of full rank. In particular, we analyze the true eigen-structure of $P$ in the rank-deficient scheme and provide an explicit condition under which the nodes from different communities are separable. It turns out that the approximation error bound in terms of the spectral norm, namely, $\|\tilde{A} - P\|_2$, attains the minimax optimal rate in SBMs (Gao, Lu, and Zhou 2015; Gao and Ma 2020) under mild conditions, indicating that the optimization error from randomization, namely, $\|A - \tilde{A}\|_2$, is dominated by the statistical error from SBMs, namely, $\|A - P\|_2$. The misclassification error bounds are identical to the original spectral clustering (Lei and Rinaldo 2015) and are optimal provided that the community number $K$ is fixed (Ahn, Lee, and Suh 2018). We also generalize the results to degree-corrected block models—an extension of SBMs incorporating the degree heterogeneity (Karrer and Newman 2011).

The contributions of this article are as follows. First, we utilize randomization tools to obtain two kinds of randomized spectral clustering algorithms and theoretically study the resulting algorithm under the SBMs. The results provide statistical insights of randomization on spectral clustering. From the statistical perspective, the randomization does not deteriorate the error bound of $\|\tilde{A} - P\|_2$, because the latter already attains
the minimax optimal rate in SBMs. Second, extending the full-
rank assumption in most works on SBMs, we also study the rank-
deficient SBMs, analyze the true eigen-structure of these models, and provide sufficient conditions under which the spectral clustering may succeed, which is rarely mentioned in SBMs works and of independent interest. Third, we develop a new R models, and provides sufficient conditions under which the spec-
rank assumption in most works on SBMs, we also study the

The remainder of this article is organized as follows. Section 2 defines the notation, introduces and analyzes the SBM and spectral clustering in more detail. Section 3 includes the random projection-based and random sampling-based spectral clustering schemes that we consider. Section 4 presents the theoretical results. Section 5 contains extensions to degree-corrected block models. Section 6 reviews and discusses related works. Sections 7 and 8 display the simulation and real experiments that verify the theoretical results and show the effectiveness of the proposed methods. Section 9 concludes with discussion. Proofs are provided in the supplementary materials.

2. Preliminaries

In this section, we provide some notation and briefly introduce the SBMs and the spectral clustering algorithm. In particular, the rationality of spectral clustering under SBMs is analyzed.

2.1. Notation

Let \( M_{n,K} \) be the set of all \( n \times K \) matrices that have exactly one 1 and \( K - 1 \) 0’s in each row. Any \( \Theta \in M_{n,K} \) is called a membership matrix where each row represents the community membership of a node in a network with \( K \) communities; for example, node \( i \) belongs to community \( g_i \in \{1, \ldots, K\} \) if and only if \( \Theta_{ig_i} = 1 \). For \( 1 \leq k \leq K \), let \( G_k = G_k(\Theta) = \{i \in [n] : g_i = k\} \), where \([n] = \{1, 2, \ldots, n\}\). \( G_k \) consists of nodes with their community membership being \( k \), and denote \( n_k = |G_k| \). For any matrix \( A_{n \times n} \) and \( I, J \subseteq [n], A_{I \times J} \) and \( A_{J \times I} \) denote the submatrix of \( A \) consisting of the corresponding rows and columns, respectively. \( \|A\|_F \) and \( \|A\|_{\infty} \) denote the Frobenius norm and the element-wise maximum absolute value of \( A \), respectively. We use \( \|\cdot\|_2 \) to denote the Euclidean norm of a vector and the spectral norm of a matrix. In addition, \( \text{diag}(A) \) denotes the matrix with its diagonal elements being the same as those of \( A \) and non-diagonal elements being 0’s.

2.2. Stochastic Block Model

The SBM introduced by Holland, Laskey, and Leinhardt (1983) is a class of probabilistic model for networks with well-defined communities. For a potential network with \( n \) nodes and \( K \) communities, the model is parameterized by the membership matrix \( \Theta \in M_{n,K} \), and the link probability matrix \( B \in [0,1]^{K \times K} \) where \( B \) is symmetric, and the entry of \( B \); for example, \( B_{kl} \), represents the edge probability between the community \( l \) and \( k \). Here, \( B \) is not necessarily of full rank, and we assume rank(\( B \)) = \( K' \leq K \). Given \( \Theta \) and \( B \), the network adjacency matrix \( A = (a_{ij})_{1 \leq i, j \leq n} \in [0,1]^{n \times n} \) is generated as

\[
a_{ij} = \begin{cases} \text{Bernoulli}(B_{gg}) & \text{if } i < j, \\ 0, & \text{if } i = j, \\ a_{ji}, & \text{if } i > j. \end{cases}
\]  

(2.1)

Define \( P = \Theta B \Theta^T \), then it is easy to see that \( P \) is the population version of \( A \) in the sense that \( E(A) = P - \text{diag}(P) \). Under the SBMs, the goal of community detection is to use the adjacency matrix \( A \) to recover the membership matrix \( \Theta \) up to column permutations.

2.3. Spectral Clustering

Spectral clustering is a popular and simple algorithm for com-

probability model for networks with well-defined

random projection-based and random sampling-based spectral clustering schemes that we consider. Section 4 presents the theoretical results. Section 5 contains extensions to degree-corrected block models. Section 6 reviews and discusses related works. Sections 7 and 8 display the simulation and real experiments that verify the theoretical results and show the effectiveness of the proposed methods. Section 9 concludes with discussion. Proofs are provided in the supplementary materials.

2. Preliminaries

In this section, we provide some notation and briefly introduce the SBMs and the spectral clustering algorithm. In particular, the rationality of spectral clustering under SBMs is analyzed.

2.1. Notation

Let \( M_{n,K} \) be the set of all \( n \times K \) matrices that have exactly one 1 and \( K - 1 \) 0’s in each row. Any \( \Theta \in M_{n,K} \) is called a membership matrix where each row represents the community membership of a node in a network with \( K \) communities; for example, node \( i \) belongs to community \( g_i \in \{1, \ldots, K\} \) if and only if \( \Theta_{ig_i} = 1 \). For \( 1 \leq k \leq K \), let \( G_k = G_k(\Theta) = \{i \in [n] : g_i = k\} \), where \([n] = \{1, 2, \ldots, n\}\). \( G_k \) consists of nodes with their community membership being \( k \), and denote \( n_k = |G_k| \). For any matrix \( A_{n \times n} \) and \( I, J \subseteq [n], A_{I \times J} \) and \( A_{J \times I} \) denote the submatrix of \( A \) consisting of the corresponding rows and columns, respectively. \( \|A\|_F \) and \( \|A\|_{\infty} \) denote the Frobenius norm and the element-wise maximum absolute value of \( A \), respectively. We use \( \|\cdot\|_2 \) to denote the Euclidean norm of a vector and the spectral norm of a matrix. In addition, \( \text{diag}(A) \) denotes the matrix with its diagonal elements being the same as those of \( A \) and non-diagonal elements being 0’s.

2.2. Stochastic Block Model

The SBM introduced by Holland, Laskey, and Leinhardt (1983) is a class of probabilistic model for networks with well-defined communities. For a potential network with \( n \) nodes and \( K \) communities, the model is parameterized by the membership matrix \( \Theta \in M_{n,K} \), and the link probability matrix \( B \in [0,1]^{K \times K} \) with rank(\( B \)) = \( K' \leq K \), suppose the eigenvalue decomposition of \( P = \Theta B \Theta^T \) is \( U_{n \times K} \Sigma_{K' \times K'} U_{K' \times K}^T \). Define \( \Delta = \text{diag}(\sqrt{n_1}, \ldots, \sqrt{n_K}) \) and denote the eigenvalue decomposition of \( \Delta \Sigma \) by \( L_{K \times K'} \Sigma_{K' \times K'} L_{K' \times K}^T \). Then the following arguments hold.

(a) If \( B \) is of full rank, that is, \( K' = K \), then for \( \Theta_{ia} = \Theta_{ja} \), we have \( U_{ia} = U_{ja} \), while for \( \Theta_{ia} \neq \Theta_{ja} \), we have \( \|U_{ia} - U_{ja}\|_2 = \sqrt{(n_\Theta)^{-1} + (n_\Theta)^{-1}} \).

(b) If \( B \) is rank deficient, that is, \( K' < K \), then for \( \Theta_{ia} = \Theta_{ja} \), we have \( U_{ia} = U_{ja} \), while for \( \Theta_{ia} \neq \Theta_{ja} \), if \( \Delta^{-1} L \)’s rows are mutually distinct such that there exists a deterministic sequence \( \{\xi_n\}_{n \geq 1} \) satisfying

\[
\min_{k \neq l} \| L_{k\kappa} \|_{\sqrt{\xi_n}} - \| L_{l\kappa} \|_{\sqrt{\xi_n}} \|_2 \geq \xi_n > 0, \quad (A1)
\]

then \( \|U_{ia} - U_{ja}\|_2 = \| L_{ia} \|_{\sqrt{\xi_n}} - \| L_{ja} \|_{\sqrt{\xi_n}} \|_2 \geq \xi_n > 0 \).

Lemma 1 says that when \( B \) is of full rank, two rows of \( U \) are identical if and only if the corresponding nodes are in the same
community. These results have already been obtained in Lei and Rinaldo (2015) and Rohe, Chatterjee, and Yu (2011), among others. While when \( B \) is rank deficient, we additionally assume (A1) holds in order to make sure that two rows of \( U \) are separable when the corresponding nodes are in distinct communities. The next lemma provides an explicit condition on \( B \) that suffices for (A1). In particular, the within-community probabilities should interact with the assumptions of SBMs.

**Lemma 2.** For an SBM with \( K \) communities parameterized by \( \Theta \in \mathcal{M}_{n,K} \) and \( B \in [0,1]^{K \times K} \) with \( \text{rank}(B) = K' < K \), suppose the eigenvalue decomposition of \( P = \Theta B \Theta^\top \) is \( \Sigma U \Sigma^\top \). If there exists two deterministic sequences \( \{\eta_n\}_{n \geq 1} \) and \( \{\eta_n\}_{n \geq 1} \) such that

\[
\min_{1 \leq k \leq K} B_{kk} + B_{k\ell} - 2B_{k\ell} \geq \eta_n > 0, \tag{2.2}
\]

and for any \( 1 \leq i \leq K' \), \( 0 < \Sigma_i \leq \eta_n \), then (A1) holds with

\[
\xi_n = \sqrt{\eta_n / \eta_n}.
\]

Lemma 1 and 2 indicate that the spectral clustering could work well if the \( K' \) leading eigenvectors of \( A \) are close to those of the population \( P \). While when \( n \) is large, the full eigenvalue decomposition is time consuming. In the following sections, we will make use of the recently developed randomization techniques—namely, the random projection and the random sampling, to accelerate the spectral clustering. In the meanwhile, we will theoretically study how the randomized spectral clustering methods interact with the assumptions of SBMs.

**Algorithm 1** Spectral clustering for \( K \) clusters

**Input:** Cluster number \( K \), target rank \( K' \), adjacency matrix \( A \in \mathbb{R}^{n \times n} \).

**Output:** Estimated membership matrix \( \hat{\Theta} \in \mathcal{M}_{n,K} \) and centroids \( \hat{X} \in \mathbb{R}^{K' \times K'} \); Estimated eigenvectors \( \hat{U} = \hat{\Theta} \hat{X} \);

1. Find the \( K' \) leading eigenvectors \( \hat{U} \) of \( A \) corresponding to the \( K' \) largest eigenvalues of \( A 
2. Treat each row of \( \hat{U} \) as a point in \( \mathbb{R}^{K'} \) and run the Lloyd's algorithm on these points with \( K \) clusters. Let (\( \hat{\Theta}, \hat{X} \)) be the solution.

### 3. Randomized Spectral Clustering

In this section, we use the randomization techniques to derive two kinds of randomized spectral clustering—namely, random projection-based spectral clustering and random sampling-based spectral clustering.

#### 3.1. Randomized Spectral Clustering via Random Projection

Recall that \( A \) is generated from a low-rank matrix \( P = \Theta B \Theta^\top \), hence \( A \) inherits a low-rank structure naturally. Therefore, if one can make use of such low-rank structure to derive a smaller matrix that captures the essential information of \( A \), then the eigenvalue decomposition of the smaller matrix can help to derive that of \( A \), which in turn reduces the computational cost. Fortunately, randomization is a powerful tool for performing such low-rank matrix approximation (Halko, Martinsson, and Tropp 2011; Witten and Candès 2015; Martinsson 2016). These techniques utilize some amounts of randomness to compress the input matrix to obtain a low-rank factorization efficiently, which is called random projection. In this section, we introduce the random projection strategy in the context of eigenvalue decomposition.

Let us see how the random projection can help reduce the time for the eigenvalue decomposition of adjacency matrix \( A \). For a symmetric matrix \( A \in \mathbb{R}^{n \times n} \) with target rank \( K' \), we aim to find an orthonormal basis \( Q \in \mathbb{R}^{n \times K'}(K' \leq n) \) such that

\[
A \approx QQ^\top A QQ^\top := \tilde{A}^\top P,
\]

where \( \tilde{A}^\top P \) is essentially a low-rank approximation of \( A \). Before constructing \( Q \), we here provide some insights. \( Q \in \mathbb{R}^{n \times K'} \) can be thought as a low-rank approximation of the column (row) space of matrix \( A \). To see this, suppose the eigendecomposition of \( A \) is \( A = \tilde{U}_n \Sigma_n \tilde{U}_n^\top \), where \( m \) is the rank of \( A \) and \( \tilde{U} \) represents the column (row) space of \( A \). Then, when \( Q = \tilde{U} \) and \( m = K' \), it is straightforward to see \( A \approx QQ^\top A QQ^\top \).

In addition, \( QQ^\top \) is a projection operator which projects any vector \( x \in \mathbb{R}^n \) to the column space of \( Q \), that is,

\[
||x - QQ^\top x||_2^2 = \min_{y \in \mathbb{R}^{K'}} ||x - Qy||_2^2. \quad Q \text{ can be obtained using the following steps (Halko, Martinsson, and Tropp 2011):}
\]

**Step 1:** Form a random test matrix \( \Omega = (\omega_1, \ldots, \omega_K) \in \mathbb{R}^{n \times K} \), where \( \{\omega_i\}_{i=1}^K \) are \( n \)-dimensional random vectors independently drawn from a distribution.

**Step 2:** Form the sketch matrix \( Y = (y_1, \ldots, y_K) = A \Omega \in \mathbb{R}^{n \times K} \).

**Step 3:** Obtain \( Q \) via the QR decomposition \( Y = QR \).

Once \( Q \) is obtained, we can perform the eigenvalue decomposition on the smaller matrix \( C := QT^\top A Q \in \mathbb{R}^{K' \times K'} \), and then post process it to obtain the approximate eigenvectors of \( A \). In this way, the computational cost of the original spectral clustering could be largely reduced when we incorporate the aforementioned steps into **Algorithm 1** to provide the approximate eigenvectors of the adjacency matrix. We call this procedure random projection-based spectral clustering.

The random test matrix \( \Omega \) can be generated in various ways, specifically, the entries of \( \omega_i \) can be iid standard Gaussian, uniform, and Rademacher distributions, among many others. The oversampling strategy is often used to improve the empirical performance of the randomized low-rank approximation (Halko, Martinsson, and Tropp 2011; Witten and Candès 2015; Martinsson 2016). As most data matrices do not have exact rank \( K' \), it is desirable to use \( l := K' + r \) random projections instead of exact \( K' \) projections to form the random sketch of \( A \). In practice, \( r = \{5, 10\} \) often suffices to make sure that the obtained basis \( Q \) is close to the best possible basis, namely, the \( K' \) leading eigenvectors of \( A \), with high probability (Martinsson 2016). Besides the oversampling scheme, the power iteration is another way to improve the quality of low-rank approximation.
For some data matrices, the eigenvalues decay slowly that may lead to information loss. Thus, instead of forming the sketch $Y$ on the basis of $A$, several authors incorporate $q$ steps of a power iteration before constructing the sketch matrix $Y$. Formally, it is defined as

$$Y := (A A^T)^q A \Omega = A^{2q+1} \Omega.$$ 

In practice, $q = 1$ or $q = 2$ often suffices to make the spectrum decay fast (Halko, Martinsson, and Tropp 2011). We summarize the random projection-based spectral clustering procedure with such power iteration and the aforementioned sampling strategies in Algorithm 2.

**Remark 1.** The time complexity of Algorithm 2 is dominated by the matrix multiplications when forming $Y$ and $C$ in Step 2 and Step 4, which take $O((2q+1)n^2 (K' + r))$ and $O(n^2 (K' + r))$ time, respectively. In particular, the time complexity of Step 2 can be improved to $O((2q+1)n^2 \log(K' + r))$ by using structured random test matrices, for example, the subsampled random Fourier transform (Halko, Martinsson, and Tropp 2011; Ericson et al. 2019). Moreover, the matrix-vector multiplications in Step 2 can be parallelized to further reduce the computation cost.

**Algorithm 2** Randomized spectral clustering via random projection

**Input:**
Cluster number $K$, target rank $K'$, adjacency matrix $A \in \mathbb{R}^{n \times n}$, oversampling parameter $r$, and exponent $q$.

**Output:**
Membership matrix $\tilde{\Theta}^{\text{TP}} \in \mathbb{M}_{n,K}$ and centroids $\tilde{X}^{\text{TP}} \in \mathbb{R}^{K \times K'}$; $U^{\text{TP}} = \tilde{\Theta}^{\text{TP}} \tilde{X}^{\text{TP}}$.

1. Draw an $(K + r)$ random test matrix $\Omega$.
2. Form the matrix $Y = A^{2q+1} \Omega$.
3. Construct $Q$ via orthonormalizing the columns of $Y$, that is, $Y := QR$.
4. Form $C = Q^T A Q$ and denote $\tilde{A}^{\text{TP}} \equiv QCQ^T$.
5. Compute the eigenvalue decomposition of the small matrix: $C = U \Sigma U^T$.
6. Set $U^{\text{TP}}$ to be the column subset of $QU$ corresponding to the $K'$ largest values of $\Sigma$.
7. Treat each row of $U^{\text{TP}}$ as a point in $\mathbb{R}^{K'}$ and run the Lloyd’s algorithm on these points with $K$ clusters. Let $(\tilde{\Theta}^{\text{TP}}, \tilde{X}^{\text{TP}})$ be the solution.

### 3.2. Randomized Spectral Clustering via Random Sampling

The random sampling strategy is to first do element-wise sampling from the adjacency matrix $A$, and then use fast iterative methods, say orthogonal iteration or Lanczos iteration, to find a nearly-optimal best rank $K'$ approximation of $A$. The motivation is that in spectral clustering, we aim to find the first $K'$ eigenvectors of $A$, or the best rank $K'$ approximation of $A$. And there exist many fast iterative methods for computing such low-rank matrix approximation; see Calvetti, Reichel, and Sorensen (1994), Baglama and Reichel (2005), Allen-Zhu and Li (2016) and Lehoucq (1995), among many others. The time complexity of iterative methods is generally proportional to the number of nonzero elements of $A$ multiplied by the number of iterations. Hence, if we sample the elements of $A$ in some way to obtain a sparser matrix, then the time for computing its rank $K'$ approximation will be largely reduced. In the meantime, we hope that the sampling scheme does not deteriorate the accuracy too much. In the sequel, we introduce the random sampling procedure and the corresponding randomized spectral clustering.

We adopt a simple sampling strategy to obtain a sparsified version of $A$. That is, randomly select pairs $(i,j)$’s of the adjacency matrix $A$ independently with probability $p_{ij}$’s, and the randomized sparsified matrix $\tilde{A}^s$ is defined as

$$\tilde{A}^s_{ij} = \left\{ \begin{array}{ll} \frac{A_{ij}}{p_{ij}}, & \text{if } (i,j) \text{ is selected,} \\ 0, & \text{if } (i,j) \text{ is not selected,} \end{array} \right.$$ 

for each $i < j$, and $\tilde{A}^s_{ij} = \tilde{A}^s_{ji}$ for each $i > j$. Once $\tilde{A}^s$ is obtained, we can apply an iterative algorithm for the eigenvalue decomposition of $\tilde{A}^s$ to attain the nearly-optimal rank $K'$ approximation of $\tilde{A}^s$ such that

$$\tilde{A}^s \approx U^s_{K' \times K'} \Sigma^s_{K' \times K'} (U^s_{K' \times K'})^T_{K' \times n} = \tilde{A}^s.$$ 

Then the Lloyd’s algorithm can be applied on the rows of $U^s_{K' \times K'}$ to find the clusters. Let $(\tilde{\Theta}^{s\text{TP}}, \tilde{X}^{s\text{TP}})$ be the solution. For reference, we summarize these steps in Algorithm 3.

**Remark 2.** The sampling strategy is element-specific. The simplest choice is that $p_{ij} = p$ for all pairs of $(i,j)$. Note that it is equivalent to sampling 1’s with probability $p$ and sampling 0’s with probability $p' \quad (p' < p)$. Another choice is to set $p_{ij}$ proportional to $\|A_{ij}\|_2$ which enables that the edges from high-degree nodes would remain with higher probability, but computing $\|A_{ij}\|_2$ brings additional time cost. In addition, for real applications where certain edges or all edges of certain nodes are forced to remain in $A'$, one can use the element-dependent sampling strategy.

**Remark 3.** It should be noted that the iteration algorithms in Step 2 of Algorithm 3 yields the nearly-optimal solution instead of the exactly-optimal rank $K'$ approximation and it is acceptable to work with a nearly-optimal low-rank approximation. In the theoretical analysis, we treat Step 2 as a black box and suppose the best rank $K'$ approximation is obtained. We mainly deal with approximation error induced by Step 1.

### 4. Theoretical Analysis

In this section, we theoretically justify the performance of two randomization schemes on spectral clustering under the model set-up of SBMs. Specifically, for each method, we evaluate its performance from the following three aspects. First, we derive an upper bound on how the randomized matrix $\tilde{A}^{\text{TP}}$ (or $\tilde{A}^{s\text{TP}}$) deviates from the population adjacency matrix of SBMs. Then, we use these results to bound the misclassification error rate of the randomized spectral clustering algorithms. At last, we use the estimated clusters to obtain an estimate of $B$, and provide its theoretical bounds.
Algorithm 3 Randomized spectral clustering via random sampling

**Input:**
- Cluster number $K$, target rank $K'$, adjacency matrix $A \in \mathbb{R}^{n \times n}$, sampling probability matrix $\bar{P} = (\bar{p}_i)$;

**Output:**
- Membership matrix $\tilde{\Theta}^{rs} \in \mathbb{M}_{n,K}$ and centroids $\tilde{X}^{rs} \in \mathbb{R}^{K \times K'}$;
- $\tilde{U}^{rs} = \tilde{G}^{rs} \tilde{X}^{rs}$;

1. For each pair $(i,j)(i < j)$, randomly select pair $(i,j)$ of $A$ with probability $\bar{p}_{ij}$. Form the sparsified matrix $\tilde{A}$ according to (3.1).
2. Apply an iterative algorithm to obtain the nearly-optimal rank $K'$ approximation of $\tilde{A}$ such that $\tilde{A} \approx U_{n \times K'}^{rs} \Sigma_{K' \times K'}^{rs} \left(U_{n \times K'}^{rs}\right)^\top_{K' \times n} := \tilde{A}^{rs}$.
3. Treat each row of $U^{rs}$ as a point in $\mathbb{R}^{K'}$ and run the Lloyd’s algorithm on these points with $K$ clusters. Let $(\hat{\Theta}^{rs}, \hat{X}^{rs})$ be the solution.

### 4.1. Random Projection

The following notes and notation would be used throughout this section. Let $A$ be a $n \times n$ adjacency matrix generated from a SBM with $K$ communities parameterized by $\Theta \in \mathbb{M}_{n,K}$ and $B \in [0,1]^{K \times K}$ with rank $B = K' (K' \leq K)$. Denote the eigenvalue decomposition of $P = \Theta B \Theta^\top$ by $U_{n \times K'} \Sigma_{K' \times K'} \left(U_{n \times K'} \right)^\top_{K' \times n}$. Let $\sigma_n$ and $\gamma_n$ be the largest and smallest nonzero eigenvalue of $P$. Let $\tilde{\Theta}^{tp}$ be the output of Algorithm 2 with the target rank being $K'$, the oversampling and the power parameter being respectively $\tau$ and $q$, and the test matrix $\Omega$ generating iid standard Gaussian entries. The following theorem provides the deviation of $\tilde{A}^{tp}$ from $P$.

**Theorem 1.** If

\[
\max_{kl} B_{kl} \leq \alpha_n \text{ for some } \alpha_n \geq c_0 \log n/n, \tag{A2}
\]

and

\[
r \geq 4, r \log r \leq n, K' + r \leq n, q = c_1 \cdot n^{1/\tau}, \tag{A3}
\]

for some constant $c_0, c_1 > 0$ and any $\tau > 0$, then for any $s > 0$, there exists a constant $c_2 = c_2(s,c_0,c_1)$ such that

\[
\|\tilde{A}^{tp} - P\|_2 \leq c_2 \sqrt{n \alpha_n}, \tag{4.1}
\]

with probability at least $1 - 6r^{-r} - 2n^{-s}$.

The deviation of $\tilde{A}^{tp}$ from $P$ arises from two sources, one is the deviation of $\tilde{A}^{tp}$ from $A$ (optimization error), and the other is the deviation of $A$ from $P$ (statistical error). To bound the statistical error $\|A - P\|_2$, we pose condition (A2), a weak condition on the population network sparsity, which has been used to obtain a sharp bound of $\|A - P\|_2$ (Chin, Rao, and Vu 2015; Lei and Rinaldo 2015; Gao et al. 2017). To bound the optimization error $\|\tilde{A}^{tp} - A\|_2$, we use the result in Halko, Martinsson, and Tropp (2011) and pose condition (A3) on the order of the oversampling parameter $r$ and the power parameter $q$. It essentially indicates that the optimization error caused by random projection is dominated by the statistical error caused by sampling $A$ from $P$. Note that $q = c_1 \cdot n^{1/\tau}$ is mild because $\tau$ can be sufficiently large. Under (A2) and (A3), the bound in (4.1) attains the minimax optimal rate under the SBMs (Gao, Lu, and Zhou 2015; Gao and Ma 2020). Thus, in the sense of the spectral norm, the randomized matrix $\tilde{A}^{tp}$ and the nonrandomized matrix $A$ behave the same provided that $A$ is generated from an SBM, and thus, the randomization pays no price theoretically ignoring the conditions that we imposed. Moreover, (A2) could be removed if one consider regularized population adjacency matrix (Qin and Rohe 2013) or using other trimming steps (Le, Levina, and Vershynin 2015). (A3) could be relaxed if one use more advanced methods, say Clarkson and Woodruff (2017), Hu et al. (2021) and Martinsson and Tropp (2020).

With the derivation of $\tilde{A}^{tp}$ from $P$ at hand, we are ready to justify the clustering performance of Algorithm 2. We consider the following metric that measures the sum of the fractions of the misclustered nodes within each community,

\[
L_1(\tilde{\Theta}, \Theta) = \min_{j \in E_k} \sum_{1 \leq k \leq K} (2n_k)^{-1} \| (\tilde{\Theta} f) G_{k*} - \Theta G_{k*} \|_0. \tag{4.2}
\]

where $\tilde{\Theta}$ is an estimate of $\Theta$, and $E_k$ is the set of all $K \times K$ permutation matrices. The following theorem provides an upper bound on $L_1$.

**Theorem 2.** Suppose that (A1), (A2), and (A3) hold, and there exists an absolute constant $c_3 > 0$ such that

\[
\frac{K' n \alpha_n}{\gamma_n^2 \delta_n^2 \min n_k} \leq c_3, \tag{A4}
\]

then with probability larger than $1 - 6r^{-r} - 2n^{-s}$ for any $s > 0$, there exist subsets $S_k \in G_k$ for $k = 1, \ldots, K$ such that

\[
L_1(\tilde{\Theta}^{tp}, \Theta) \leq \sum_{k=1}^K \frac{|S_k|}{n_k} \leq c_3 \frac{K' n \alpha_n}{\gamma_n^2 \delta_n^2 \min n_k}. \tag{4.5}
\]

Moreover, for $G = \bigcup_{k=1}^K (G_k \setminus S_k)$, there exists a $K \times K$ permutation matrix $J$ such that

\[
\tilde{\Theta}^{tp}_{G*} J = \Theta_{G*}. \tag{4.6}
\]

The proof of Theorem 2 follows that in Lei and Rinaldo (2015). (A1) is required only when $K' < K$. (A2) and (A3) ensure the results of Theorem 1 hold. (A4) is a technical condition which ensures the bound in (4.5) vanishes and provides the range of parameters $(K, n, \alpha_n, \gamma_n, \delta_n)$ in which the result is appropriate. (A4) is satisfied automatically if the bound in (4.5) is $o(1)$. $S_k$ is actually the set of nodes in $G_k$ that are misclustered. $\delta_n$ measures the minimum distance of every pair of rows of true eigenvectors for nodes from different communities (recall Lemma 1). As expected, larger $\delta_n$ and $\min n_k$ indicate better misclassification error rate. In particular, following Theorem 1,
the bound in (4.5) is identical to that of the nonrandomized spectral clustering when $K^* = K$ (Lei and Rinaldo 2015).

The bound in (4.5) is not explicit as $\gamma_n$ is related to $n$. To illustrate, we now consider a simple case. Suppose a SBM parameterized by $(\Theta, B)$ is generated with balanced communities size $n/K$, and

\[ P = \Theta B \Theta^T = \Theta(\alpha_n \lambda I_K + \alpha_n (1 - \lambda) 1_K 1_K^T) \Theta^T, \quad (4.7) \]

where $1_K$ represents a $K$ dimensional vector of 1's and $\lambda$ is a constant. In the case, $\gamma_n = n\alpha_n/K$ (Rohe, Chatterhee, and Yu 2011), and then the bound in (4.5) reduces to

\[ \sum_{k=1}^K |S_k| n_k = O(K^3/n\alpha_n). \]

Let us discuss some specific parameter settings now. For fixed $K$, $n\alpha_n$ needs to be of order $\omega(1)$, namely, $n\alpha_n \geq c$ for some constant $c$, to ensure a vanishing error bound. In such case, the bound $O(1/n\alpha_n)$ is optimal in the sense that there is no estimator which is weakly consistent when $n\alpha_n = O(1)$ (see, e.g., Ahn, Lee, and Suh 2018). On the other hand, when $\alpha_n = c_0 \log n / n$, $K = O((\log n)^{1/3})$ is required to ensure a vanishing misclassification error rate. It should be noted that since the pure spectral clustering generally could not attain the optimal misclassification error rate under SBMs (Gao et al. 2017) except some simple case ($K = 2$, within-community and between-community probability being $\log n / n$ and $\log n / n$) considered in Abbe et al. (2020), our randomized version also has limitations in terms of misclassification rate. While the algorithms in Gao et al. (2017) that attain the statistical optimal error rate has higher computational complexity than the randomized spectral clustering we considered here do. The current error rate would be improved if one study more refined proof techniques of pure spectral clustering or develop variants of spectral clustering that has better error rates but without increasing the time complexity.

Remark 4. In the proof of Theorem 2, we made an assumption that the $k$-means algorithm finds the optimal solution as in Rohe, Chatterhee, and Yu (2011). Alternatively, one can use more delicate $(1 + \varepsilon)$-approximate $k$-means (Matoušek 2000; Kumar, Sabharwal, and Sen 2004) to bridge the gap, where one can find a good approximate solution within a constant fraction of the optimal value.

In the sequel, we discuss how we can utilize the estimated link probability matrix $\hat{\Theta}^P$ and $\hat{A}^P$ to estimate the link probability matrix $B$. Without loss of generality, we assume that the permutation matrix $j$ in (4.6) is $I_{K \times K}$. Noting that

\[ B_{ij} = \sum_{1 \leq i \leq K} \frac{p_{ij} \Theta_{ij} \Theta_{ji}}{\sum_{1 \leq j \leq K} \Theta_{ij} \Theta_{ji}}, \quad 1 \leq q, l \leq K, \]

thus, it is reasonable to estimate $B$ by the following

\[ \hat{B}^P = (\hat{B}^P)_{1 \leq q, l \leq K}, \]

\[ \hat{B}^P_{ij} = \frac{1 \sum_{1 \leq i \leq K} \hat{A}^P_{ij} \Theta_{ij} \Theta_{ji}}{\sum_{1 \leq i \leq K} \hat{A}^P_{ij} \Theta_{ij} \Theta_{ji}}, \quad 1 \leq q, l \leq K. \]

The following theorem provides a theoretical bound for the estimator $\hat{B}^P$.

**Theorem 3.** Suppose that (A2), (A3), and (A4) hold, then with probability larger than $1 - 6e^{-t} - 2K^2 n^{-s}$ for any $s > 0$, there exists constant $c_4 > 0$ so that,

\[ \| \hat{B}^P - B \|_\infty \leq c_4 \frac{\sqrt{K + t \sqrt{n\alpha_n}} + \sqrt{K \alpha_n}}{\min n_k} \left( 1 + (1 - \Phi_n)^{-1} + \frac{2\max n_k}{\min n_k} \left( 1 - \Phi_n \right)^{-2} \right), \]

(4.8)

with $\Phi_n := c_3^{-1} \frac{1}{\sqrt{q} \log n}$, where $\delta_n = \delta_{1n}$ (see (4.3)) when $K' = K$ and $\delta_n = \delta_{2n}$ (see (4.4)) when $K' < K$.

Let us illustrate the bound in (4.8) more explicitly. As a simple example, we consider the specific case in (4.7). Suppose further that $\alpha_n = c_0 \log n/n$, and then the bound in (4.8) reduces to

\[ O \left( \frac{K^3/2 \log n}{n} \left( 1 + (1 - K^3/\log n)^{-1} \right) \right). \]

It turns out that $K = O((\log n)^{1/3})$ would lead to a vanishing bound. We note that Tang, Cape, and Priebe (2021) established the asymptotic normality results for the estimation of $B$ in SBMs, where they assume $n\alpha_n = \omega(\sqrt{n})$. In particular, when $B$ is of full-rank, $n\alpha_n = O(\sqrt{n})$ and the community size is balanced, the error rate for $B$ is $O(K^3/\sqrt{n})$ in Tang, Cape, and Priebe (2021), tighter than $O(\sqrt{n})$ in our work, which is partially because that we study the randomized spectral clustering while they considered the original spectral clustering. Note that the parameter range of $\alpha_n$ in this work is more friendly than theirs. In addition, it would be interesting to study the asymptotic properties of $\hat{B}$ under the setting of randomized spectral clustering.

### 4.2. Random Sampling

Similar to the random projection-based spectral clustering, we will derive theoretical results on the random sampling method from three aspects—namely, the deviation of $\hat{A}^s$ from $P$, the misclassification error rate, and the deviation of $\hat{B}^s$ from $B$, where $\hat{A}^s$ is an analog of $\hat{B}^P$ with the estimators therein replaced by the counterparts under the random sampling scheme.

The SBM set-up is the same with that in Section 4.1. We here recall some notation specific to the random sampling scheme. Let $\hat{A}^s$ be the intermediate output in Algorithm 3 with the target rank being $K'$, that is, the best rank-$K'$ approximation of the sparsified matrix $A^s$ whose elements $(i, j)$'s are sampled from $A$ with probability $p_j$'s. The next theorem provides an upper bound for the deviation of $\hat{A}^s$ from $P$.

**Theorem 4.** Suppose that (A2) holds and assume

\[ p_j \geq p_{\min}, \quad \text{for all } 1 \leq i < j \leq n. \]

Define

\[ I_1 = \min \left\{ n\alpha_n \sqrt{\frac{\max |\sum_j \frac{1}{p_j}|}{p_{\min}}}, \right\} \]

I_1 is a constant.
then there exist constants $c_5 > 0$ and $c_6 > 0$ such that
\[
\|\tilde{A}^\tau - P\|_2 \leq c_5
\]
\[
\max\left\{ l_1, \sqrt{\frac{\log n}{p_{\min}}}, \sqrt{n \alpha n^{-1}}, \sqrt{2 \log \max(1, \frac{1}{p_{\min}})^2} \right\} := \Psi_n,
\]
with probability larger than $1 - c_6 n^{-v}$, where constant $v > 0$ depends on $c_5$.

It should be noted that the bound in (4.10) is not obtained by simple combination of the $\|A - P\|_2$ and $\|\tilde{A}^\tau - A\|_2$. Instead, we make use of the low-rank nature of $P$, which measures the worst overall sampling probability of all edges of certain node may affect the bound. In particular, when $\alpha_n$ is fixed, and $p_{11}, p_{22}, \ldots, p_{nn}$'s are highly heterogeneous for each fixed $i$, $l_1$ reduces to $\sqrt{\max_i \sum_j p_{ij}}$. It should be noted that when $p_{ij}$'s are uniform and fixed, the RHS of (4.10) reduces to $n \alpha$, being the same with the best concentration bound of the full adjacency matrix $A$ around its population $P$ (Lei and Rinaldo 2015; Gao, Lu, and Zhou 2015). In this sense, the sampled matrix $\tilde{A}^\tau$ can be regarded as a network sampled from the same SBM generating $A$, although the elements of $\tilde{A}^\tau$ are not binary.

The following theorem justifies the clustering performance of the randomized spectral clustering via the random sampling (Algorithm 3).

**Theorem 5.** Suppose that (A1), (A2), and (A5) hold, and assume there exists an absolute constant $c_7 > 0$ such that,
\[
\frac{K' \Psi_n^2}{\nu^{2(n-1)} \min n_k} \leq c_7
\]
where $\delta_1 := \delta_{1n}$ (see (4.3)) when $K' = K$ and $\delta_2 := \delta_{2n}$ (see (4.4)) when $K' < K$. Then with probability larger than $1 - c_6 n^{-v}$ for some $v > 0$, there exist subsets $S_k \in G_k$ for $k = 1, \ldots, K$ such that
\[
L_1(\tilde{\Theta}^\tau, \theta) \leq \sum_{k=1}^{K} \frac{|S_k|}{n_k} \leq c_1 \frac{K' \Psi_n^2}{\nu^{2(n-1)} \min n_k}
\]
where recall that $\Psi_n$ is defined in (4.10). Moreover, for $G := \cup_{k=1}^{K} (G_k \setminus S_k)$, there exists a $K \times K$ permutation matrix $J$ such that
\[
\tilde{\Theta}^\tau_{G_k} = \Theta_{G_k}.
\]

The proof is similar to that of Theorem 2, hence we omit it. Under the assumption of SBM in (4.7) and let $b$ be fixed; then similar to the random projection scheme, the bound in (4.11) reduces to $O(\nu^{3(n-1)} / np_{\min})$, which is $o(1)$ under the parameter setup that $\alpha_n = \log n / n$ and $K = o((\log n)^{1/3})$. Also, the current bound could be improved potentially; see our discussion after Theorem 2.

Next, we turn to the estimation of the link probability matrix $B$. Similar to the random projection setting, we define the following plug-in estimator $\tilde{B}^\tau = (\tilde{B}^\tau_{ij})_{1 \leq i, j \leq K}$ for $B$,
\[
\tilde{B}^\tau_{ij} := \frac{\sum_{1 \leq i, j \leq n} \tilde{A}^\tau_{ij} \tilde{\Theta}^\tau_{ij}}{\sum_{1 \leq i, j \leq n} \tilde{\Theta}^\tau_{ij}} = 1, 1 \leq q, l \leq K.
\]
The following theorem provides an upper bound for the deviation of $\tilde{B}^\tau = (\tilde{B}^\tau_{ij})_{1 \leq q, l \leq K}$ from $B$.

**Theorem 6.** Suppose that (A2), (A5), and (A6) hold, then with probability larger than $1 - c_6 K n^{-v}$ for some $v > 0$, there exists $c_8 > 0$ that,
\[
\|\tilde{B}^\tau - B\|_\infty \leq c_8 \left( \sqrt{K' + r \sqrt{\nu n}} + \sqrt{K' \alpha n^2} \right)
\]
\[
\left( 1 + (1 - \Psi_n)^{-1} + \frac{2 \max n_k}{n_k} \right)
\]
where recall that $\Psi_n$ is defined in (4.10).

We omit the proof since it is similar to that of Theorem 3. We can discuss the bound (4.13) in a similar way to those in the random projection scheme. For example, under the special case of SBM in (4.7), let $\alpha_n = \log n / n$ and $p$ be fixed, then the bound (4.13) reduces to the one in (4.9). Thus, $K = o((\log n)^{1/3})$ suffices to make sure that the RHS of (4.13) vanishes when $n$ goes to infinity.

5. Extensions

Standard SBMs often fail to capture the property of networks with strong degree heterogeneity. As a remedy, in this section we extend our results to degree-corrected stochastic block models (DC-SBMs) coupled with the randomized spherical spectral clustering.

5.1. Degree-Corrected Stochastic Block Models

Similar to the SBMs, the DC-SBMs (Karrer and Newman 2011) are parameterized by the membership matrix $\Theta \in M_{n,K}$, and the link probability matrix $B \in [0, 1]^{K \times K}$ where $B$ is not necessary of full rank and denote $\text{rank}(B) = K'(K' \leq K)$. To account for the degree heterogeneity, the DC-SBMs additionally introduce the node propensity parameter $\vartheta \in \mathbb{R}_+^n$. With these set-up, the population adjacency matrix is defined as $P := \text{diag}(\vartheta) \Theta \Theta^\top \text{diag}(\vartheta)$. To make the parameters identifiable, we follow Lei and Rinaldo (2015) to assume that $\max_{i \in G_k} \vartheta_i = 1$. To facilitate further analysis, let $\phi_k$ be an $n \times 1$ vector that is consistent with $\vartheta$ on $G_k$ and zero otherwise. Let $\Omega = \text{diag}(\vartheta_{f1}, \ldots, \vartheta_{fK})$, and let $\hat{B} = \Omega B \Omega^\top$. The following lemma reveals the eigen-structure of the population matrix $P$.

**Lemma 3.** For a DC-SBM with $K$ communities parameterized by $\Theta \in M_{n,K}$, $B \in [0, 1]^{K \times K}$ and $\vartheta \in \mathbb{R}_+^n$, we suppose that $\text{rank}(B) = K'(K' \leq K)$ and the eigenvalue decomposition of $P = \text{diag}(\vartheta) \Theta \Theta^\top \text{diag}(\vartheta)$ is $U_{n \times K'} \Sigma_{K' \times K'} U_{n \times K'}^\top$. Denote the eigenvalue decomposition of $\hat{B}$ by $H_{n \times K'} \Sigma_{K' \times K'} H_{n \times K'}^\top$. For any two vectors $a$ and $b$, $\text{cos}(a, b)$ is defined to be $a^\top b / \|a\|_2 \|b\|_2$. Then the following arguments hold.

(a) If $B$ is of full rank, that is, $K' = K$, then for any $\Theta_{f_k}$, $\text{cos}(U_{f_k}, U_{f_j}) = 1$, and for $\Theta_{f_k} \neq \Theta_{f_j}$, $\text{cos}(U_{f_k}, U_{f_j}) = 0$.
(b) If $B$ is rank deficient, that is, $K' < K$, then for any $\Theta_{f_k} = \Theta_{f_j}$, $\text{cos}(U_{f_k}, U_{f_j}) = 1$, and for any $\Theta_{f_k} \neq \Theta_{f_j}$, if $H'$s
The randomized spherical spectral clustering is readily available when we replace the input adjacency matrix $A$ in Algorithm 4 by the randomized counterpart $\tilde{A}$ ($\tilde{A}^{TP}$ or $\tilde{A}^T$). With slight abuse of notation, the output is denoted by $\tilde{\Theta}$ ($\tilde{\Theta}^{TP}$ or $\tilde{\Theta}^T$).

**Remark 5.** The spherical spectral clustering algorithms have been studied by several authors; see Lei and Rinaldo (2015) and Qin and Rohe (2013), among others. In particular, Lei and Rinaldo (2015) remove the zero rows of $\bar{U}$ and use $k$-median instead of $k$-means for technical reasons. Differently, we let the zero rows of $\bar{U}$ be untreated and still use the $k$-means on the normalized vectors. Note that besides $k$-means based algorithms, one could use other clustering algorithms, say subspace clustering (Vidal, Ma, and Sastry 2005; Liu et al. 2012; Terada 2014), directly on the un-normalized eigenvectors.

### 5.3. Misclassification Analysis

Note that the approximation error bounds $||\tilde{A} - P||_2$ ($\tilde{A}$ can be $\tilde{A}^{TP}$ or $\tilde{A}^T$; see (4.1) and (4.10)) only make use of the low-rank nature of $P$, hence they remain the same under the DC-SBMs. The following theorem provides the misclassification error rate of randomized spherical spectral clustering on DC-SBMs, where output $\tilde{\Theta}$ represents $\tilde{\Theta}^{TP}$ and $\tilde{\Theta}^{TS}$ respectively when $\tilde{A} = \tilde{A}^{TP}$ and $\tilde{A} = \tilde{A}^T$.

**Theorem 7.** For a DC-SBM with $K$ communities parameterized by $\Theta \in \mathbb{M}_{n,K}$, $B \in [0,1]^{K \times K}$ and $\vartheta \in \mathbb{R}^n_+$, such that rank($B$) = $K'$ ($K' < K$) and the eigenvalue decomposition of $P = \text{diag}(\vartheta)\Theta B \Theta^T \text{diag}(\vartheta)$ is $U_{n \times K'} \Sigma_{K' \times K'} U_{K' \times K'}^T$. Recall that $\Omega = \text{diag}(|\varphi_1|, \ldots, |\varphi_K|)$, and $\tilde{B} = \Omega \varphi \Omega^2$. If there exists deterministic positive sequences $(\eta_i')_{n \geq 1}, (\xi_n')_{n \geq 1}, (\tau_n)_{n \geq 1}$, and $(\beta_n)_{n \geq 1}$ such that

$$\min_{1 \leq k \leq K} \tilde{B}_{kk} \xi_n' \leq \xi_n' \beta_n'$$

and for any $1 \leq i \leq K'$, $0 < \xi_n' \leq \Sigma_{ii} < \tau_n$, and $0 < \min_{1 \leq k \leq K} \tilde{B}_{kk} \leq \max_{1 \leq k \leq K} \tilde{B}_{kk} < \beta_n'$, then (A7) holds with

$$\varepsilon_n' = \sqrt{1 - \frac{\eta_i'}{\tau_n \beta_n'}},$$

Compared with Lemma 1, we see that for the DC-SBMs, not the distances but the angles between the rows of true eigenvector $U$ reveal whether the corresponding nodes are in the same community.

#### 5.2. Randomized Spherical Spectral Clustering

In light of Lemma 3, to make the spectral clustering valid on DC-SBMs, we need to normalize the rows of eigenvectors before performing the $k$-means. In this way, the angle-based results in Lemma 3 can be transformed to the distance-based counterpart, and thus, making the distance-based $k$-means valid. The resulting algorithm is called spherical spectral clustering; see Algorithm 4.

**Algorithm 4** Spherical spectral clustering for $K$ clusters

**Input:**
- Cluster number $K$, target rank $K'$, adjacency matrix $A \in \mathbb{R}^{n \times n}$;

**Output:**
- Estimated membership matrix $\tilde{\Theta} \in \mathbb{M}_{n,K}$ and centroids $X \in \mathbb{R}^{K \times K'}$;
- Estimated eigenvectors $\tilde{U} = \tilde{\Theta} \tilde{X}$;

1. Find the $K'$ leading eigenvectors $\tilde{U}$ of $A$ corresponding to the $K'$ largest eigenvalues of $A$.
2. Normalize each row of $\tilde{U}$ and denote the resulting matrix by $\tilde{U}'$, where the rows with Euclidean norm 0’s are removed the same.
3. Treat each row of $\tilde{U}'$ as a point in $\mathbb{R}^{K'}$ and run the Lloyd’s algorithm on these points with $K$ clusters. Let $(\tilde{\Theta}, \tilde{X})$ be the solution.

The randomized spherical spectral clustering is readily available when we replace the input adjacency matrix $A$ in Algorithm 4 by the randomized counterpart $\tilde{A}$ ($\tilde{A}^{TP}$ or $\tilde{A}^T$). With slight abuse of notation, the output is denoted by $\tilde{\Theta}$ ($\tilde{\Theta}^{TP}$ or $\tilde{\Theta}^T$).

**Remark 5.** The spherical spectral clustering algorithms have been studied by several authors; see Lei and Rinaldo (2015) and Qin and Rohe (2013), among others. In particular, Lei and Rinaldo (2015) remove the zero rows of $\bar{U}$ and use $k$-median instead of $k$-means for technical reasons. Differently, we let the zero rows of $\bar{U}$ be untreated and still use the $k$-means on the normalized vectors. Note that besides $k$-means based algorithms, one could use other clustering algorithms, say subspace clustering (Vidal, Ma, and Sastry 2005; Liu et al. 2012; Terada 2014), directly on the un-normalized eigenvectors.

### 5.3. Misclassification Analysis

Note that the approximation error bounds $||\tilde{A} - P||_2$ ($\tilde{A}$ can be $\tilde{A}^{TP}$ or $\tilde{A}^T$; see (4.1) and (4.10)) only make use of the low-rank nature of $P$, hence they remain the same under the DC-SBMs. The following theorem provides the misclassification error rate of randomized spherical spectral clustering on DC-SBMs, where output $\tilde{\Theta}$ represents $\tilde{\Theta}^{TP}$ and $\tilde{\Theta}^{TS}$ respectively when $\tilde{A} = \tilde{A}^{TP}$ and $\tilde{A} = \tilde{A}^T$.

**Theorem 7.** For a DC-SBM with $K$ communities parameterized by $\Theta \in \mathbb{M}_{n,K}, B \in [0,1]^{K \times K}$ with rank($B$) = $K'$ ($K' < K$) and $\vartheta \in \mathbb{R}^n_+$, such that rank($B$) = $K'$ ($K' < K$) and the eigenvalue decomposition of $P = \text{diag}(\vartheta)\Theta B \Theta^T \text{diag}(\vartheta)$ is $U_{n \times K'} \Sigma_{K' \times K'} U_{K' \times K'}^T$. Recall that $\Omega = \text{diag}(|\varphi_1|, \ldots, |\varphi_K|)$, and $\tilde{B} = \Omega \varphi \Omega^2$. If there exists deterministic positive sequences $(\eta_i')_{n \geq 1}, (\xi_n')_{n \geq 1}, (\tau_n)_{n \geq 1}$, and $(\beta_n)_{n \geq 1}$ such that

$$\min_{1 \leq k \leq K} \tilde{B}_{kk} \xi_n' \leq \xi_n' \beta_n'$$

and for any $1 \leq i \leq K'$, $0 < \xi_n' \leq \Sigma_{ii} < \tau_n$, and $0 < \min_{1 \leq k \leq K} \tilde{B}_{kk} \leq \max_{1 \leq k \leq K} \tilde{B}_{kk} < \beta_n'$, then (A7) holds with

$$\varepsilon_n' = \sqrt{1 - \frac{\eta_i'}{\tau_n \beta_n'}},$$

Compared with Lemma 1, we see that for the DC-SBMs, not the distances but the angles between the rows of true eigenvector $U$ reveal whether the corresponding nodes are in the same community.

#### 5.2. Randomized Spherical Spectral Clustering

In light of Lemma 3, to make the spectral clustering valid on DC-SBMs, we need to normalize the rows of eigenvectors before performing the $k$-means. In this way, the angle-based results in Lemma 3 can be transformed to the distance-based counterpart, and thus, making the distance-based $k$-means valid. The resulting algorithm is called spherical spectral clustering; see Algorithm 4.

**Algorithm 4** Spherical spectral clustering for $K$ clusters

**Input:**
- Cluster number $K$, target rank $K'$, adjacency matrix $A \in \mathbb{R}^{n \times n}$;

**Output:**
- Estimated membership matrix $\tilde{\Theta} \in \mathbb{M}_{n,K}$ and centroids $X \in \mathbb{R}^{K \times K'}$;
- Estimated eigenvectors $\tilde{U} = \tilde{\Theta} \tilde{X}$;

1. Find the $K'$ leading eigenvectors $\tilde{U}$ of $A$ corresponding to the $K'$ largest eigenvalues of $A$.
2. Normalize each row of $\tilde{U}$ and denote the resulting matrix by $\tilde{U}'$, where the rows with Euclidean norm 0’s are removed the same.
3. Treat each row of $\tilde{U}'$ as a point in $\mathbb{R}^{K'}$ and run the Lloyd’s algorithm on these points with $K$ clusters. Let $(\tilde{\Theta}, \tilde{X})$ be the solution.
incorporating the bound of sense. Larger min ativemethods for fast eigen-decomposition.

In this section, we review and discuss the literature that is closely related to the current work. We classify them into three groups: spectral clustering, randomization techniques, and iterative methods for fast eigen-decomposition.

The community detection is one of the fundamental problems in network analysis. The SBMs and their variants have been useful tools for modeling networks with communities and thus, being widely studied (Abbe 2018). In particular, a multitude of researches focus on spectral clustering and its variants, see Arroyo and Levine (2021), Chin, Rao, and Vu (2015), Fishkind et al. (2013), Joseph and Yu (2016), Lei and Rinaldo (2015), Li, Levine, and Zhu (2020a), Lyzinski et al. (2014), Paul and Chen (2020), Qin and Rohe (2013), Rohe, Chatterjee, and Yu (2011), Tang, Cape, and Priebe (2021), Su, Wang, and Zhang (2019), Yang et al. (2020), Yun and Proutiere (2014) and Yun and Proutiere (2016), and references therein, among which weak (strong) consistency, namely the fraction (number) of misclustered nodes decreases to zero as \( n \) grows, are well-established. Compared with most of these works, the current work has novelty in terms of both algorithms and also theoretics. In respect of algorithms, the randomized spectral clustering algorithms can deal with networks with up to millions number of nodes, showing the advantage over original spectral clustering with full eigenvalue decomposition. In respect of theoretics, the approximation error bound \( \| \tilde{A} - P \|_2 \) is optimal under mild conditions though we use randomized adjacency matrix \( \tilde{A} \). As a by-product, we generalize the common assumption \( \text{rank}(B) = K \) in SBMs and DC-SBMs to \( \text{rank}(B) \leq K \), which is of independent interest and rarely mentioned in the works of literature except Tang, Cape, and Priebe (2021); Fishkind et al. (2013), and a few others.

There are also various prior works on spectral clustering using randomized methods, see Liao, Couillet, and Mahoney (2020), Sakai and Imiya (2009), Sinha (2018), Tremblay et al. (2016), Tremblay and Loukas (2020), Wang, Gittens, and Mahoney (2019b), and Yan, Huang, and Jordan (2009), among others. For example, Sakai and Imiya (2009) developed fast spectral clustering algorithms by using random projection and random sampling techniques in order to reduce the data dimensionality and cardinality. Yan, Huang, and Jordan (2009) provided a general framework for fast spectral clustering where a distortion-minimizing local transformation is first applied to the data to reduce the dimensionality. Tremblay et al. (2016) proposed the compressive spectral clustering using the randomized techniques in graph signal processing. Compared with this line of works, the merits of this work lie in that we study the effect of randomization from the statistical point of view—under the framework of SBMs and DC-SBMs. The current methods can obtain optimal error for \( \| \tilde{A} - P \|_2 \) under mild conditions, indicating that the optimization error induced by random projection or random sampling are dominated by the statistical error induced by the randomness of networks from SBMs and DC-SBMs. It should be noted that the structure of SBMs and DC-SBMs facilitates bounding the approximation error in the random sampling regime. The resulting bound is tighter than those obtained by simply combining the optimization error bound \( \| \tilde{A}^{rs} - A \|_2 \) in Achlioptas and McSherry (2007) and the statistical error bound \( \| \tilde{A} - P \|_2 \) in Lei and Rinaldo (2015). Note that Li, Levine, and Zhu (2020b) also studied the deviation of \( \tilde{A}^{rs} \) from \( P \) but in the context of network cross-validation. It turns out that \( K \leq n/\log n \) is additionally required therein to ensure that the concentration bound of \( \tilde{A}^{rs} \) meets that of the full adjacency matrix \( A \), provided that \( p \) is fixed.

Iterative methods are widely used for partial eigen-decomposition and there are fruitful works in this line; see Allen-Zhu and Li (2016), Baglama and Reichel (2005), Calvetti, Reichel, and Sorensen (1994) and Lehoczky (1995), among others. We illustrate the merits of this work as follows. Actually, in the random sampling scheme, we use the iterative methods of (Calvetti, Reichel, and Sorensen 1994; Qiu and Mei 2019) as our baseline method and study how sampling could be used to further accelerate the partial eigen-decomposition. While for the random projection scheme, it has the following advantages (Halko, Martinsson, and Tropp 2011). First, the random projection-based methods are scalable because the matrix-vector operations can be done via multi-threading and distributed computing, which has been exploited in the R package (RClust) of this work. Second, the random projection-based methods have low communication costs as it only requires few passes over the input matrix. Further, the communication costs could be reduced by considering single-pass version (Tropp et al. 2016). At last, our experiments show that the randomized methods are faster than iterative methods while achieving satisfactory performance provided that the network’s scale is super-large, say millions of nodes.

7. Numerical Results

In this section, we empirically compare the finite sample performance of the randomized spectral clustering, namely, the random projection and the random sampling, with the original spectral clustering, where we use uniform sampling in the random sampling scheme for computational convenience. We will start with a simple SBM model to test the effect of \( n, K, \alpha \) on the approximation error, misclassification error, estimation error for \( B \), respectively. Then we extend our model setting to more complex models. At last, we test the effect of hyper parameters, including the power parameter \( q \) and the oversampling parameter \( r \) in the random projection scheme, and the sampling parameter \( p \) in the random sampling scheme.

7.1. Theoretical Bounds Evaluation

To be consistent with Section 4, we use the following three metrics to evaluate the theoretical performance of each method. The first one is the spectral derivation of the “approximated”
Figure 1. The average effect of \( n \) on the three metrics over 20 replications. (a)–(c) correspond to the approximation error for \( P \), the misclassification error for \( \Theta \), and the estimation error for \( B \), respectively. The other parameters \( K = 3, \alpha_n = 0.2, \alpha_n(1 - \lambda) = 0.1, r = 10, q = 2, p = 0.7 \), and \( \Omega \) had iid standard Gaussian entries, respectively.

Figure 2. The average effect of \( \alpha \) on the three metrics over 20 replications. (a)–(c) correspond to the approximation error for \( P \), the misclassification error for \( \Theta \), and the estimation error for \( B \), respectively. The other parameters \( n = 1152, K = 3, \lambda = 0.5, r = 10, q = 2, p = 0.7 \), and \( \Omega \) had iid standard Gaussian entries, respectively.

Figure 3. The average effect of \( K \) on the three metrics over 20 replications. (a)–(c) correspond to the approximation error for \( P \), the misclassification error for \( \Theta \), and the estimation error for \( B \), respectively. The other parameters \( n = 1152, \alpha_n = 0.2, \alpha_n(1 - \lambda) = 0.1, r = 10, q = 2, p = 0.7 \), and \( \Omega \) had iid standard Gaussian entries, respectively.

adjacency matrix \( \hat{A} \) from the population adjacency matrix \( P \), namely, \( \| \hat{A} - P \|_2 \), where \( \hat{A} \) can be \( \hat{A}^r, \hat{A}^p \) or \( A \). The second metric is the sum of the fractions of misclustered nodes within each true cluster, namely,

\[
\min_{J \in E_k} \sum_{1 \leq k \leq K} (2n_k)^{-1} \|(\hat{\Theta})J_{G_k} - \Theta G_k\|_0,
\]

where \( \hat{\Theta} \) can be \( \hat{\Theta}^p, \hat{\Theta}^s \) or \( \hat{\Theta} \). The third metric is the derivation of the estimated link probability matrix \( \hat{B} \) from the true link probability matrix \( B \), namely, \( \| \hat{B} - B \|_\infty \), where \( \hat{B} \) can be \( \hat{B}^p, \hat{B}^s \), or the counterpart corresponding to the original spectral clustering. Throughout this section, the SBMs parameterized by \( (\Theta, B) \) were homogeneously generated in the following way,

\[
P = \Theta B \Theta^T = \Theta (\alpha_n \lambda I_K + \alpha_n(1 - \lambda)1_K 1_K^T) \Theta^T,
\]
where $1_K$ represents a $K$ dimensional vector of 1’s and $\lambda$ is a constant, and the community sizes are balanced to be $n/K$. To see how the above mentioned metrics change with $n$, $K$, $\alpha_n$, we conduct the following four experiments.

**Experiment 1.** In this experiment, we aim to evaluate the effect of $n$ on the three metrics. To that end, we let $n$ vary while keeping the other parameters fixed at $K = 3$, $\alpha_n = 0.2$, $\alpha_n(1 - \lambda) = 0.1$, $q = 2$, $r = 10$, $p = 0.7$. The random test matrix in the random projection scheme was generated with iid standard Gaussian entries, respectively. Figure 1 shows the average results of 20 replications, where “nonrandom” refers to the original spectral clustering. Recall that the error bound for $P$ increases with order $O(\sqrt{n})$, the error bound for $\Theta$ decreases with order $O(1/\alpha_n)$, and the error bound for $B$ vanishes as $n$ goes to infinity. As expected, from Figure 1 we can see that the randomized methods perform worse than the original spectral clustering when $n$ is small, say $n < 600$, but they become almost identical when $n$ becomes large, say $n > 800$, which is actually the focus of this paper (see Figure 1(b) and (c)). As for the approximation error, we see that the random projection and the random sampling perform better than the original spectral clustering (see Figure 1(a)), which is partially because of the constants’ effects.

**Experiment 2.** In this experiment, we evaluate the effect of $\alpha_n$ on the three metrics. We fix the sample size for the moment, and focus on the influence of the maximum link probability $\alpha$. Specifically, we let $\alpha$ vary and the between cluster probability was set as $\alpha(1 - 0.5)$ varying with $\alpha$. The sample size $n$ was fixed at 1152. The other parameters were the same as those in Experiment 1. Figure 2 displays the average results of 20 replications. By the theoretical results, we know that the error bound for $P$ increases with order $O(\sqrt{\alpha})$, the error bound for $\Theta$ decreases with order $O(1/\alpha_n)$, and the error bound for $B$ decreases ultimately after some increase at the beginning as $\alpha$ increases. The empirical results in Figure 2 coincide with the theoretical results in some sense. The error for $P$ increases slowly with $\alpha_n$, while the error for $\Theta$ and $B$ both decrease eventually with $\alpha_n$. In addition, the gap between the randomized and the original spectral clustering in Figure 2(b) and (c) closes as $\alpha$ increases.

**Experiment 3.** In this experiment, we test the effect of $K$ on the three metrics. Specifically, we let $K$ vary, the within cluster probability $\alpha = 0.2$, and the between cluster probability $\alpha(1 - 0.5) = 0.1$, respectively. The other parameters were the same as those in Experiment 2. The average results of 20 replications are shown in Figure 3. The theoretical bounds indicate that the error bound for $\Theta$ increases with order $O(K^3)$, and the error bound for $B$ increases with $K$. As expected, the empirical results support the theoretical findings (see Figure 3(b) and (c)). The error for $\Theta$ and $B$ both increases with $K$. While for the approximation error, recall that our randomized $A$ attains the minimax optimal rate which does not rely on $K$ (see Theorem 3.6 of Gao, Lu, and Zhou (2015)). Empirically, from Figure 3(a) we see that the approximation error for $P$ changes slowly as $K$ increases, which is partially due to the randomness throughout the experimental procedure.

**Experiment 4.** In the above three experiments, we fixed all the other parameters except the one that we pay attention to. Indeed, in view of the theoretical bounds, all the parameters can vary with $n$. To see the so-called high-dimensional performance of each method, in this experiment, we consider a simple setting that the within cluster and between cluster probabilities decrease with $n$ according to $\alpha_n = 2/\sqrt{n}$ and $\alpha_n(1 - 0.5) = 1/\sqrt{n}$, respectively. In such a setting, to ensure the decreasing trend of the misclustered error, $K$ should be of smaller order than $n^{1/6}$, which is rather small for $n$ smaller than, say, 1000. Hence we set $K = 2$ for simplicity. The other parameters were the same as those in Experiment 2. Figure 4 shows the average curves for each method in terms of three metrics. As expected, the misclassification error and the error for $B$ both decrease as $n$ increases, showing the high-dimensional feature of the theoretic methods. In addition, the performance of randomized methods become close to that of the original spectral clustering as $n$ increases.

### 7.2. Model Extensions

Besides the simple SBMs we considered in Section 7.1, we here consider the following six more complex models.

- **Model 1.** (Full-rank SBM with random $B$): $K = 3$, and the elements of $B$ are generated randomly according to $B_{ij} \sim$
Figure 5. Averaged results of each method over 20 replications on Models 1–3. Each column corresponds to a model. The first, second, and third row corresponds to the approximation error, the misclassification error, and the estimation error for link probability matrix, respectively.
Uniform(0.2, 0.3) and $B_{ij} \sim$ Uniform(0.01, 0.1), the community sizes are balanced.

- **Model 2.** (Full-rank SBM with random $B$ and unbalanced communities): The parameter set-up is identical to that of Model 1 except that the proportions of the number nodes within of each community over that of the whole nodes are $\frac{1}{6}, \frac{1}{2}, \frac{1}{3}$, respectively.

- **Model 3.** (Rank-deficient SBM): $K = 3$, and the community sizes are balanced. The link probability matrix $B := C C^T$ where

\[
C := \begin{bmatrix}
\frac{2\sin 0}{6} & \frac{2\cos 0}{3} \\
\frac{3\sin \frac{\pi}{2}}{6} & \frac{5\cos \frac{\pi}{2}}{6} \\
\end{bmatrix}
\]

- **Model 4.** (Full-rank DC-SBM): $K = 3$, and the elements of $B$ are generated randomly according to $B_{ii} \sim$ Uniform(0.4, 0.6) and $B_{ij} \sim$ Uniform(0.01, 0.2). Within each true cluster $k$, $\theta_{i \in G_k}$’s are iid 0.2 with probability (w.p.) 0.8, and 1 w.p. 0.2. $\theta$’s are then normalized such that its maximum value is 1 within each true cluster.

- **Model 5.** (Full-rank DC-SBM with more heterogeneity): Except for the node propensity parameter $\theta$, the parameter set-up is identical to that of Model 4. The $\theta$’s are generated as follows. Within each true cluster $k$, $\theta_{i \in G_k}$’s are iid with its element being 0.1 w.p. 0.4, being 0.2 w.p. 0.4, and being 1 with probability 0.2. $\theta$’s are then normalized such that its maximum value is 1 within each true cluster.

- **Model 6.** (Rank-deficient DC-SBM): The parameter set-up is identical to that of Model 4 except the formulation of $B$. Particularly, $B$ is the same with that in Model 3.

Figures 5 and 6 display the averaged results over 20 replications of model 1–3 and model 4–6, respectively. It can be seen that for all models we tested, the clustering performance of the randomized spectral clustering algorithms become close to that of the original spectral clustering as the sample size $n$ increases, coinciding with theoretical results.
Figure 7. Effects of the parameters $r, q, \Omega$ in the random projection scheme. Each row corresponds to the effect of one parameter with the others fixed. Each column corresponds to a measure for the clustering performance. The other parameters are fixed at $n = 1152, K = 3$, and the within cluster probability $\alpha = 0.2$. 
In the above experiments, we only considered assortative networks where nodes tend to be connected with those in the same community, which is mainly because that we require the link probability matrix to be diagonally dominant in some sense (Lemmas 2 and 4). For the disassortative networks, it would be our future work to study their clustering methods specifically in the rank deficient setting, though it is suggest that the absolute eigenvalue would help finding the proper communities in full rank setting (Rohe, Chatterjee, and Yu 2011).

### 7.3. Additional Experiments

To see how hyper parameters $r, q, p$ and the distribution of test matrix affect the performance of corresponding method, we here conduct another series of experiments. Specifically, to remove the computational cost of finding the best permutation matrix over the permutation matrix set, we use $F_1$ score ($F_1$), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) (Hubert and Arabie 1985; Manning, Raghavan, and Schütze 2010) to justify the clustering performance of each method. These indexes measure the similarity of two clusters, and here we refer to the estimated and true clusters, from different perspectives. The larger these indexes, the better the clustering algorithm performs. The parameters were basically set as $n = 1152$, $K = 3$, and the within cluster probability $\alpha = 0.2$. To see the effect of other parameters, we varied the oversampling parameter $r \in \{0, 4, 8, 12\}$, the power parameter $q \in \{2, 4, 6\}$, the sampling rate $p \in \{0.6, 0.7, 0.8, 0.9\}$, and the distribution of test matrix $\Omega$ was generated as iid Gaussian (standard), uniform (from $-1$ to $1$), and Rademacher (take values $+1$ and $-1$ with equal probability). And for each setting, we let the between cluster probability vary. Figures 7 and 8 show the averaged results of the random projection scheme and the random sampling scheme, respectively. As expected, larger $r$, $q$ and $p$ lead to better clustering performance but at the cost of computational efficiency. One should choose these parameters according to the problem at hand. In addition, among the distribution of $\Omega$ we tested, it has little effect on the resulting clustering performance of random projection.

### 8. Real Data Examples

In this section, we numerically evaluate the merits of randomized spectral clustering in terms of accuracy and efficiency. Specifically, we first compare the clustering accuracy of each method on four small-scale real networks, using the original spectral clustering as the baseline method. After that, we examine the computational efficiency as well as the relative error of

(continued)
each method on four large-scale networks, where we compare randomized methods with several iterative methods.

### 8.1. Accuracy Evaluation

In this section, we test the effectiveness of randomized spectral clustering on four network datasets, including the European e-mail network (Leskovec, Kleinberg, and Faloutsos 2007; Yin et al. 2017), the political blog network (Adamic and Glance 2005), the statistician coauthor network and the statistician citation network (Ji and Jin 2016), where the first two datasets have ground truth community assignments and the last two have no ground truth community assignment. Table 1 shows the basic statistics about the networks, where for the first two networks, the target rank is set as the number of true clusters, while for the last two networks, the target rank follows Ji and Jin (2016). For the datasets with ground truth labels, we computed F1, NMI, and ARI (Hubert and Arabie 1985; Manning, Raghavan, and Schütze 2010) between the estimated clusters and the true clusters for each of the three methods, namely, the random projection, the random sampling, and the original spectral clustering, respectively. While for the datasets without ground truth labels, we computed F1, NMI, and ARI between the clusters estimated by the randomized spectral clustering and the clusters estimated by the original spectral clustering. Our aim is to show that randomized algorithms perform comparably to the original spectral clustering. Hence for the datasets with ground truth labels, the smaller gap of F1, NMI, and ARI between randomized and original spectral clustering indicate the better match between these methods. While for the datasets without ground truth labels, larger F1, NMI, and ARI indicate the better match. For the random projection scheme, the oversampling parameter $r = 10$, the power parameter $q = 2$, and the random test matrix has iid Gaussian entries. And for the random sampling scheme, we test two cases, namely, $p = 0.7$ and 0.8. Table 2 summarizes the average performance of these methods over 20 replications with the standard deviations in the parentheses. From Table 2, we see that all the methods perform very similarly to each other in terms of F1, NMI, and ARI, and the results are rather stable, which shows the effectiveness of randomized methods.

### 8.2. Efficiency Evaluation

In this section, we examine the computational efficiency of randomized methods for partial eigenvalue decomposition on four large-scale real undirected networks, including DBLP co-

---

**Figure 9.** The pairwise comparison of the clustering results of six methods on four large-scale networks. The relative clustering performance are measured by ARI. Larger ARI, that is, larger circles in the figure, indicates that the clustering results of the two methods are more close.

**Table 4.** Median time (seconds) of each method for computing the (approximated) eigenvectors of four real network adjacency matrices over 20 replications, where for the random sampling, the time with the sampling time included and excluded (shown in the parentheses) are reported, respectively.

| Networks  | Random projection | Random sampling | irlba | svds | svdr | partial_eigen |
|-----------|-------------------|-----------------|-------|------|-----|-------------|
| DBLP      | 0.369             | 0.280(0.248)    | 0.341 | 0.411| 6.132| 0.346        |
| Youtube   | 2.037             | 2.302(2.204)    | 2.487 | 3.043| 35.595| 9.111        |
| Internet  | 2.773             | 2.072(1.774)    | 3.404 | 3.332| 30.900| 7.706        |
| LiveJournal| 13.213            | 7.207(6.216)    | 15.179| 20.077| 106.166| 15.080       |
laboration network, YouTube social network, Internet topology network, LiveJournal social network (Leskovec, Kleinberg, and Faloutsos 2005; Yang and Leskovec 2015). These four networks are large-scale with up to millions of nodes and tens of millions of edges. Table 3 shows the basic statistics about the networks, where the target rank corresponds to a network is $k$ if there exists a large gap between the $k$-th and $(k + 1)$-th largest (in absolute value) approximated eigenvalues. We compare the performance of our methods with iterative methods, including the implicitly restarted Lanczos algorithm (Calvetti, Reichel, and Sorensen 1994) (svds in the R package RSpectra (Qiu and Mei 2019)), the augmented implicitly restarted Lanczos bidiagonalization algorithms (Baglama and Reichel 2005) (irlba and partial_eigen in the R package irlba (Baglama, Reichel, and Lewis 2019)). In addition, we also compare our implementation with the randomized methods implemented in svdr in R package irlba (Baglama, Reichel, and Lewis 2019). Note that the full eigenvalue decomposition always fails in such large-scale data setting.

Table 4 shows the median computational time of each method over 20 replications, where all computations are done on a machine with Intel Core i9-9900K CPU 3.60GHz, 32GB memory, and 64-bit WS operating-system, and R version 4.0.4 is used for all computations. For the random projection-based method, the power parameter is 2 and the oversampling parameter is 10. For the random sampling-based method, the sampling probability is 0.7. We can see from Table 4 that the our methods shows great advantage over compared methods especially when the network scale is large. In particular, the random sampling-based method is efficient no matter the sampling time is included or not.

Figure 9 shows the pairwise comparison of the clustering results of six methods on these four networks. The relative clustering performance are measured by ARI. It turns out that the random projection-based method and the random sampling-based method yield similar results with other compared methods, though the random sampling-based method seems to behave slightly different.

Overall, as indicated by our theory and experiments, the randomized methods could bring high efficiency while slightly sacrificing the accuracy. In real world applications, one should balance the accuracy-efficiency tradeoff via selecting appropriate hyper parameters according to the real setting.

9. Conclusion

In this article, we used randomized sketching techniques to accelerate the spectral clustering when facing large-scale networks, say networks with millions of nodes, and studied how well the resulting algorithms perform under the SBMs and DC-SBMs. We studied two randomized spectral clustering algorithms. The first one is random projection-based, which reduces the computational cost by projecting the columns and rows of the adjacency matrix to a lower-dimensional space. The second one is random sampling-based, which samples the edges to obtain a sparsified adjacency matrix, and thus, reducing the computational cost of partial eigen-decomposition. In the framework of SBMs, we studied these two randomized spectral clustering algorithms in terms of the approximation error that measures the deviation of the randomized adjacency matrix $\tilde{A}$ from the population matrix $P$, the misclassification error that measures the fraction of the number of misclustered nodes over the total number of nodes, and the estimation error for the link probability matrix $B$. In particular, we considered a more generalized content where $\text{rank}(B)$ could be smaller than the cluster number $K$. Under mild conditions, the approximation error turns out to be statistically optimal, which shows that the randomized matrix behaves as if it was sampled from the SBM. We also extend theoretical results to DC-SBMs. Experimental results showed the merits of randomized spectral clustering on networks with up to millions of nodes. For practical convenience, we developed an R package Rclust.

There are many ways that the content in this article can be extended. First, we studied the weak consistency of the pure spectral clustering without any regularization or refinement, and we mainly used the Davis–Kahan theorem to study the eigenvector perturbation. There exist several works on trimming or refining the pure spectral clustering to help the refined spectral clustering achieve the information-theoretical limit of the exact recovery (strong consistency) or minimax optimal rate of the partial recovery (weak consistency) under SBMs; see Gao et al. (2017) and Yun and Proutiere (2016), among others. It would be interesting to study whether one could use similar treatments on the randomized spectral clustering in order to improve its theoretical performance while without increasing the time cost. On the other hand, a few works study the entry-wise perturbation of eigenvectors very recently; see Cape, Tang, and Priebe (2019), Tang, Cape, and Priebe (2021), Abbe et al. (2020) and Su, Wang, and Zhang (2019), among others. It would be important and insightful to study the entry-wise perturbation of eigenvectors after randomization, and also study the related strong consistency in SBMs. Second, although the approximation error is minimax optimal under SBMs, it would be important to develop advanced randomization techniques in order to weaken the condition on $\alpha_n$, $q$, $p_j$’s. In addition, we mainly focused on the adjacency matrix sampled from the SBMs. It would be interesting to generalize the results to the Laplacian matrix and other network generating models, say – the latent space model, and the graphon models, among others. Finally, it would be important to study the estimation of the model parameters $K$ and rank($B$) (Fishkind et al. 2013; Ma, Su, and Zhang 2021).

Supplementary Materials

Proofs and Code are included. R package is available at https://github.com/XiaoGuo-stat/Rclust.

Acknowledgments

We are grateful to the editor, associate editor, and two reviewers for their consideration of our paper and for their helpful suggestions, which led to an improved version of this paper. We also thank Professor Yixuan Qiu for his great help in developing the R package.

Funding

Our research is partially supported by National Natural Science Foundation of China (No.U1811461), National Natural Science Foundation of China.
Li, T., Levina, E., and Zhu, J. (2020a), “Community Models for Partially Observed Networks from Surveys,” arXiv preprint arXiv:2008.03652. [896]

(2020b), “Network Cross-Validation by Edge Sampling,” Biometrika, 107, 257–276. [888, 896]

Liao, Z., Coupillet, R., and Mahoney, M. W. (2020), “Sparse Quantized Spectral Clustering,” arXiv preprint arXiv:2010.01376. [896]

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. (2012), “Robust Recovery of Subspace Structures by Low-Rank Representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 171–184. [895]

Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A., Priebe, C. E. (2014), “Perfect Clustering for Stochastic Blockmodel Graphs via Adjacency Spectral Embedding,” Electronic Journal of Statistics, 8, 2905–2922. [896]

Ma, P., Mahoney, M. W., and Yu, B. (2015), “A Statistical Perspective on Algorithmic Leveraging,” The Journal of Machine Learning Research, 16, 861–911. [887]

Ma, S., Su, L., and Zhang, Y. (2021), “Determining the Number of Communities in Degree-Corrected Stochastic Block Models,” Journal of Machine Learning Research, 22, 1–63. [904]

Mahoney, M. W. (2011), “Randomized Algorithms for Matrices and Data,” Foundations and Trends in Machine Learning, 3, 123–224. [887]

Mahoney, M. W., and Drineas, P. (2009), “Cur Matrix Decompositions for Improved Data Analysis,” Proceedings of the National Academy of Sciences, 106, 697–702. [887]

Manning, C., Raghavan, P., and Schütze, H. (2010), “Introduction to Information Retrieval,” Natural Language Engineering, 16, 100–103. [902, 903]

Martinsson, P. G. (2016), “Randomized Methods for Matrix Computations,” arXiv preprint arXiv:1607.01649. [887, 888, 890]

Martinsson, P. G., and Tropp, J. (2020), “Randomized Numerical Linear Algebra: Foundations & Algorithms,” arXiv preprint arXiv:2002.01387. [892]

Matoušek, J. (2000), “On Approximate Geometric k-Clustering,” Discrete & Computational Geometry, 24, 61–84. [893]

Newman, M. (2018), Networks, Oxford: Oxford University Press. [887]

Paul, S., and Chen, Y. (2020), “Spectral and Matrix Factorization Methods for Consistent Community Detection in Multi-Layer Networks,” The Annals of Statistics, 48, 230–250. [896]

Planci, M., and Wainwright, M. J. (2016), “Iterative Hessian sketch: Fast and Accurate Solution Approximation for Constrained Least-Squares,” The Journal of Machine Learning Research, 17, 1842–1879. [887]

Qi, T., and Rohe, K. (2013), “Regularized Spectral Clustering Under the Degree-Corrected Stochastic Blockmodel,” in Advances in Neural Information Processing Systems, pp. 3120–3128. [888, 889, 896]

Qi, Y., and Mei, J. (2019), RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems. Available at https://CRAN.R-project.org/package=RSpectra. [896, 904]

Raskutti, G., and Mahoney, M. W. (2016), “A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares,” The Journal of Machine Learning Research, 17, 7508–7538. [887]

Rohe, K., Chatterjee, S., and Yu, B. (2011), “Spectral Clustering and the High-Dimensional Stochastic Block Model,” The Annals of Statistics, 39, 1878–1915. [888, 890, 893, 896, 902]

Sakai, T., and Imiya, A. (2009), “Fast Spectral Clustering with Random Projection and Sampling,” in International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 372–384. Springer. [896]

Sarkar, P., and Bickel, P. J. (2015), “Role of Normalization in Spectral Clustering for Stochastic Blockmodels,” The Annals of Statistics, 43, 962–990. [888]

Sinha, K. (2018), “K-means Clustering Using Random Matrix Sparsification,” in International Conference on Machine Learning, pp. 4684–4692. PMLR. [896]

Su, L., Wang, W., and Zhang, Y. (2019), “Strong Consistency of Spectral Clustering for Stochastic Block Models,” IEEE Transactions on Information Theory, 66, 324–338. [888, 896, 904]

Tang, M., Cape, J., and Priebe, C. E. (2021), “Asymptotically Efficient Estimators for Stochastic Blockmodels: The Naive mle, the Rank-Constrained mle, and the Spectral,” Bernoulli, Preprint. [888, 893, 896, 904]

Terada, Y. (2014), “Strong Consistency of Reduced k-means Clustering,” Scandinavian Journal of Statistics, 41, 913–931. [895]

Tremblay, N., and Loukas, A. (2020), “Approximating Spectral Clustering via Sampling: A Review,” Sampling Techniques for Supervised or Unsupervised Tasks, pp. 129–183. [896]

Tremblay, N., Puy, G., Grisonval, R., and Vandergheynst, P. (2016), “Compressive Spectral Clustering,” in International Conference on Machine Learning, pp. 1002–1011. PMLR. [896]

Tropp, J. A., Yurtsever, A., Udell, M., and Cevher, V. (2016), “Randomized Single-View Algorithms for Low-Rank Matrix Approximation,” arXiv preprint, arXiv:1609.00048. [896]

Vidal, R., Ma, Y., and Sastry, S. (2005), “Generalized Principal Component Analysis (gPCA),” IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 1945–1959. [895]

Von Luxburg, U. (2007), “A Tutorial on Spectral Clustering,” Statistics and Computing, 17, 395–416. [888, 889]

Wang, H. (2019), “More Efficient Estimation for Logistic Regression with Optimal Subsamples,” Journal of Machine Learning Research, 20, 1–59. [887]

Wang, H., Zhu, R., and Ma, P. (2018), “Optimal Subsampling for Large Sample Logistic Regression,” Journal of the American Statistical Association, 113, 829–844. [887]

Wang, H., Yang, M., and Stufken, J. (2019a), “Information-Based Optimal Subdata Selection for Big Data Linear Regression,” Journal of the American Statistical Association, 114, 393–405. [887]

Wang, S., Gittens, A., and Mahoney, M. W. (2017), “Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging,” The Journal of Machine Learning Research, 18, 8039–8088. [887]

Wang, S., Gittens, A., and Mahoney, M. W. (2019a), “Scalable Kernel k-means Clustering with Nyström Approximation: Relative-Error Bounds,” Journal of Machine Learning Research, 20, 1–49. [896]

Witten, R., and Candès, E. (2015), “Randomized Algorithms for Low-Rank Matrix Factorizations: Sharp Performance Bounds,” Algorithmica, 72, 264–281. [887, 888, 890]

Yan, D., Huang, L., and Jordan, M. I. (2009), “Fast Approximate Spectral Clustering,” in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 907–916. [896]

Yang, C., Priebe, C. E., Park, Y., and Marchette, D. J. (2020), “Simultaneous Dimensionality and Complexity Model Selection for Spectral Graph Clustering,” Journal of Computational and Graphical Statistics, 30, 422–441. [888, 896]

Yang, J., and Leskovec, J. (2015), “Defining and Evaluating Network Communities Based on Ground-Truth,” Knowledge and Information Systems, 42, 181–213. [904]

Yin, H., Benson, A. R., Leskovec, J., and Gleich, D. F. (2017), “Local Higher-Order Graph Clustering,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 555–564. ACM. [903]

Yun, S.-Y., and Proutiere, A. (2014), “Accurate Community Detection in the Stochastic Block Model via Spectral Algorithms,” arXiv preprint arXiv:1412.7335. [896]

Yun, S.-Y., and Proutiere, A. (2016), “Optimal Cluster Recovery in the Labeled Stochastic Block Model,” in Advances in Neural Information Processing Systems, pp. 965–973. [896, 904]

Zhou, J., Tu, Y., Chen, Y., and Wang, H. (2017), “Estimating Spatial Autocorrelation with Sampled Network Data,” Journal of Business & Economic Statistics, 35, 130–138, 2017. [887]